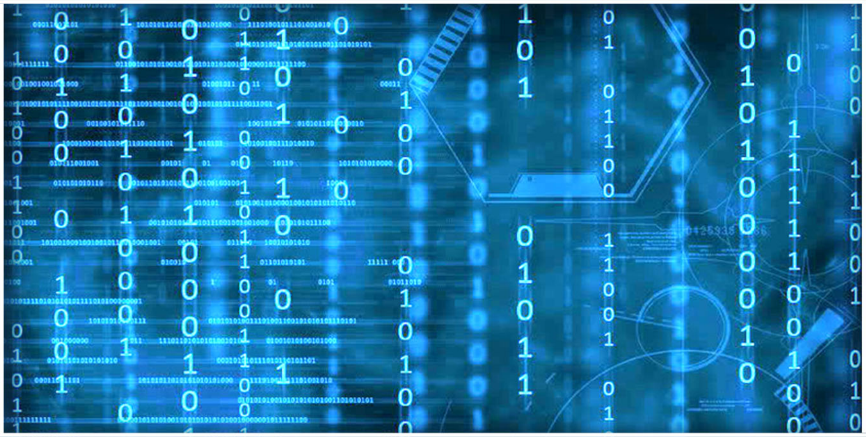
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 7 Issue 2

February 2016



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

SAI

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA

Volume 7 Issue 2 February 2016
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

CONTENTS

Paper 1: Fruit Fly Optimization Algorithm for Network-Aware Web Service Composition in the Cloud
Authors: Umar SHEHU, Ghazanfar SAFDAR, Gregory EPIPHANIOU
PAGE 1 – 11

Paper 2: Role of Security in Social Networking
Authors: David Hiatt, Young B. Choi
PAGE 12 – 15

Paper 3: The SVM Classifier Based on the Modified Particle Swarm Optimization
Authors: Liliya Demidova, Evgeny Nikulchev, Yulia Sokolova
PAGE 16 – 24

Paper 4: 3D Servicescape Model: Atmospheric Qualities of Virtual Reality Retailing
Authors: Aasim Munir Dad, Professor Barry Davies, Dr. Asma Abdul Rehman
PAGE 25 – 38

Paper 5: Hidden Markov Models (HMMs) and Security Applications
Authors: Rubayyi Alghamdi
PAGE 39 – 47

Paper 6: Toward Secure Web Application Design: Comparative Analysis of Major Languages and Framework Choices
Authors: Stephen J. Tipton, Young B. Choi
PAGE 48 – 53

Paper 7: Pattern Visualization Through Detection Plane Generation for Macroscopic Imagery
Authors: Hanan Hassan Ali Adlan
PAGE 54 – 57

Paper 8: Hierarchical Compressed Sensing for Cluster Based Wireless Sensor Networks
Authors: Vishal Krishna Singh, Manish Kumar
PAGE 58 – 67

Paper 9: A Model for Classification Secondary School Student Enrollment Approval Based on E-Learning Management System and E-Games
Authors: Hany Mohamed El-katary, Essam M. Ramzy Hamed, Safaa Sayed Mahmoud
PAGE 68 – 73

Paper 10: A Random-Walk Based Privacy-Preserving Access Control for Online Social Networks
Authors: You-sheng Zhou, En-wei Peng, Cheng-qing Guo
PAGE 74 – 79

Paper 11: Pricing Schemes in Cloud Computing: An Overview
Authors: Artan Mazrekaj, Isak Shabani, Besmir Sejdiu
PAGE 80 – 86

Paper 12: Analysis of a Braking System on the Basis of Structured Analysis Methods
Authors: Ben Salem J., Lakhoua M.N., El Amraoui L.
PAGE 87 – 92

Paper 13: Contributions to the Analysis and the Supervision of a Thermal Power Plant
Authors: Lakhoua M.N, Glaa R., Ben Hamouda M., El Amraoui L.
PAGE 93 – 101

Paper 14: Effective Teaching Methods and Proposed Web Libraries for Designing Animated Course Content: A Review
Authors: Rajesh Kumar Kaushal, Dr. Surya Narayan Panda
PAGE 102 –106

Paper 15: An Augmented Reality Approach to Integrate Practical Activities in E-Learning Systems
Authors: EL KABTANE Hamada, EL ADNANI Mohamed, SADGAL Mohamed, MOURDI Youssef
PAGE 107 – 117

Paper 16: Designing of Hydraulically Balanced Water Distribution Network Based on GIS and EPANET
Authors: RASOOLI Ahmadullah, KANG Dongshik
PAGE 118 – 125

Paper 17: Corporate Responsibility in Combating Online Misinformation
Authors: Fadi Safieddine, Wassim Masri, Pardis Pourghomi
PAGE 126 – 132

Paper 18: A Framework for Classifying Unstructured Data of Cardiac Patients: A Supervised Learning Approach
Authors: Iqra Basharat, Ali Raza Anjum, Mamuna Fatima, Usman Qamar, Shoab Ahmed Khan
PAGE 133 – 141

Paper 19: Optimum Route Selection for Vehicle Navigation
Authors: Dalip, Vijay Kumar
PAGE 142 – 148

Paper 20: Skip List Data Structure Based New Searching Algorithm and Its Applications: Priority Search
Authors: Mustafa Aksu, Ali Karcı
PAGE 149 – 154

Paper 21: Self-Organized Hash Based Secure Multicast Routing Over Ad Hoc Networks
Authors: Amit Chopra, Dr. Rajneesh Kumar
PAGE 155 – 163

Paper 22: Ranking Documents Based on the Semantic Relations Using Analytical Hierarchy Process
Authors: Ali I. El-Dsouky, Hesham A. Ali, Rabab S. Rashed
PAGE 164 – 173

Paper 23: The Effect of Parallel Programming Languages on the Performance and Energy Consumption of HPC Applications
Authors: Muhammad Aqib, Fadi Fouad Fouz
PAGE 174 – 179

Paper 24: Image Transmission Model with Quality of Service and Energy Economy in Wireless Multimedia Sensor Network
Authors: Benlabbes Haouari, Benahmed Khelifa, Beladgham Mohammed
PAGE 180 – 184

Paper 25: Dual Security Testing Model for Web Applications
Authors: Singh Garima, Kaushik Manju
PAGE 185 – 191

Paper 26: Paper Automatic Keyphrase Extractor from Arabic Documents

Authors: Hassan M. Najadat, Ismail I. Hmeidi, Mohammed N. Al-Kabi, Maysa Mahmoud Bany Issa

PAGE 192 – 199

Paper 27: A Topic Modeling Based Solution for Confirming Software Documentation Quality

Authors: Nouh Alhindawi, Obaida M. Al-Hazaimeh, Rami Malkawi, Jamal Alsakran

PAGE 200 – 206

Paper 28: A New Hybrid Network Sniffer Model Based on Pcap Language and Sockets (Pcapsocks)

Authors: Azidine GUEZZAZ, Ahmed ASIMI, Yassine SADQI, Younes ASIMI, Zakariae TBATOU

PAGE 207 – 214

Paper 29: Choosing a Career Based Personality Matching: A Case Study of King Abdulaziz University

Authors: Nahla Aljojo

PAGE 215 – 221

Paper 30: Tracking Items Through Rfid and Solving Heterogeneity Problems During a Collaboration Between Port Companies

Authors: Mehdi ABID, Benayad NSIRI, Yassine SERHANE, Haitam AGHARI

PAGE 222 – 228

Paper 31: Frequency Domain Analysis for Assessing Fluid Responsiveness by Using Instantaneous Pulse Rate Variability

Authors: Pei-Chen Lin, Chia-Chi Chang, Hung-Yi Hsu, Tzu-Chien Hsiao

PAGE 229 – 233

Paper 32: Arabic Studies' Progress in Information Retrieval

Authors: Essam Hanandeh, Hayel Khafajah

PAGE 234 – 238

Paper 33: Preliminary Study of Software Performance Models

Authors: Issamjebreen, Mohammed Awad

PAGE 239 – 242

Paper 34: Implementing Project Management Category Process Areas of CMMI Version 1.3 Using Scrum Practices, and Assets

Authors: Ahmed Bahaa Farid, A. S. Abd Elghany, Yehia Mostafa Helmy

PAGE 243 – 252

Paper 35: Development of System Architecture for E-Government Cloud Platforms

Authors: Margulan Aubakirov, Evgeny Nikulchev

PAGE 253 – 258

Paper 36: Comparative Study for Software Project Management Approaches and Change Management in the Project Monitoring & Controlling

Authors: Amira M. Gaber, Sherief Mazen, Ehab E. Hassanein

PAGE 259 – 264

Paper 37: N-ary Relations of Association in Class Diagrams: Design Patterns

Authors: Sergievskiy Maxim

PAGE 265 – 268

Paper 38: Towards Domain Ontology Creation Based on a Taxonomy Structure in Computer Vision

Authors: Mansouri fatimaezzahra, Sadgal mohamed, Elfazziki abdelaziz, Benchikhi loubna

PAGE 269 – 278

Paper 39: The Discovery of the Implemented Software Engineering Process Using Process Mining Techniques

Authors: Mostafa Adel Zayed, Ahmed Bahaa Farid

PAGE 279 – 286

Paper 40: A Novel Neural Network Based Method Developed for Digit Recognition Applied to Automatic Speed Sign Recognition

Authors: Hanene Rouabeh, Chokri Abdelmoula, Mohamed Masmoudi

PAGE 287 – 292

Paper 41: A Secure Cloud Computing Architecture Using Homomorphic Encryption

Authors: Kamal Benzekki, Abdeslam El Fergougui, Abdelbaki El Belrhiti El Alaoui

PAGE 293 – 298

Paper 42: Nonquadratic Lyapunov Functions for Nonlinear Takagi-Sugeno Discrete Time Uncertain Systems Analysis and Control

Authors: Ali Bouyahya, Yassine Manai, Joseph Haggège

PAGE 299 – 308

Paper 43: Design of Intelligent Control System of Transformer Oil Temperature

Authors: Caijun Xu, Liping Zhang, Yuchen Chen, Zhifeng Liu

PAGE 309 – 313

Paper 44: Design of Wireless Temperature Measuring System Based on the nRF24l01

Authors: Song Liu, Zhiqiang Yuan, Yuchen Chen

PAGE 314 – 317

Paper 45: Risk Diffusion Modeling and Vulnerability Quantification on Japanese Human Mobility Network from Complex Network Analysis Point of View

Authors: Kiyotaka Ide, Hiroshi Sato, Tran Quang Hoang Anh, Akira Namatame

PAGE 318 – 327

Paper 46: Detection and Defense Against Packet Drop Attack in MANET

Authors: Tariq Ahamad

PAGE 328 – 331

Paper 47: The Enhanced Arabchat: An Arabic Conversational Agent

Authors: Mohammad Hijjawi, Zuhair Bandar, Keeley Crockett

PAGE 332 – 347

Paper 48: Understanding a Co-Evolution Model of Business and IT for Dynamic Business Process Requirements

Authors: Muhammad Asif Khan

PAGE 348 – 352

Paper 49: Performance Evaluation of Loss Functions for Margin Based Robust Speech Recognition

Authors: Syed Abbas Ali, Maria Andleeb, Raheela Asif, Danish-ur-Rehman

PAGE 353 – 361

Paper 50: Improving Quality of Vietnamese Text Summarization Based on Sentence Compression
Authors: Ha Nguyen Thi Thu, Tu Nguyen Ngoc, Cuong Nguyen Ngoc, Hiep Xuan Huynh
PAGE 362 – 366

Paper 51: Virtual Identity Approaches Evaluation for Anonymous Communication in Cloud Environments
Authors: Ibrahim A.Gomaa, Emad Abd-Elrahman, Mohamed Abid
PAGE 367 – 376

Paper 52: A Multi-Criteria Decision Method in the DBSCAN Algorithm for Better Clustering
Authors: Abdellah IDRISSI, Altaf ALAOUI
PAGE 377 – 384

Paper 53: CAT5:A Tool for Measuring the Maturity Level of Information Technology Governance Using COBIT 5 Framework
Authors: Souhaïl El ghazi El Houssaïni, Karim Youssfi, Jaouad Boutahar
PAGE 385 – 391

Paper 54: A Platform to Support the Product Servitization
Authors: Giovanni Di Orio, Oliviu Matei, Sebastian Scholze, Dragan Stokic, José Barata, Claudio Cenedese
PAGE 392 – 400

Paper 55: On Shear Wave Speed Estimation for Agar-Gelatine Phantom
Authors: Hassan M. Ahmed, Nancy M. Salem, Ahmed F. Seddik, Mohamed I. El Adawy
PAGE 401 – 409

Paper 56: A Variant of Genetic Algorithm Based Categorical Data Clustering for Compact Clusters and an Experimental Study on Soybean Data for Local and Global Optimal Solutions
Authors: Abha Sharma, R. S. Thakur
PAGE 410 – 419

Paper 57: Improving Throughput and Delay by Signaling Modification in Integrated 802.11 and 3G Heterogeneous Wireless Network
Authors: Majid Fouladian, Mohammad Ali Pourmina, Faramarz Hendessi
PAGE 420 – 428

Paper 58: Constructing Relationship Between Software Metrics and Code Reusability in Object Oriented Design
Authors: Manoj H.M, Dr. Nandakumar A.N
PAGE 429 – 438

Paper 59: An Ensemble of Fine-Tuned Heterogeneous Bayesian Classifiers
Authors: Amel Alhussan, Khalil El Hindi
PAGE 439 – 448

Paper 60: Cloud-Based Processing on Data Science for Visualization
Authors: Ahmad Ashari, A Min Tjoa, Mardhani Riasetiawan
PAGE 449 – 454

Paper 61: Comparative Study of Robust Control Strategies for a Dfig-Based Wind Turbine
Authors: Mohamed BENKAHLA, Rachid TALEB, Zinelaabidine BOUDJEMA
PAGE 455 – 462

Paper 62: Android Malware Detection & Protection: A Survey
Authors: Saba Arshad, Munam Ali Shah, Abid Khan, Mansoor Ahmed
PAGE 463 – 475

Paper 63: An Improved Transformer for LLC Resonant Inverter for Induction Heating Applications
Authors: Amira ZOUAOUI-KHEMAKHEM, Ferid KOURDA, Hamed belloumi
PAGE 476 – 483

Paper 64: An Empirical Investigation of Predicting Fault Count, Fix Cost and Effort Using Software Metrics
Authors: Raed Shatnawi, Wei Li
PAGE 484 – 491

Paper 65: Issues Elicitation and Analysis of CMMI Based Process Improvement in Developing Countries Theory and Practice
Authors: Shahbaz Ahmed Khan Ghayyur, Ahmed Noman Latif, Muhammad Daud Awan, Malik Sikandar Hayat Khiyal
PAGE 492 – 504

Paper 66: Energy Efficient Clustering Using Fixed Sink Mobility for Wireless Sensor Networks
Authors: Muhammad Ali Khan, Arif Iqbal Umar, Babar Nazir, Noor ul Amin, Shaukat Mehmood, Kaleem Habib
PAGE 505 – 510

Paper 67: A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools
Authors: D. P. Acharjya, Kauser Ahmed P
PAGE 511 – 518

Paper 68: A Cloud-Based Platform for Democratizing and Socializing the Benchmarking Process
Authors: Fuad Bajaber, Amin Shafaat, Omar Batarfi, Radwa Elshawi, Abdulrahman Altalhi, Ahmed Barnawi, Sherif Sakr
PAGE 519 – 530

Paper 69: Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews
Authors: Mahmoud Al-Ayyoub, Aya Nuseir, Ghassan Kanaan, Riyad Al-Shalabi
PAGE 531 – 539

Paper 70: Weighted G1-Multi-Degree Reduction of B´ezier Curves
Authors: Abedallah Rababah, Salisu Ibrahim
PAGE 540 – 545

Paper 71: Abnormalities Detection in Digital Mammography Using Template Matching
Authors: Ahmed M. Farrag
PAGE 546 – 550

Paper 72: An Anti-Pattern-based Runtime Business Process Compliance Monitoring Framework
Authors: Ahmed Barnawi, Ahmed Awad, Amal Elgammal, Radwa Elshawi, Abduallah Almalaise, Sherif Sakr
PAGE 551 – 572

Paper 73: Comprehensive Survey on Dynamic Graph Models
Authors: Aya Zaki, Mahmoud Attia, Doaa Hegazy, Safaa Amin
PAGE 573 – 582

Paper 74: Evolutionary Algorithms Based on Decomposition and Indicator Functions: State-of-the-art Survey

Authors: Wali Khan Mashwani, Abdellah Salhi, Muhammad Asif jan, Muhammad Sulaiman, Rashida Adeeb Khanum, Abdulmohsen Algarni

PAGE 583 – 593

Paper 75: Implementation and Evaluation of a Secure and Efficient Web Authentication Scheme using Mozilla Firefox and WAMP

Authors: Yassine SADQI, Ahmed ASIMI, Younes ASIMI, Zakaria TBATOU, Azidine GUEZZAZ

PAGE 594 – 603

Paper 76: Indoor Navigation System based on Passive RFID Transponder with Digital Compass for Visually Impaired People

Authors: A. M. Kassim, T. Yasuno, H. Suzuki, H. I. Jaafar, M. S. M. Aras

PAGE 604 – 611

Paper 77: Smart Cities: A Survey on Security Concerns

Authors: Sidra Ijaz, Munam Ali Shah, Abid Khan, Mansoor Ahmed

PAGE 612 – 625

Paper 78: Sorting Pairs of Points Based on Their Distances

Authors: Mohammad Farshi, Abolfazl Poureidi, Zorieh Soltani

PAGE 626 – 630

Paper 79: Theoretical and numerical characterization of continuously graded thin layer by the reflection acoustic microscope

Authors: Ahmed Markou, Hassan Nounah, Lahcen Mountassir

PAGE 631 – 636

Paper 80: Toward Information Diffusion Model for Viral Marketing in Business

Authors: Lulwah AlSuwaidan, Mourad Ykhlef

PAGE 637 – 646

Paper 81: A New Threshold Based Penalty Function Embedded MOEA/D

Authors: Muhammad Asif Jan, Nasser Mansoor Tairan, Rashida Adeeb Khanum, Wali Khan Mashwani

PAGE 647 – 655

Paper 82: Threshold Based Penalty Functions for Constrained Multiobjective Optimization

Authors: Muhammad Asif Jan, Nasser Mansoor Tairan, Rashida Adeeb Khanum, Wali Khan Mashwani

PAGE 656 – 667

Paper 83: Regression-Based Feature Selection on Large Scale Human Activity Recognition

Authors: Hussein Mazaar, Eid Emary, Hoda Onsi

PAGE 668 – 674

Paper 84: Reflected Adaptive Differential Evolution with Two External Archives for Large-Scale Global Optimization

Authors: Rashida Adeeb Khanum, Nasser Tairan, Muhammad Asif Jan, Wali Khan Mashwani, Abdel Salhi

PAGE 675 – 683

Fruit Fly Optimization Algorithm for Network-Aware Web Service Composition in the Cloud

Umar SHEHU
Department of Computer Science and Technology University of Bedfordshire Luton, UK

Ghazanfar SAFDAR
Department of Computer Science and Technology University of Bedfordshire Luton, UK

Gregory EPIPHANIOU
Department of Computer Science and Technology University of Bedfordshire Luton, UK

***Abstract*—Service Oriented Computing (SOC) provides a framework for the realization of loosely coupled service oriented applications. Web services are central to the concept of SOC. Currently, research into how web services can be composed to yield QoS optimal composite service has gathered significant attention. However, the number and spread of web services across the cloud data centers has increased, thereby increasing the impact of the network on composite service performance experienced by the user. Recently, QoS-based web service composition techniques focus on optimizing web service QoS attributes such as cost, response time, execution time, etc. In doing so, existing approaches do not separate QoS of the network from web service QoS during service composition. In this paper, we propose a network-aware service composition approach which separates QoS of the network from QoS of web services in the Cloud. Consequently, our approach searches for composite services that are not only QoS-optimal but also have optimal QoS of the network. Our approach consists of a network model which estimates the QoS of the network in the form of network latency between services on the cloud. It also consists of a service composition technique based on fruit fly optimization algorithm which leverages the network model to search for low latency compositions without compromising service QoS levels. The approach is discussed and the results of evaluation are presented. The results indicate that the proposed approach is competitive in finding QoS optimal and low latency solutions when compared to recent techniques.**

Keywords*—Web Services; Service Composition; QoS; Network Latency; Cloud; Fruit Fly Algorithm

I. INTRODUCTION

Service Oriented Computing (SOC) is a paradigm for building loosely coupled distributed systems [14]. It encapsulates the functional units of a distributed system into web services which perform specific tasks and are easily reusable in other software systems. The true value of SOC lies in its ability to compose different services to complete more complex user requests. Web services are central to the realization of SOC. A web service is defined as a network-accessible object that provides some functionality [9]. Web services are characterized by functional and non-functional attributes [13]. The functional attribute dictates what kind of task a web service is meant to perform e.g. credit card validation. On the other hand, the non-functional attribute, also known as Quality of Service (QoS), indicates service's level of quality. The QoS attribute is mainly used to differentiate services having similar functional attributes. Service providers normally advertise services together with their QoS levels as part of a Service Level Agreement (SLA). Some service QoS attributes advertised includes cost, response time and reputation [15].

A. Qos-aware service composition

QoS-based web service composition has received a lot of attention from the research community. It allows organizations to share their business processes with other service providers to facilitate delivery of service functionalities which satisfy complex user requests.

In many situations, a single service may not be able to satisfy a user's request. For instance, a web service is incapable of performing both credit card validation and hotel booking for a user attempting to plan a business trip. During such situations web services from different service providers are combined to meet the user's requirements. This is where service composition comes into play. It is the process of aggregating web services having disparate functionalities into a composite service. The composition of services is performed via their functional and QoS attributes. The functional attributes define what a service is capable of doing while QoS attributes represent the non-functional or quality aspects of a service e.g. service availability, reputation, response time, etc. The QoS attributes are used for composition only when the services involved have comparable functionalities. The goal of service composition is to search for a combination of services that leads to optimal QoS levels. During the composition process, QoS attributes for services are aggregated according to a composition's workflow pattern [2] into a composite service. Once a user request is issued, service composition breaks down the request into smaller tasks organized according to one of several workflow patterns [2]. Within each task, a number of functionally similar services offered by different service providers are made available for the aggregation process. In the next stage, a service is selected per task to form part of a composite service. As a result of the high number of services that currently exist on the Internet, there are large numbers of possible composition paths that can be formed. For example, suppose there exist eight tasks within a user request and each task can be executed by twenty possible web services. Then the total number of possible composition paths will be 20^8 (or 25.6 Billion. This constitutes a research problem of how to compose services in a short time such that the resulting composite service presented to the user has optimal QoS level and meets QoS constraints.

QoS-based web service composition problem (SCP) has been defined as an NP-Hard problem [5]. The basis for our research was first presented in [6] where the authors propose an approach to tackling the service composition problem based on Linear Integer Programming (LIP). Their approach makes use of local search techniques in finding services for each function, and then combines these services into a composite service. Currently, Heuristic approaches have been developed to tackle the SCP. One such work is presented in [7] which introduces a heuristic approach based on genetic algorithm to solve the SCP. Their genetic algorithm (GA) encodes composite services as gene and then makes use of evolutionary selection, crossover and mutation operation to search for optimal compositions. Another work [8] employs a different heuristic approach based on particle swarm optimization algorithm (PSO) to solve the SCP. Their solution differs from [7] in the way the algorithm covers the search space in finding optimal compositions. Instead of utilizing genetic operators, PSO encodes compositions as particles which travel in the search space by updating their positions and velocities using the characteristics of the best particle. . In [19] a heuristic based on ant colony algorithm is presented to tackle the SCP where the authors consider multiple QoS attributes in finding composite services that meet user expectations. Many other heuristic approaches such as in [9] [10] [17] have been developed. While the service composition problem can be tackled optimally using meta-heuristics such as genetic, particle swarm and ant colony optimization algorithms, we adopt fruit fly optimization (FOA) algorithm to the SCP. To the best of our knowledge FOA has not been applied to the SCP before.

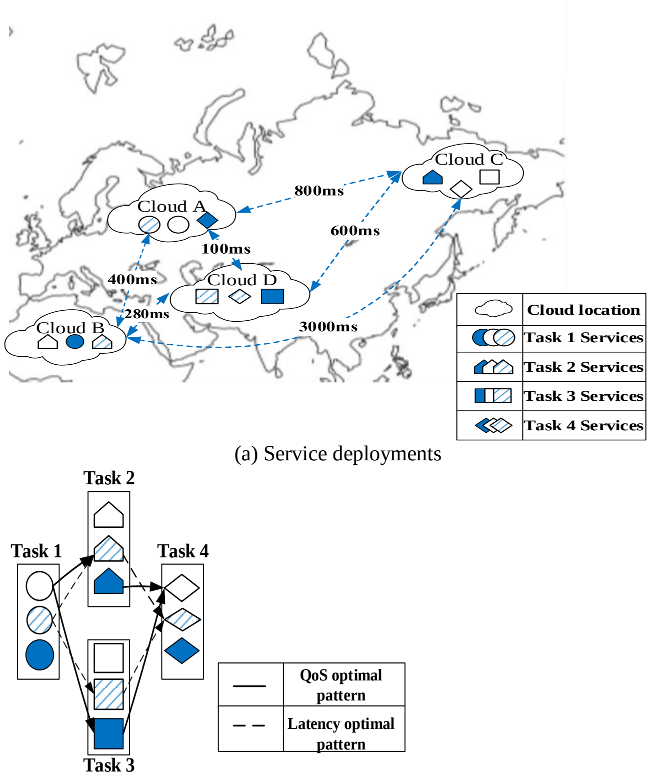
B. *Service composition in the cloud*

Cloud computing provides a platform for enterprises (Service providers) to deploy web services on cloud data centers so that internet users can access service functionalities. Generally, service composition in the cloud takes place across three layers: the infrastructure layer, platform layer and software layer. Once a user request is made, the request is sent to web services which are deployed on the software layer and offered by Software-as-a-Service (SaaS) providers. Underneath the software layer, service registries and composition engine are deployed on the platform layer. The service registry functions as a general repository for storing web services while the composition engine is responsible for aggregating web service into composite services. The infrastructure layer houses the datacenters necessary to support both platform and software layers. Several web services exist on the cloud. For example companies like Amazon and Microsoft provide cloud-based infrastructure and storage services via Amazon Web Services (AWS) and Windows Azure platforms respectively. Within the cloud, users can access different web services from literally any part of the world. Currently, there exist a large number of cloud-based data centers located across the globe. This has exponentially increased the number of possible network paths that can be formed in the search space. Another important issue in the cloud is the need for composite services to meet the QoS guarantees specified in the SLA between services providers and users. This will allow service providers to maximize their earnings while ensuring that user experiences of their services is optimized. Therefore QoS-aware web service composition is critical to the delivery of quality cloud-based composite services on to customers.

When cloud-based web services from different locations are aggregated into a composite service, QoS of the network cannot be ignored. In situations where web services participating in a composition process are small in number, QoS of the network may not significantly affect the performance of a composite service. This is not the case when composition is taking place between large numbers of web services. QoS of the network usually defines the quality of network path between web service data centers. The quality of network paths can be measured using different network QoS metrics such as network latency or round-trip time (RTT) [3], network bandwidth [27], packet loss [25], jitter [26], etc. However, network latency is mainly used to represent QoS of the network because RTT measurements are easier to obtain than other metrics. In addition, occurrences of other metrics are relatively rare. In the cloud, the RTT of network path between web services is usually obtained using geographical location information already stored within each service’s data center. Also, the RTT between a web service’s data center and a user is obtained simply by measuring the RTT of the network path between the user and data center. RTT defines the time it takes for packets of data to move from a source to a destination and then back to the source.

Ideally, network latency is accounted for in the service provider’s service level agreement (SLA) [16] as part of response time QoS attribute. However, this representation can greatly differ from the true network latency that services are physically experiencing. As such, this may lead to sub optimal performance of a composite service from the user’s perspective even if it has been advertised in the SLA as having an optimal response time. Therefore network latency is important in determining the realistic network performance of a composite service in the cloud. To further illustrate this point, [4] claims that a network latency of 20ms can lead to a 15 percent decrease in Google cloud service response times. Similarly, 500ms latency can negatively impact the performance of Amazon web services. Considering another example depicted in Fig. 1 in which services are deployed on different cloud locations. Assuming the network latency between the locations are the values illustrated in Fig. 1a. Current service composition approaches do not separate web service QoS from QoS of the network. Therefore, they will ordinarily search for a solution having optimal service QoS. We assume this optimal solution is depicted by the path with a solid line in Fig. 1b. The end-to-end network latency between cloud data centers in the optimal solution is 3780ms (i.e. 400ms + 280ms + 3000ms). In comparison, executing services in the path of the dashed lines will result in a lower end-to-end latency of 780ms (400ms + 280ms + 0ms). Hence, it is expected that the composite service indicated in dashed lines will have a more positive impact on the user’s experience than the one indicated by the solid lines as long as service QoS constraints are not compromised in the process. This constitutes another motivation for our work as most recent works are incapable in finding optimal latency compositions without compromising web service QoS as defined by the cloud service providers.

Few studies have investigated impact QoS of the network on performance of service composition in the cloud. One such study is proposed in [11] where the authors develop a network-aware genetic algorithm that automatically optimizes compositions in the cloud. In their work they make use of a locality-sensitive hashing scheme coupled with a generic network coordinate system to find services that are close to certain network locations on the cloud. A similar study [18] employ an enhanced genetic algorithm which leverages KD-trees to search for services constituting low latency network paths within the cloud. Another approach in [23] presents a genetic algorithm that tackles service composition in a cloud-based geo-distributed network. In [24] an Ant colony optimization approach to service composition in cloud is proposed. Their approach makes use of greedy search coupled with ant colony algorithm to find minimum number of clouds that will partake in successful service composition. Our previous work in [1] introduces two network-aware composition algorithms for multi-objective QoS optimization in the cloud. One is based on a GA that utilizes k-means cluster to perform mutation of composite services by replacing them with other services in closer network locations. The second algorithm is based on multi-population PSO. PSO separates the particles into two populations. One population operates on web service QoS search space while the other population operates on network QoS search space. Particles from both populations then combine their best characteristics to form new particles.



(a) Service deployments

(b) QoS-optimal vs Latency-optimal patterns

Fig. 1. Service deployments on the cloud

Both algorithms were fed by state-of-the-art network coordinate system to determine network distance estimation between services in the cloud. Also, both approaches search for low latency compositions while adhering to strict user QoS constraints. However they make use of a similar network model which decomposes known latency measurements into coordinates prior to estimation process. The model then converts coordinates back to latency representation after estimation process and before they are fed to GA and PSO.

Despite the development of several meta-heuristic algorithms for solving SCP on the cloud, most of these techniques are not naturally compatible with network QoS metrics. For example previous algorithms [1, 11, 18] had to use special structures and computations which allow them to work with network models and perform latency-centric QoS optimization. This approach adds to the complexity and computation cost of the algorithms presented. It is therefore necessary to discover algorithms that are better suited to working with network latency while performing QoS optimization. To this end we present an enhanced fruit fly algorithm known as NFOA to search for low latency compositions with near optimal QoS. As a new meta-heuristic optimization algorithm, FOA is inspired by the behaviour of fruit flies in searching for food. FOA is easy to implement and consists of few adjustable parameters. Due to these merits, FOA has been successfully used in solving several NP-Hard optimization problems such as neural network optimization [21], financial distress [12] and more recently in scheduling problems [22]. A core characteristic of fruit fly optimization algorithm is its ability to encode solutions in form of two-dimensional network coordinates [12]. This property sets FOA apart from other meta-heuristic algorithms because it allows FOA to seamlessly work with network QoS metrics that are correlated to network coordinates such as network latency. The network coordinates employed by NFOA are obtained from our network model.

In summary, we propose an approach to network-aware service composition in the cloud. Our approach consists of the following contributions:

i. We adopt a network model that is supplemented by a state of the art network coordinate system. The model will be used to predict the network positions of web service data centers in the cloud.

ii. We present an enhanced fruit fly algorithm by adding capability to find services whose network positions are closer to each other and to the users while ensuring QoS is optimized. Consequently these services will result in low latency compositions. We compare our algorithm against current approaches. The results of our experiment demonstrate that our algorithm is competitive when compared to recent approaches in finding near optimal compositions.

The remainder of the paper is organized as follows. Section 2 formulates the service composition problem. Section 3 presents our network model and fruit fly algorithm for network-aware web service composition. Section 4 discusses the evaluation results of our approach. Section 5 concludes this paper.

II. Problem Formulation

Service composition forms its basis from workflow management systems [2] where a complex user request is exposed as set of tasks that require one or more services to be completed. Hence the following definitions are used in this paper:

Definition 1 (*Service*). Service is a single unit meant for solving a particular functionality or task that is part of a user request. Services are published in the cloud by the service provider.

Definition 2 (*Service class*). Service class is a group of services having similar functionality but different QoS levels.

Definition 3 (*Candidate service*). Candidate service is a service that is part of a service class.

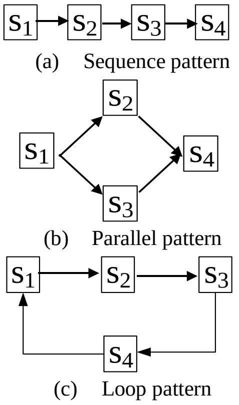
Definition 4 (*QoS attribute*). QoS attribute defines a given quality aspect of service. Some popular QoS attributes for services include cost, response time, reputation, reliability, etc.

TABLE I. Examples of Service QoS Attributes

QoS Attribute	Description
Cost	Amount payable in monetary value for the execution of service.
Reputation	Users' average rank of a service based on their experiences
Response time	Time it takes to process a user request from the point it is made up till the point it is received.

Definition 5 (*Network Latency*). Defined as RTT from one source data center to another and then back to the source data center. In the case of a composite service, network latency is defined as end-to-end RTT from the first service's data center in a given composite service to the last service's data center then back to the first service.

Definition 6 (*Workflow pattern*). Workflow pattern dictates the direction in which data flows from one service to another within a composite service. Some major workflow patterns include sequence, parallel, exclusive choice and loop.



(a) Sequence pattern

(b) Parallel pattern

(c) Loop pattern

Fig. 2. Workflow patterns for services s_1, s_2, s_3, and s_4

TABLE II. Types of Workflow Patterns

Workflow pattern	Synonym	Description
Sequence	Sequential routing	Executes a set of services sequentially
Parallel	AND-split	Executes a set of services simultaneously

In Table 3, *h* represents total number of executions. Formulas presented in the table are used to obtain end-to-end values for QoS attributes. For instance, if our workflow pattern is the example show in Figure 3.

TABLE III. Aggregation Formulas for End-to-End QoS of Composite Service

QoS attribute	Sequence pattern	Parallel pattern
Response time	$\sum_{i=1}^{n} RT(S_i)$	$Min\left(RT(S_1),..,RT(S_n)\right)$ *OR* $Max\left(RT(S_1),..,RT(S_n)\right)$
Reputation	$\frac{\sum_{i=1}^{n} RP(S_i)}{n}$	$Max\left(RP(S_1),..,RP(S_n)\right)$
Cost	$\sum_{i=1}^{n} C(S_i)$	$\sum_{i=1}^{n} C(S_i)$

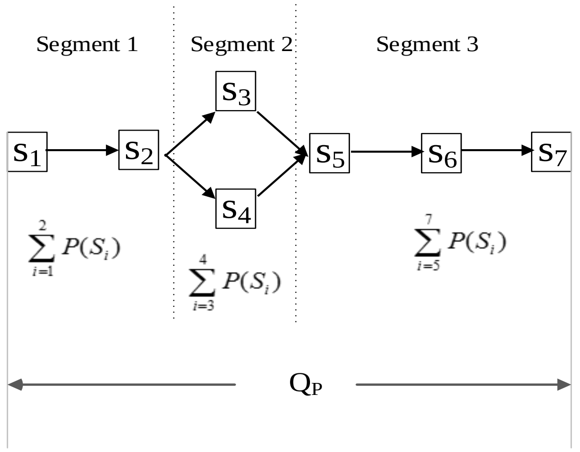


Fig. 3. Example workflow pattern

End-to-end cost (Q_P) is computed by adding cost for services in each segment of the workflow pattern. End-to-end response time on the other hand is computed by adding service response times in segments 1 and 3 to service with maximum response time in segment 2.

Definition 7 (*Service composition problem*). The Service composition problem is defined as follows:

Given a set of *n* interconnected tasks that are needed to satisfy a user requirement,

$$T = \{t_1, t_2, \ldots, t_n\}$$

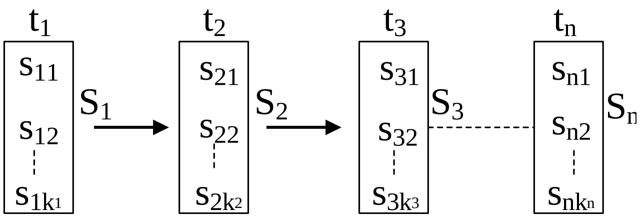


Fig. 4. Arrangement of candidate services into tasks

Each *i*-th task requires k_i number of similar services (candidate services) that have the ability to complete the task (as seen in Figure 4),

$$S_i = \left\{s_{i1}, s_{i2}, \ldots, s_{ik_i}\right\},\ \forall i \in [1..n]$$

Where *i* identifies the service class in which similar services are grouped according to their task. Our service composition problem assumes that only one candidate service is selected per service class and bound to a task. We also assume that each service class has the same number of candidate services. Once all tasks have been bound a composite service *C* is formed,

$$C = \left\{s_{1j}, s_{2j}, \ldots, s_{3j}\right\},\ \forall j \in [1..k]$$

Each service is advertised with its own QoS level in the SLA. In this study we consider the cost QoS attribute. As Table 3 shows, the end-to-end cost for a composite service is computed by aggregating individual cost for each service that forms part of the composite service. Thus,

$$Q_P(C) = \sum_{i=1}^{n} P(s_{ij}) \qquad (1)$$

Where Q_P represent end-to-end QoS value for composition cost. Also *P* represents candidate service QoS value for cost.

Cost is normalized in the range [0 1] using (2). Where f_p is normalized cost, $Max_p(P(S_i))$ and $Min_p(P(S_i))$ represent maximum and minimum QoS values for service class *i* respectively.

$$f_p(C) = \sum_{i=1}^{n}\left(\frac{Max_p(P(S_i))\ -\ Q_p(P(s_{ij}))}{Max_p(P(S_i)) - Min_p(P(S_i))}\right) \qquad (2)$$

With respect to the QoS of the network, we assume that each web service is deployed on its own cloud data center for the sake of simplicity. Then end-to-end network latency for a composite service is defined as a vector of network coordinates (E).

$$E(C) = \left\{\left[x_{1j}, y_{1j}\right], \left[x_{2j}, y_{2j}\right], \ldots, \left[x_{nj}, y_{nj}\right]\right\}$$
$$\forall j \in [1..k]$$

Where $[x, y]$ is the network coordinate of a service in the Cloud.

The values of $[x, y]$ coordinates are obtained from the estimation of RTT by our network model.

Each service that is part of a composite service is represented by two dimensional network position as seen in Figure 5. Where x_{ij} and y_{ij} are x-axis and y-axis coordinates of a service s_{ij}.

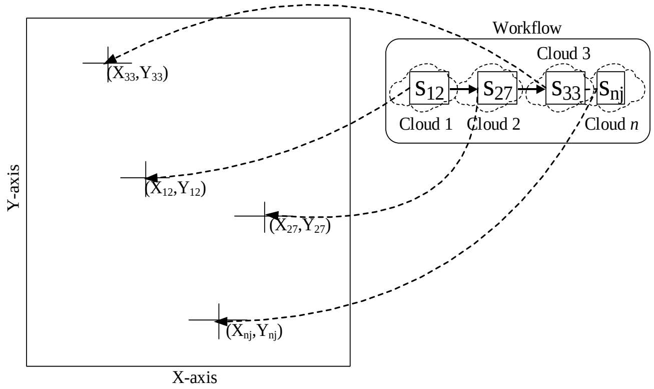


Fig. 5. Services and their network positions

We model our optimization problem as a single objective optimization problem where the goal is to optimize fitness value (F);

$$F = function(f_p, E) \quad (3)$$

Hence our service composition problem is to find a composite service that has optimal cost and near optimal network latency between constituent service network paths in terms of their network positions (E). Ideally this composite service will have selected a set of services deployed on Cloud locations that have the shortest end-to-end RTT without compromising cost QoS.

III. Network-Aware Service Composition Algorithm

A. Basic concept of fruit fly algorithm

The fruit fly optimization algorithm is a new type of evolutionary algorithm proposed in 2011. The algorithm mimics the behaviour of a fruit fly when it is searching for food as shown in Figure 6. A fruit fly is characterized by its acute sensing and perception abilities. This is said to be as a result of its osphresis organs [12]. Via the organs, a fruit fly is able to perceive food particles from several kilometres away. Once a fruit fly smells the presence of food, it closes in on the direction of the food in a hoping fashion. Each time the fly hops to a possible location, it tries to determine the next hoping direction that will take it to closer to the food source. Based on the behaviour exhibited by the fruit fly. We describe the steps required by the fruit fly optimization algorithm.

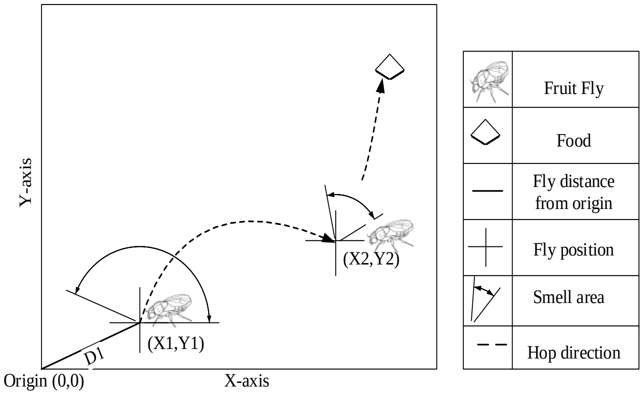


Fig. 6. Food searching pattern of fruit fly

1) Initialize population

X and *Y* axes ($\bar{x}$, $\bar{y}$) for a fruit fly swarm are first initialized*;*

$$\bar{x} = Init(X_{axis}) \quad (4)$$

$$\bar{y} = Init(Y_{axis})$$

Then individual positional coordinates of each fruit fly is initialized. For a fruit fly *i*,

$$x_i = \bar{x} + rand() \quad (5)$$

$$y_i = \bar{y} + rand()$$

2) Estimate Distance and Smell concentration judgment value

Given that the exact position of the food is initially unknown, each fruit fly computes its distance (*g*) from origin (0,0) using (6), then the smell concentration judgment value (*v*) for every fruit fly is computed as the inverse of distance.

$$g_i = \sqrt{x_i^{\,2} + y_i^{\,2}} \quad (6)$$

$$v_i = \frac{1}{g_i} \quad (7)$$

3) Determine fitness value

The fitness value, also known as Smell concentration judgment function, is calculated as a function of smell concentration value (*g*).

$$F_i = function(v_i) \quad (8)$$

4) Determine best fruit fly

Compare fitness values of all fruit flies in swarm and determine fruit fly with the best fitness value.

$$[best_F \quad best_{index}] = \max(F) \quad (9)$$

5) Store attributes of best fruit fly

In order to compare fitness of best fruit fly against other fitness values subsequent iterations, the best fitness is stored in memory,

$$Fit_{best} = best_F \quad (10)$$

Then the positions of the best fruit fly are stored as new X and Y axes for the fruit fly swarm,

$$\bar{x} = X(best_{index}) \quad (11)$$

$$\bar{y} = Y(best_{index})$$

Best positions are used to update each fruit fly in the swarm according to equation (5). In the next iteration, steps 2 to 5 are repeated until either the maximum number of iterations is reached, or optimization is achieved.

B. Fruit fly algorithm for network-aware service composition

We proposed an enhanced network-aware fruit fly optimization algorithm called NFOA. Before we discuss our algorithm, we introduce the network model that feeds the algorithm with the network positions of web services that will take part in the composition process.

1) Network model

Traditionally, RTTs of network paths between a number of data centers are measured by physically sending ping packets between the data centers. This can be both time and resource intensive. For instance, given *T* number of interconnected tasks within a composite service (as seen in Figure 7), there exists *O* (n^2) network paths that can be formed between them. Some research has been done to discover more efficient techniques for determining RTT between Internet nodes. Some of the techniques are based on Euclidean distance models (EDM) [29, 30] while others are based on matrix factorization models (MF) [31, 32].

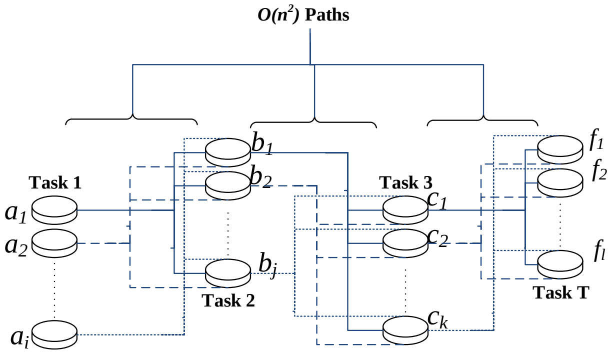


Fig. 7. Network of *n* web service nodes and *O* (n^2) paths for a sequence of *T* sub-tasks in a workflow

EDM employs central landmark servers which are responsible for making direct RTT pings and measuring path latencies. MF on the other hand employ a more accurate and decentralized method which allows each Internet node to estimate its path RTTs.

We adopt a network model based on MF that efficiently estimates network latency between services in a cloud network. The model consists of a state of the art MF-based network coordinate system [31] that predicts RTT between web services on the cloud. The adopted network coordinate system works by only measuring RTT from each service location to a small subset of *k* neighbouring service locations on the cloud e.g. from S_{11}'s data center (Cloud 1) to S_{21}'s data center (Cloud 2) in Figure 8. The measurements are then used to estimate un-measured RTT to other locations in the form of network positions e.g. from S_{11}'s data center (Cloud 1) to S_{23}'s data center (Cloud 3). In mathematical terms, MF finds estimates of row matrix *X* and transposed column matrix *Y* that minimizes estimation error (ε) which is the difference between measured RTT values and predicted RTT values. The *X* and *Y* matrices represent two dimensional network positions of services in the cloud.

$$\varepsilon = D-(X*Y^T) \quad (12)$$

Where $D^{j\cdot}$ defines an RTT matrix of both know and unknown measurements, while ($X*Y^T$) defines predicted RTT in the form of network positions. Once RTT between all service locations have been determined, *X* and *Y* network positions are fed to our fruit fly algorithm to find low latency and QoS optimal compositions. The MF algorithm for RTT estimation is outlined below.

Algorithm 1 MF Algorithm

```
Input:  D, max_iter, k
Ouput:  Dnew
[X, Y] = function MF(D)
{       for(i =1: maxIter)
              for(j =1: maxCS)
                  $X \leftarrow$ rand(x)
                     $Y \leftarrow$ rand(y)
                      $\varepsilon \leftarrow w\,[D-(X*Y^T)]^2$
                  if ( $\varepsilon$ is minimised)
                         $Dnew \leftarrow X*Y^T$
                          return
                      endif
              endfor
        endfor
}
```

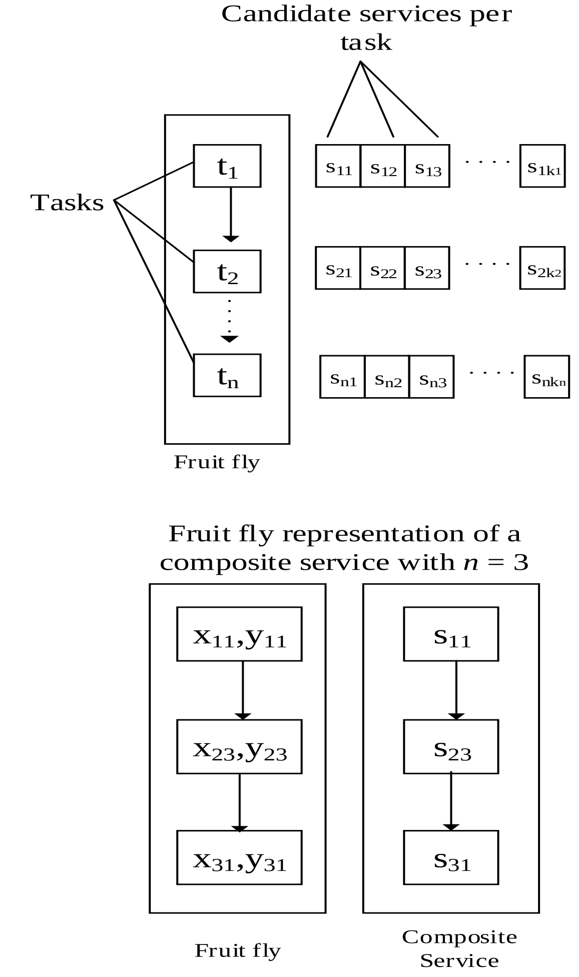


Fig. 9. Encoding a composite service as a fruit fly using NFOA

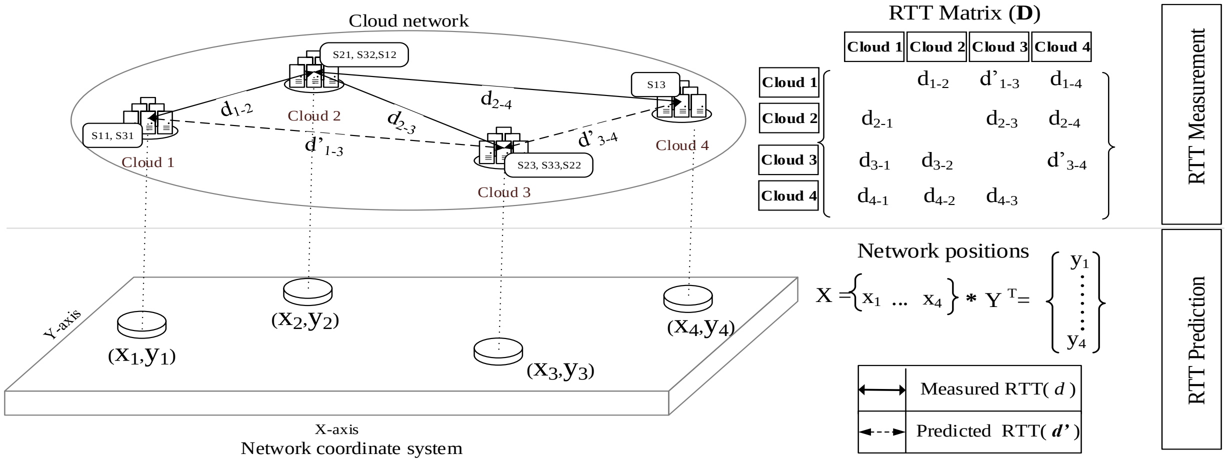


Fig. 8. RTT estimation process

2) *NFOA Algorithm*

In this section we present network-aware fruit fly algorithm to tackle our service composition problem.

a) Initialize population

Firstly, each fruit fly in the swarm is initialized as a possible composite service. In this case, a fruit fly is encoded as a set of service coordinates where each service coordinate represents the network position of service within the cloud as seen in Figure 9.

b) Determine end-to-end Vector of network coordinates and Cost QoS

Instead of randomly assigning coordinates to each service (as seen in step 1 of basic fruit fly algorithm), NFOA assigns network coordinates fed by our network model to each service. These coordinates will be a representation of the RTT between each service location in the cloud. Hence,

$$x_i = X \qquad (13)$$
$$y_i = Y$$

Where X and Y represent network coordinates for a service obtained from MF algorithm.

Using this procedure, a vector of network positions (E) is obtained for each fruit fly by aggregating the network positions of each service within the fruit fly. Then each fruit fly determines its end-to end smell concentration value (G) by combining individual smell concentration values (g) for all n services in a fruity fly.

$$G = \sum_{i=1}^{n} g_i \quad (14)$$

The next step involves determination of end-to-end cost (f_p) by aggregating individual service QoS levels according to Equation (2).

c) Estimation of end-to-end smell concentration judgment function

Smell concentration judgment function is estimated for each service in a fruit fly (according to (7)) and then combined into end-to-end smell concentration judgment function for the composite service.

$$V = \sum_{i=1}^{n} v_i \quad (15)$$

d) Computation of fitness value

Both end-to-end smell concentration judgment function and end-to-end cost are used to compute the fitness value (F) for a fruit fly thus;

$$F = \frac{V}{f_p} \quad (16)$$

The last step involves storing the fruit fly with best fitness and then updating the coordinates of the each fly in the population with that of the best fly. The process is repeated until maximum number of iteration is reached. Below outlines our NFOA algorithm.

Algorithm 2 NFOA Algorithm

```
Input: T, C, O, maxgen, pop_size, D
Ouput: bestFly
Randomly generate fruit fly positions ← pop
pop ← MF (D)
while (gen ≠ maxgen)
    {
        G ← Dist (pop)
        V ← Smell (pop)
        $f_p$ ← Smell_Function (pop)
        F← V/ $f_p$
        bestFly ← pop[min (F)]
        pop ←bestFly + rand()
        endwhile
    }
```

IV. Experimental Results

Evaluations were run on a PC with Intel Core i7 processor with 2.8 GHZ CPU and 8GB RAM. Our algorithm and simulations were done on MATLAB 2014 environment. Meridian RTT dataset [20] was used to simulate a network of 650 unique data centers spread out in the cloud. Each location represents a web service position on the cloud. For the sake of simplicity, a sequence workflow pattern of 13 tasks and 20 candidate services per task is considered. This pattern considered is meant to simulate a realistically large service environment. Also, a single user location is considered in our cloud network. In our simulation, we consider the cost QoS attribute, although any other QoS attribute could be considered as this will not affect our experiments. Cost QoS values for every service is generated randomly with a Gaussian distribution within the range [1, 40].

NFOA algorithm is compared against state of the art service composition methods based on Genetic Algorithm (GA) [5] and Particle swarm optimization algorithm (PSO) [21]. Both GA and PSO are fed by our network model in order to estimate RTT of their solutions for the sake of comparing their optimality against NFOA. Table 4 presents the environment settings for our test algorithms.

TABLE IV. Algorithm Settings

Parameters	GA	PSO	NFOA	MF
Population size	200	200	200	260
Number of generation	200	200	200	50
Crossover probability	0.9	-	-	-
Mutation probability	0.5	-	-	-
Tour size	2	-	-	-
Network model	-MF	MF	MF	-
Distribution index	20	-	-	-
Crossover operator	Single crossover	-	-	-
Mutation operator	Standard mutation	-	-	-

Number of Tasks	13	13	13	13
Number of Candidate services	20	20	20	20
Number of neighbours that measure RTT with each service				5

1) *Fitness*

We run our test algorithms over 200 generations. From Figure 10, we discover that the fitness value for NFOA converges after 100 generations. Also NFOA finds solution with the best fitness among the three test algorithms. This shows that NFOA's natural ability to work with network coordinates makes it an ideal choice in searching for solutions with low latency and optimal QoS. The Figure also shows that NFOA has the ability to find a global solution and avoid being trapped in local optimum. This is attributed to the update strategy employed by NFOA which ensures that updates to *x* and *y* network coordinates are widely distributed across the network coordinate search space. Table 5 shows the best fitness values obtained over five runs. The result demonstrates that NFOA obtained the best fitness in three of the five runs as highlighted by the bold values.

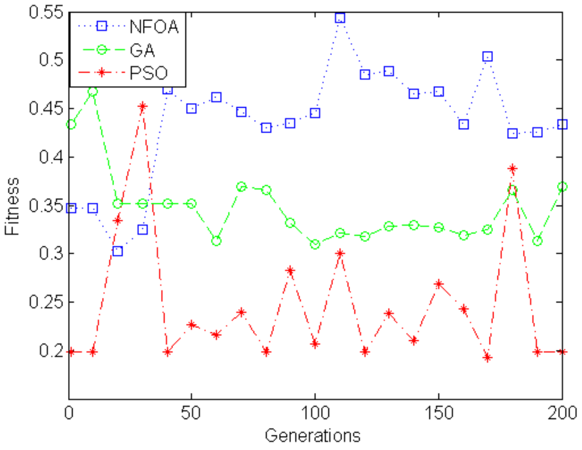


Fig. 10. Fitness versus Generation

TABLE V. EXPERIMENTAL RESULT FOR FITNESS

Runs	NFOA.	GA	PSO
1	0.3467	**0.4341**	0.3273
2	0.4502	0.3520	0.2264
3	**0.4458**	0.3099	0.2065
4	**0.4679**	0.45207	0.2686
5	**0.4336**	0.3690	0.3966

2) *Network latency*

In this experiment, we evaluate the network latency (RTT) solutions for each generation. Typically, the best algorithm will indicate the lowest RTT. From Figure 11, it is observed that the RTT converged at100-th iteration for NFOA which represents the best RTT while it converges at much higher values for GA and PSO. This further demonstrates NFOA's superiority to other algorithms to in finding low latency solutions.

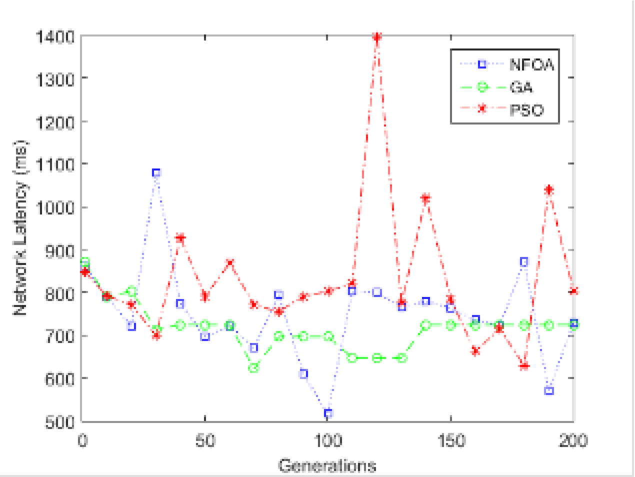
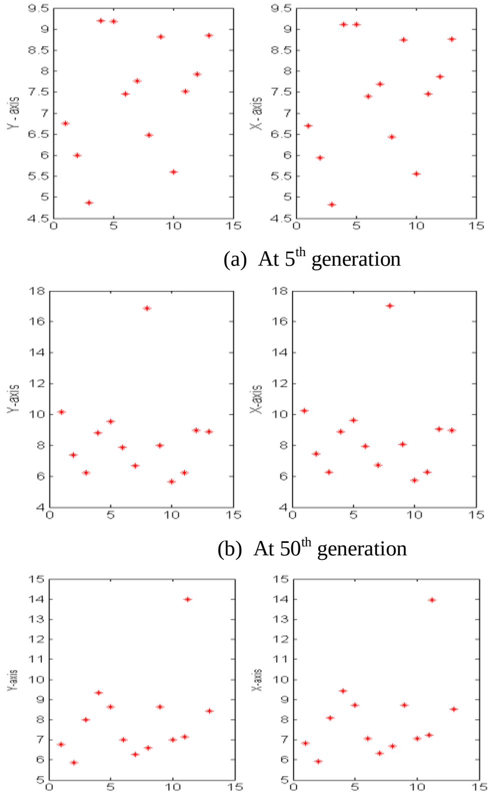


Fig. 11. Network latency versus Iterations

3) *Best Fruit fly path*

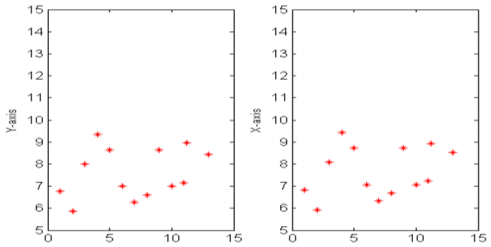
Figure 12 shows the best fruit fly's path to optimization. Each point on the plot reflects the network positions of each web service that forms part of the best composite service. The graph demonstrates that, upon reaching the 200-th generation, QoS-optimal services that have shorter RTT from each other are constituted into the best fruit fly.



(a) At 5th generation

(b) At 50th generation

(c) At 100th generation



(d) At 200th generation

Fig. 12. Path of the best fruit fly

4) *Computation time*

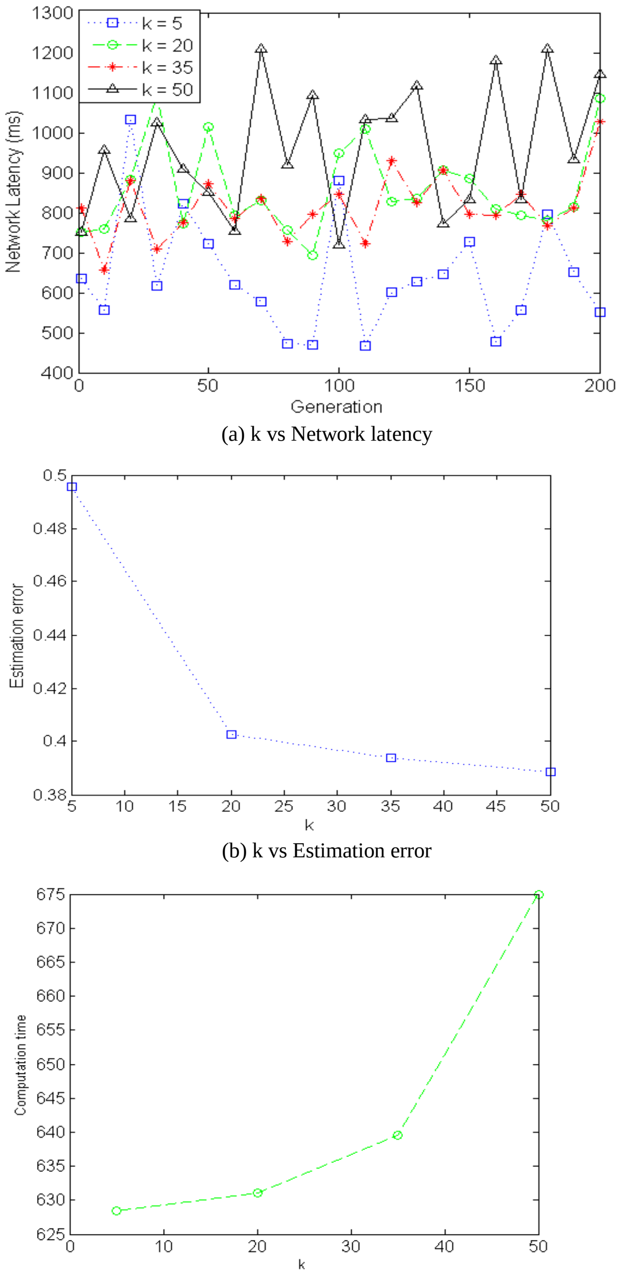
As for computational efficiency, Table 6 shows that NFOA has the fastest average computation time when compared to GA and PSO. This is because since NFOA is already naturally built to handle optimization using network coordinates, it does not require additional structures and computations to work with our network model. This is not the case with PSO and GA which require additional computations that further worsen their execution times.

TABLE VI. AVERAGE COMPUTATION TIMES (IN SECONDS) OF THE FOUR ALGORITHMS

NFOA.	PSO	GA
48.42s	60.055s	109.24s

5) *Number of RTT-measured neighbours*

This experiment evaluates the impact of number of RTT-measured neighbours (*k*) on estimation error (ε), computation time and quality of NFOA's solutions. The estimation error will give us an idea of how accurate our compositions' predicted RTTs are compared to their actual RTTs. In this experiment, we vary the value of *k* between 5 and 50 neighbours per service. In Figure 13(a), it is observed that as *k* is increased (i.e. the more neighbours each service measures its RTT to) the higher the latency value of the compositions. The reason for this effect can be seen from Figure 13 (b) which shows the variation of estimation error (ε) with *k*. When the value of *k* is set to 5, it means that each service will measure RTT with small number (5) of its neighbours and then predict RTT with all the other services. This will ultimately reduce the prediction accuracy (i.e. increase the estimation error) for each composite service. On the other hand, setting *k* to 50 means increasing the number of the measured RTT paths to 50. This will lead to a higher prediction accuracy (i.e. lower estimation error) for each composite service. This result means that even if composite service latencies are lower when *k* is set at values below 20, they are the least accurate representations of true network latency of the compositions when compared to values above 20.



(a) k vs Network latency

(b) k vs Estimation error

(c) k vs Computation time

Fig. 13. Effect of *k* on network latency, estimation error, and computation time

Figure 13 (c) shows the linear variation between *k* and Computation time. If *k* is too high then computation time for NFOA algorithm will increase and vice versa. Based on these observations, the best setting for *k* should be between 20 and 35.

V. Conclusion

In this paper we propose an enhanced fruit fly optimization algorithm called NFOA that performs network-aware web service composition in the cloud. Fruit fly optimization a new approach for finding best solutions by mimicking the behaviour of the fruit fly. The number of services distributed on the Cloud has increased. Therefore the QoS of network has become important in determining performance of a composite service. We define a network model that estimates network latency in the form of service network positions with the aid of a network coordinate system based on matrix factorization called MF. MF measures RTT between a service and a small number of its neighbours then estimates the unknown RTT with other services in the cloud. MF feeds network positions of services to NFOA which uses them directly to find composite services with low latency and near-optimum web service QoS. Experimental simulations have shown that NFOA is superior to other meta-heuristic techniques in finding solutions with optimum fitness and latency.

References

[1] U. Shehu; G. Ali Safdar; G. Epiphaniou; “Network-aware Composition for Internet of Thing Services*” in Transactions on Networks and Communications* vol.3, no.1, pp 45-58 February 2015

[2] Jaeger, M.C.; Rojec-Goldmann, G.; Muhl, G., "QoS aggregation for Web service composition using workflow patterns," *Enterprise Distributed Object Computing Conference, 2004. EDOC 2004. Proceedings. Eighth IEEE International* , vol., no., pp.149,159, 20-24 Sept. 2004

[3] Rony Kay; “Pragmatic Network Latency Engineering Fundamental Facts and Analysis,” *cPacket Networks on* vol., no., pp.1-13, 2009

[4] http://www.wired.com/2012/09/layers-of-latency/

[5] G. Canfora, M. D. Penta, R. Esposito, and M. L. Villani. An approach for QoS-aware service composition based on genetic algorithms. *In GECCO ’05: Proceedings of the 2005 conference on Genetic and evolutionary computation*, vol., no., pp 1069–1075, New York, NY, USA, 2005. ACM

[6] L. Zeng; B. Benatallah; M. Dumas; J. Kalagnanam; Q. Z. Sheng;, ”Quality Driven Web Services Composition,” *In WWW '03: Proceedings of the 12th international conference on World Wide Web*, vol., no., pp., 2003.

[7] Yilmaz, A.E.; Karagoz, P.; "Improved Genetic Algorithm Based Approach for QoS Aware Web Service Composition," *Web Services (ICWS), 2014 IEEE International Conference on* , vol., no., pp.463,470, 2014

[8] W. Yang; C. Zhang; ”A Hybrid Particle Swarm Optimization Algorithm for Service Selection Problem in the Cloud*” International Journal of Grid Distribution Computing,* vol.7, no.4, pp.1-10, 2014

[9] A. Sawczuk da Silva; H. Ma; M. Zhang; “A GP Approach to QoS-Aware Web Service Composition and Selection” *in Springer Simulated Evolution and Learning* vol.8886, no., pp.180-191 2014

[10] X. Wu; T. Wang; X. Qian; C. Zeng; “Multi-QoS aware automatic service composition” *in Springer Wuhan University Journal of Natural Sciences,* vol.19, no.4, pp. 307-314, August 2014

[11] Adrian, K.; Fuyuki I.; Shinichi Honiden,"Towards network-aware service composition in the cloud," In *Proceedings of the 21st international conference on World Wide Web* (WWW '12). ACM, New York, NY, USA, on, vol., no., pp.959-968, 2012.

[12] Wen-Tsao Pan; “A new Fruit Fly Optimization Algorithm: Taking The Financial Distress Model As An Example” *In Elsevier Knowledge-Based Systems* vol 26 no. pp.69-74 2012

[13] J. O’Sullivan; D. Edmond; A. T. Hofstede; “What’s in a service?” *In Distrib. Parallel Databases*, vol. 12, nos. 2–3, pp. 117–133, 2002

[14] L. Wengin; “Towards a Resilient Service-Oriented Computing based on Ad-hoc web Service Compositions in Dynamic Environments”, *INSA Lyon*, vol., no., pp.4-5, March 2014

[15] U.Shehu; G. Epiphaniou; G. Safdar;”A Survey of QoS-Aware Web Service Composition Techniques”, *In International Journal of Computer Applications v*ol.89, no.12, march 2014

[16] Landi, G.; Metsch, T.; Neves, P.M.; Mueller, J.; Edmonds, A.; Secondo Crosta, P., "SLA Management And Service Composition of Virtualized Applications In Mobile Networking Environments," *Network In Operations and Management Symposium (NOMS) IEEE* , vol., no., pp.1,8, 5-9 May 2014

[17] A. Younes; M. Essaaidi; A. Moussaoui;”SFL Algorithm for QoS-based Cloud Service Composition”, *In International Journal of Computer Applications*, vol.97, no.17, pp.42-49, July 2014

[18] Klein, A.; Ishikawa, F.; Honiden, S., "SanGA: A Self-Adaptive Network-Aware Approach to Service Composition," in *Services Computing, IEEE Transactions on* , vol.7, no.3, pp.452-464, July-Sept. 2014

[19] Hui Liu; Dong Xu; Huaikou Miao, "Ant Colony Optimization Based Service Flow Scheduling with Various QoS Requirements in Cloud Computing," in *Software and Network Engineering (SSNE), 2011 First ACIS International Symposium on* , vol., no., pp.53-58, 19-20 Dec. 2011

[20] Wong, B.; Slivkins, A.; Sirer, E.; “Meridian: A lightweight network location service without virtual coordinates,” *In: Proc. the ACM SIGCOMM.*, vol., no., pp., 2005

[21] Ludwig, S.A., "Applying Particle Swarm Optimization to Quality-of-Service-Driven Web Service Composition," *Advanced Information Networking and Applications (AINA), 2012 IEEE 26th International Conference on* , vol., no., pp.613,620, 26-29 March 2012

[22] W.T. Pan; “Using Modified Fruit Fly Optimization Algorithm To Perform The Function Test And Case Studies,” *Connect. Sci.,* vol.25, no., pp. 151–160, 2013

[23] X. Zheng; L. Wang; S. Wang;”A Novel Fruit Fly Optimization Algorithm For The Semiconductor Final Testing Scheduling problem” Tsinghua National Laboratory for Information Science and Technology, vol.57, no., pp.95-103, 2014

[24] D. Wang; Y. Yang; Z. Mi;”A Genetic-based Approach to Web Service Composition in Geo-distributed Cloud Environment,” In Elsevier Journal of Computers and Electrical Engineering, vol., no.,pp.1-12, 2014

[25] Q. Yu; L. Chen; B. Li;”Ant Colony Optimization Applied to Web Service Compositions in Cloud Computing,” In Elsevier Journal of Computers and Electrical Engineering, vol.41, no.,pp.18-27, 2015

[26] Guohui Wang; Ng, T.S.E., "The Impact of Virtualization on Network Performance of Amazon EC2 Data Center," *INFOCOM, 2010 Proceedings IEEE* , vol., no., pp.1,9, 14-19 March 2010.

[27] Kyoung Shin Park; Kenyon, R.V., "Effects of network characteristics on human performance in a collaborative virtual environment," *Virtual Reality, 1999. Proceedings., IEEE* , vol., no., pp.104,111, 13-17 Mar 1999

[28] Cong Ding; Yang Chen; Tianyin Xu; Xiaoming Fu, "CloudGPS: A scalable and ISP-friendly server selection scheme in cloud computing environments," *Quality of Service (IWQoS), 2012 IEEE 20th International Workshop on* , vol., no., pp.1,9, 4-5 June 2012.

[29] Hyuk Lim, Jennifer C. Hou, Chong-Ho Choi; “Constructing internet coordinate system based on delay measurement,” *IEEE/ACM Transactions on Networking*, vol.13, no.3, pp.513-525, 2005

[30] Hyuk Lim, Jennifer C. Hou, Chong-Ho Choi; “Constructing internet coordinate system based on delay measurement,” *IEEE/ACM Transactions on Networking*, vol.13, no.3, pp.513-525, 2005

[31] Liao, Yongjun. "Learning to predict end-to-end network performance.", PhD Thesis University of Liege Belgium, vol, no., pp.38-43, 2013.

[32] Y. Mao, L. Saul, J. M. Smith, IDES: An Internet Distance Estimation Service for Large Network, IEEE Journal on Selected Areas in Communications (JSAC), vol., no., pp.2273 – 2284, 2006.

Role of Security in Social Networking

David Hiatt
College of Arts & Sciences
Regent University
Virginia Beach, Virginia, U.S.A.

Young B. Choi
College of Arts & Sciences
Regent University
Virginia Beach, Virginia, U.S.A.

***Abstract*—In this paper, the concept of security and privacy in social media, or social networking will be discussed. First, a brief history and the concept of social networking will be introduced. Many of the security risks associated with using social media are presented. Also, the issue of privacy and how it relates to security are described. Based on these discussions, some solutions to improve a user's privacy and security on social networks will be suggested. Our research will help the readers to understand the security and privacy issues for the social network users, and some steps which can be taken by both users and social network organizations to help improve security and privacy.**

Keywords—Security; Information Security; Social Networking: CIA; Confidentiality; Integrity; Availability; PII; Social Networking Service; SNS

I. Introduction

Information security is very important these days to anyone using a computer or to any organization that employs computers and networking in their day to day operations. That is nearly everyone. Information security should be at the forefront of everyone's mind since so much of our personal information is out there on the Internet. [1] states that information security is necessary because of the risk generated when technology is used to process information because information may be disclosed in the wrong way or to the wrong person. Information security is broken up into three major areas, which are called the CIA of information security. These areas are confidentiality, integrity, and availability. Confidentiality deals with making sure only authorized people have access to the information. Integrity deals with making sure that the information is not tampered with or corrupted in any way. And finally, availability is just making sure the information can be accessed and where it is supposed to be. This is about protecting information in storage, transmission, and processing, using policy, education, and technology, according to the McCumber Cube model of information security. Many companies and organizations that are just working with day to day data are taking all precautions to prevent hackers from causing attacks and data breaches, using firewalls, intrusion detection and prevention systems, honeypots, and appropriate training and policy enacted by their security managers.

It's a different ball game when talking about social networks though. Social networking service (SNS) like Facebook are not as secure, despite the technologies implemented at their facilities or the policies put in place by their security personnel. The main reason for this is because of the information that users put on these social networks. According to [2], the staggering popularity of these social networks, which are often used by teenagers and people who do not have privacy or security on their minds, leads to a huge amount of potentially private information being placed on the Internet where others can have access to it. She goes on to say that interacting with people is not new, but this medium for doing it is relatively new. She says, "Social networking sites have become popular sites for youth culture to explore themselves, relationships, and share cultural artifacts. These sites centralize and help coordinate the interpersonal exchanges between American teens and global brands [2]." According to [3], it is very easy to communicate with others using a social network construct. He also says that all the information you post on these sites over the years builds up into a collection of information that becomes known as your profile and nearly anyone online is able to see it, especially your friends. So with the continued prevalence of social networking there is a continued risk to the security of information, but not mainly from hackers or thieves, but from the false trust that many people have when placing private information about themselves online. This is a huge risk but it can be combatted with education. [3] states that Facebook and other sites have become such a part of many of our lives and Internet usage. So compounding the huge repository of personal data online is the careless and over-trusting nature in which people, especially teenagers, share personal information online. This information may not contain actual PII (Personally Identifiable Information), but it does contain many parts that can be aggregated into a whole by an attacker. This information can also be contained in pictures that are posted; for example a picture taken in front of your house may contain the house number. It is easy to see how this can happen when people are not very attentive to their security and the security of their information. It is essential to be careful what we put online in this way; being careless can lead to information being posted that should not be available to others.

II. History and Development of Social Networking

Sharing information and communicating with people has been around for as long as people have been around. But when computers and the Internet became much more common, we saw the use of email systems and short text messages as the first popular means of communication between people [4]. This was not so dangerous because it involved the sending of one message at a time between two people only, and it was no riskier than sending other information across the web to only one person. [4] goes on to point out that more technologies like chat rooms and online games came to be, and then social media where users could share information, talk, discuss interests and likes, post pictures and video, etc. One of the first social networking sites like this was MySpace [2]. Its original audience was teenagers and the music and art scene. She points out that its popularity dropped like a rock when Facebook came online and then it became the most popular social network. Some sites, such as LinkedIn or Flickr, have a specific purpose, and some are more general. There is almost no limit to what people can post online these days, and this is a potentially scary thing. According to [5], social media "spread quickly and widely and contain large-scale information of a broad audience. However, the unstructured massive data transaction may overwhelm users with information overflow" leading to a form of chaos. [2] states that social media sites and chat rooms are basically just "organizational and software procedures that control the exchange of interpersonal information in social networking sites, text messaging, instant messenger programs, bulletin boards, online role-playing games, computer-supported collaborative work (CSCW), and online education." As mentioned in [4], sites that provide these types of services to users at little to no cost provide a lot of enticement for people and have become very popular.

[6] tells us in his article that there is a rich history for social media, which has always been a promising idea that drew many users, especially young people. The use of social media today among teens is almost universal. The success of a social platform is largely dependent on its architecture, which dictates the nature of the interactions that can occur. It is interesting to note, he says, that when conversations can be overheard by others, it gives rise to potentially much more interesting interaction among users. It is starting to become clear where the risks are in this, with the combination of social networking being so easy to access by teens and people who are not security/privacy conscious.

III. Security and Privacy Risks in Social Networking

Needless to say, social networking is not without its security risks. A great majority of social networking deals with privacy. [6] tells us that there are many information management issues with social media services, mainly in the area of privacy and personally identifiable information and how to properly store and protect it. This often makes the information available to government agencies. This is because, as [2] puts it, "social networking sites create a central repository of personal information" which continues to grow as users keep adding to it. What makes this worse is teenagers, who are less worried about privacy and security, continue giving up information about themselves willingly. This is a huge part of the problem, and a possible solution that should help to combat this will be suggested later. Sometimes this is in the name of being popular. Sometimes this is just pure carelessness. [2] says the "private versus public boundaries of social media spaces are unclear." He goes on to note that parents are often very unaware, or not caring about, what their teens are putting online.

Another main risk with the privacy and security of information in social networks is the centralized architecture. As stated previously, social media servers are a gold mine of personally identifiable information, which is freely given up, by teenagers and adult users alike. [4] says that this gives rise to grave privacy concerns and can give rise to things like identity theft and selling of user data to third parties. Users have a false sense of trust in their social network provider to protect their information, when it is often being sold to third parties or hacked by identity thieves. He goes on to point out that while Facebook added privacy settings that the user can control, their default setting is public when an account is first created. Thus, a brand new user that does not changes these settings to make them more strict is actually posting information that can be viewed by the public and non-friends. [4] continues to show that the amount of information that trusting users put in their profiles on popular social media sites can be pieced together to form a picture of the user, if you will, that contains enough information to trick their friends into thinking it's really them. An identity thief can then create a false profile of that person, re-friend all of their friends, and then trick their friends into revealing more personal information about the user. [3] calls this practice "profile cloning." He states that some thieves steal information about users from one site to create a fake profile on another. He states that information can also be tricked out of users through the use of phishing attacks, where information is gleaned from users via setting up fake Websites that ask for personal information or even passwords and social security numbers.

Various other attacks, according to [3], are engineered to either take personal information from users, or infect their system with viruses. They include click jacking in which an attacker posts a video to a user and when the user plays it, malicious code is introduced into their system, and watering hole attacks, where a developer's forum is hacked and everyone that visits the forum gets their system infected by a Trojan horse virus. Other risks include scams and cyber bullying, too. The risk any user takes on will be proportional to the amount of personal information they choose to post, and how they set their security/privacy settings.

The biggest problem here, according to [3], is that many users are not aware of the privacy settings and how to use them. They are also "not aware of the risks associated with uploading sensitive information." Studies have shown that social media sites are designed to get as many users together into one place, and many of these users are unaware of how to use the privacy settings. These sites value "openness, connecting, and sharing with others – unfortunately the very aspects which allow cyber criminals to use these sites as a weapon for various crimes [3]." He goes on to say that employees often post company information on social networks, introducing risk to the organization they work for. When you see how naïve and trusting some people are, and how much private information is stored in a central repository like a social media service, it is easy to see that this is a very big reason why attackers go after social networks.

So it is plain to see, according to [4], that even though technology and policy may be used at the social networking sites the same as any other organization, the centralized structure and the huge repository of private information gives rise to huge security gaps. These can be addressed with more policy, some common sense by users, and some architectural changes.

IV. Some Possible Solutions

The rising tide of attacks on social networks, according to [3], tell us that “social networks and their millions of users have to do a lot more to protect themselves from organized cybercrime, or risk failing to identity theft schemes, scams, and malware attacks. Understanding these risks and challenges should be addressed to avoid potential loss of private and personal information.” Also, as [7] says, “The area of internet information security is well developed and evolves continuously in response to new threats” and so it must evolve with social media too.”

[3] gives some important tips for social network users to follow to help protect themselves online. The amount of personal information posted should be limited, and not post home addresses or private contact information. This, and information about your likes and daily routine can all be pieced together by a cybercriminal. Also, think of the Internet as public. Even if privacy settings are in place, information posted can still get out there, through friends reposting, and it is stored on servers that can be hacked. Be comfortable with the public seeing whatever you are posting on social network sites. Also be skeptical and beware of strangers. Not everyone is who he or she claims to be, and they could have stolen someone’s identity to commit cybercrime. Do not use the third party applications that are often making their way around Facebook. They often install malware that tracks your online activities. Use strong passwords, use anti-virus software, and keep your software up to date to help protect against the latest security threats. For those with kids, they need to be monitored very closely because they often do not know the wise techniques of online security or don’t care to keep themselves safe. Remember that once you post something, it never goes away even if you delete it, and know what to do to report someone that you suspect may be a security threat.

This goes into some other ideas that [2] brings up in her article, which are still applicable today. One thing she says is that parents need to be much more involved in the online activity of their children, since they are not experienced or wise enough to watch out for themselves or make the best decisions. Schools are also taking some actions in this regard, with policy and supervision, but not all schools are on the same page with this. Some are just letting kids suffer the natural consequences, and warning them that college and potential employers check their social networking pages and the posting of certain content is frowned upon and could result in non-admission or non-hiring.

This problem can also be combatted with changes to architecture and policy. One such architectural change was proposed by [3] in the form of a Secure Request-Response Application Architecture. This scheme involves the ability for a user to accept or reject another user’s request for information, whether they are a friend or not. The user can also set up two different databases for information depending on how much they trust the requestor. He can then protect his most sensitive information. [2] points out that some sites are implementing better and more customizable privacy settings. Facebook has overhauled their privacy system several times to make it more user-friendly to customize settings and give users the power over who can see individual posts. While this is not a completely safe solution, it does help, as long as people are aware of the features and use them wisely. But this is no substitute for being smart about what you post online. She goes on to say that many schools are making more of an effort to teach students about the importance of online privacy in the name of greater security. [4] gives us an idea that a decentralized architecture would help keep information safer. In this type of setup, any user’s information would never be all on the same server or even at the same facility. This would do a great deal to help prevent a full retrieval of a user’s profile by an attacker. An example of a social network that employs this approach is Diaspora. And according to [8], it is essential that a good amount of risk management be done. This will help solidify the security policy in place at the organization in question.

A very extensive paper was written by [9] that details an extremely complex study outlining the effects of unfriending people in your social profile. The main idea of this approach was that every friend someone has in their profile has a certain risk factor assigned to them by an algorithm, and this is based on the habits of how they interact with the user online, and how they pass on information, or repost things, to their friends. Due to this effect, even if you post online to just your friends, there is no guarantee they will not repost it to their friends, thus allowing the post to get outside your friend circle. So once the most vulnerable friends are identified using this algorithm, they can be defriended and have the effect of making your online experience more secure and private. The math formulas that went into this calculation were extremely complicated and likely nothing that would be comprehended by the average user, but the upshot of all of this is clear; unfriend people that are leaking your information, and your time online will be safer

V. A New Proposed Solution

The biggest problem here is carelessness in what is posted online, and this is one of the easiest to solve conceptually. A possible solution is certainly not complete, but will help put a dent in the problem and reduce the amount of carelessness on the Internet, and fits in with the idea of using education as one of the three ways to secure information systems. A proposal that all social networks, including Facebook, Twitter, Flickr, LinkedIn, as well as all portable applications that serve a similar purpose is suggested to require all new users, when signing up for an account, to view a short video that discusses the topic of Internet safety, personally identifiable information, and instructs users on that network’s privacy settings. The button to submit for an account should not appear until the video has played. This way it cannot be bypassed like the legal disclaimers that people just accept blindly. Also, any current users would have to watch the video on the day it goes online in order to continue using their accounts. To expand on the idea of yearly training often used in the military, the video could be required to be viewed once a year to remind users of its importance. Such an idea is rather easy to implement with the technology of today. With much better education, we can help combat this problem, especially if we also decentralize the information storage on these social networks.

VI. Conclusion

It is fairly clear from all of this research that social networks are big security and privacy risks. They have this risk because of their centralized architecture, their huge repository of all the personally identifiable information a hacker could ever want, and the general ignorance of the populace to how to properly use privacy settings to improve their online safety. There is also a large risk because many people, especially teenagers, are extremely trusting of other people and what type of information about themselves they reveal online.

This can only be combatted in a limited way by technological means, or even by policy. [10] tells us that we should consider any information sent through social media not secure, and therefore not transmit any sensitive information through social networks. The burden falls mainly on users to be smart about what they are doing online. The best thing we can do is to be smart when online.

But with better education and some architectural changes, social networks can be used more safely. Education is the biggest part. People fall into complacency and need to be reminded of things sometimes.

Lastly, it is important that research continue in the area of how to make social networks more secure even though trusting users are placing a plethora of personally identifiable information online.

References

[1] Hekkala, R., Väyrynen, K., & Wiander, T. (2012, June). Information Security Challenges of Social Media for Companies. In *ECIS* (p. 56).

[2] Barnes, S. (2006). A privacy paradox: Social networking in the United States. *First Monday, 11*(9). doi:10.5210/fm.v11i9.1394

[3] Kumar, A., Gupta, S. K., Rai, A. K., & Sinha, S. (2013). Social Networking Sites and Their Security Issues. *International Journal of Scientific and Research Publications*, *3*(4), 3.

[4] Verma, A., Kshirsagar, D., & Khan, S. (2013). Privacy and Security: Online Social Networking. *International Journal of Advanced Computer Research*, *3*(8), 310-315.

[5] Deng, X., Bispo, C. B., & Zeng, Y. (2014). A Reference Model for Privacy Protection in Social Networking Service. *Journal Of Integrated Design & Process Science*, *18*(2), 23-44. doi:10.3233/jid-2014-0007

[6] Bertot, J. C., Jaeger, P. T., & Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government Information Quarterly*, *29*(1), 30-40.

[7] Vladlena, B., Saridakis, G., Tennakoon, H., & Ezingeard, J. N. (2015). The role of security notices and online consumer behaviour: An empirical study of social networking users. *International Journal Of Human - Computer Studies*, *80*36-44. doi:10.1016/j.ijhcs.2015.03.004

[8] Kim, H. J. (2012). Online Social Media Networking and Assessing Its Security Risks. *International Journal Of Security & Its Applications*, *6*(3), 11-18.

[9] GUNDECHA, P., BARBIER, G., JILIANG, T., & HUAN, L. (2014). User Vulnerability and Its Reduction on a Social Networking Site. *ACM Transactions On Knowledge Discovery From Data*, *9*(2), 12:1-12:25. doi:10.1145/2630421

[10] Thompson, A. F., Otasowie, I., & Famose, O. A. (2014). Evaluation of Security Issues in Social Networks. *Computing & Information Systems*, *18*(1), 6-20.

The SVM Classifier Based on the Modified Particle Swarm Optimization

Liliya Demidova
Moscow Technological Institute,
Ryazan State Radio Engineering
University
Moscow, Russia

Evgeny Nikulchev
Moscow Technological Institute
Moscow, Russia

Yulia Sokolova
Ryazan State Radio Engineering
University
Ryazan, Russia

***Abstract*—The problem of development of the SVM classifier based on the modified particle swarm optimization has been considered. This algorithm carries out the simultaneous search of the kernel function type, values of the kernel function parameters and value of the regularization parameter for the SVM classifier. Such SVM classifier provides the high quality of data classification. The idea of particles' «regeneration» is put on the basis of the modified particle swarm optimization algorithm. At the realization of this idea, some particles change their kernel function type to the one which corresponds to the particle with the best value of the classification accuracy. The offered particle swarm optimization algorithm allows reducing the time expenditures for development of the SVM classifier. The results of experimental studies confirm the efficiency of this algorithm.**

Keywords—particle swarm optimization; SVM-classifier; kernel function type; kernel function parameters; regularization parameter; support vectors

I. Introduction

Currently, for the different classification problems in various applications the SVM algorithm (Support Vector Machines, SVM), which carries out training on precedents («supervised learning»), is successfully used. This algorithm includes in the group of boundary classification algorithms [1], [2].

The SVM classifiers by the SVM algorithm have been applied for credit risk analysis [3], medical diagnostics [4], handwritten character recognition [5], text categorization [6], information extraction [7], pedestrian detection [8], face detection [9], etc.

The main feature of the SVM classifier is using of the special function called the kernel, with which the experimental data set has been converted from the original space of characteristics into the higher dimension space with the construction of a hyperplane that separates classes. A herewith two parallel hyperplanes must be constructed on both sides of the separating hyperplane. These hyperplanes define borders of classes and have been situated at the maximal possible distance from each other. It has been assumed that the bigger distance between these parallel hyperplanes gives the better accuracy of the SVM classifier. Vectors of the classified objects' characteristics which are the nearest to the parallel hyperplanes are called support vectors. An example of the separating hyperplane building in the 2D space has been shown in Fig. 1.

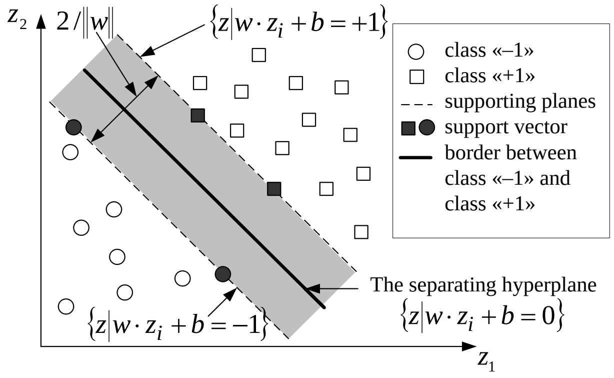

z_2
$2/\|w\|$
$\{z|w \cdot z_i + b = +1\}$
○ class «–1»
□ class «+1»
- - - supporting planes
■● support vector
— border between class «–1» and class «+1»
The separating hyperplane
$\{z|w \cdot z_i + b = 0\}$
$\{z|w \cdot z_i + b = -1\}$
z_1

Fig. 1. Linear separation for two classes by the SVM classifier in the 2D space

The SVM classifier supposes an execution of training, testing, and classification. Satisfactory quality of training and testing allows using the resulting SVM classifier in the classification of new objects.

Training of the SVM classifier assumes solving a quadratic optimization problem [1]–[3]. Using a standard quadratic problem solver for training the SVM classifier would involve solving a big quadratic programming problem even for a moderate sized data set. This can limit the size of problems which can be solved with the application of the SVM classifier. Nowdays methods like SMO [10, 11], chunking [12] and simple SVM [13], Pegasos [14] exist that iteratively compute the required solution and have a linear space complexity [15].

A solution for the problem which has been connected with a choice of the optimal parameters' values of the SVM classifier represents essential interest. It is necessary to find the kernel function type, values of the kernel function parameters and value of the regularization parameter, which must be set by a user and shouldn't change [1], [2]. It is impossible to provide implementing of high-accuracy data classification with the use of the SVM classifier without adequate solution of this problem.

Let values of the parameters of the SVM classifier be optimal, if high accuracy of classification has been achieved: numbers of error within training and test sets are minimal, moreover the number of errors within test set must not strongly differ from the number of errors within training set. It will allow excluding retraining of the SVM classifier.

In the simplest case solution of this problem can be achieved by a search of the kernel function types, values of the kernel function parameters and value of the regularization parameter that demands significant computational expenses. A herewith for an assessment of classification quality, the indicators of classification accuracy, classification completeness, etc. can be used [3].

In most cases of the development of binary classifiers, it is necessary to work with the complex, multiple extremal, multiple parameter objective function.

Gradient methods are not suitable for search of the optimum of such objective function, but search algorithms of stochastic optimization, such as the genetic algorithm [16]–[18], the artificial bee colony algorithm [19], the particle swarm algorithm [20], [21], etc., have been used. earch of the optimal decision is carried out at once in all space of possible decisions.

The particle swarm algorithm (Particle Swarm Optimization, PSO algorithm), which is based on an idea of possibility to solve the optimization problems using modeling of animals' groups' behavior is the simplest algorithm of evolutionary programming because for its implementation it is necessary to be able to determine only value of the optimized function [20], [21].

The traditional approach to the application of the PSO algorithm consists of the repeated applications of the PSO algorithm for the fixed type of the kernel functions to choose optimal values of the kernel function parameters and value of the regularization parameter with the subsequent choice of the best type of the kernel function and values of the kernel function parameters and value of the regularization parameter corresponding to this kernel function type.

Along with the traditional approach to the application of the PSO algorithm a new approach, that implements the simultaneous search for the best type of the kernel function, values of the kernel function parameters and value of the regularization parameter, is offered [22]. Hereafter, particle swarm algorithms corresponding to traditional and modified approaches will be called as the traditional PSO algorithm and the modified PSO algorithm consequently.

The objective of this paper is to fulfill a comparative analysis of the traditional and modified particle swarm algorithms, applied for the development of the SVM classifier, both on the search time of the optimal parameters of the SVM classifier and the quality of data classification.

The rest of this paper is structured as follows. Section II presents the main stages of the SVM classifier development. Then, Section III details the proposed new approach for solving the problem of the simultaneous search of the kernel function type, values of the kernel function parameters and value of the regularization parameter for the SVM classifier. This approach is based on the application of the modified PSO algorithm. Experimental results comparing the traditional PSO algorithm to the modified PSO algorithm follow in Section IV. Finally, conclusions are drawn in Section V.

II. The SMV Classifier

Let the experimental data set be a set in the form of $\{(z_1,y_1),\dots,(z_s,y_s)\}$, in which each object $z_i \in Z$ ($i=\overline{1,s}$; s is the number of objects) is assigned to number $y_i \in Y = \{+1;\ -1\}$ having a value of +1 or −1 depending on the class of object z_i. A herewith it is assumed that every object z_i is mapped to q-dimensional vector of numerical values of characteristics $z_i = (z_i^1, z_i^2, \dots, z_i^q)$ (typically normalized by values from the interval [0, 1]) where z_i^l is the numeric value of the l-th characteristic for the i-th object ($i=\overline{1,s}$, $l=\overline{1,q}$) [22]–[25]. It is necessary with use of special function $\kappa(z_i, z_\tau)$, which is called the kernel, to build the classifier $F:\ Z \to Y$, which compares to the class with number from the set $Y = \{+1;\ -1\}$ some object from the set Z.

To build «the best» SVM classifier it is necessary to realize numerous repeated training and testing on the different randomly generated training and test sets with following determination of the best SVM classifier in terms of the highest possible classification quality provision. The test set contains the part of data from the experimental data set. The size of the test set must be equal to 1/10 – 1/3 of the size of the experimental data set. The test set doesn't participate in the control of parameters of the SVM-classifier. This set is used for check of classifier's accuracy. The SVM classifier with satisfactory training and testing results can be used to classify new objects [22].

The separating hyperplane for the objects from the training set can be represented by equation $\langle w, z\rangle + b = 0$, where w is a vector-perpendicular to the separating hyperplane; b is a parameter which corresponds to the shortest distance from the origin of coordinates to the hyperplane; $\langle w, z\rangle$ is a scalar product of vectors w and z [1–3]. The condition $-1 < \langle w, z\rangle + b < 1$ specifies a strip that separates the classes. The wider the strip, the more confidently we can classify objects. The objects closest to the separating hyperplane, are exactly on the bounders of the strip.

In the case of linear separability of classes we can choose a hyperplane so that there is no any object from the training set between them, and then maximize the distance between the hyperplanes (width of the strip) $2/< w, w >$, solving the problem of quadratic optimization [1], [2]:

$$\begin{cases} \langle w, w\rangle \to \min, \\ y_i \cdot (\langle w, z_i\rangle + b) \ge 1, \quad i=\overline{1,S}. \end{cases} \tag{1}$$

The problem of the separating hyperplane building can be reformulated as the dual problem of searching a saddle point of the Lagrange function, which reduces to the problem of quadratic programming, containing only dual variables [1], [2]:

$$\begin{cases} -L(\lambda) = -\sum\limits_{i=1}^{S}\lambda_i + \\ \qquad + \frac{1}{2}\cdot\sum\limits_{i=1}^{S}\sum\limits_{\tau=1}^{S}\lambda_i\cdot\lambda_\tau\cdot y_i\cdot y_\tau\cdot\kappa(z_i,z_\tau)\to\min\limits_{\lambda}, \\ \qquad\qquad \sum\limits_{i=1}^{S}\lambda_i\cdot y_i = 0, \\ \qquad\qquad 0\le\lambda_i\le C,\ i=\overline{1,S}, \end{cases} \quad (2)$$

where λ_i is a dual variable; z_i is the object of the training set; y_i is a number (+1 or −1), which characterize the class of the object z_i from the experimental data set; $\kappa(z_i, z_\tau)$ is a kernel function; C is a regularization parameter ($C > 0$); S is a quantity of objects in the experimental data set; $i = \overline{1, S}$.

In training of the SVM classifier it is necessary to determine the kernel function type $\kappa(z_i, z_\tau)$, values of the kernel parameters and value of the regularization parameter C, which allows finding a compromise between maximizing of the gap separating the classes and minimizing of the total error. A herewith typically one of the following functions is used as the kernel function $\kappa(z_i, z_\tau)$ [1], [3], [26]:

- linear function: $\kappa(z_i, z_\tau) =< z_i, z_\tau >$;
- polynomial function: $\kappa(z_i, z_\tau) = (< z_i, z_\tau > +1)^d$;
- radial basis function:

$\kappa(z_i, z_\tau) = exp(-< z_i - z_\tau, z_i - z_\tau > /(2\cdot\sigma^2))$;

- sigmoid function: $\kappa(z_i, z_\tau) = th(k_2 + k_1 \cdot < z_i, z_\tau >)$,

where $< z_i, z_\tau >$ is a scalar product of vectors z_i and z_τ; d [$d \in N$ (by default $d = 3$)], σ [$\sigma > 0$ (by default $\sigma^2 = 1$)], k_2 [$k_2 < 0$ (by default $k_2 = -1$)] and k_1 [$k_1 > 0$ (by default $k_1 = 1$)] are some of parameters; th is a hyperbolic tangent.

These kernel functions allow dividing the objects from different classes.

As a result of the SVM classifier training the support vectors must be determined. These vectors are closest to the hyperplane separating the classes and contain all information about the classes' separation. The main problem dealing with the training of the SVM classifier, is the lack of recommendations for the choice of value of the regularization parameter, the kernel function type and values of the kernel function parameters, which can provide the high accuracy of objects' classification. This problem can be solved with the use of various optimization algorithms, in particular with the use of the PSO algorithm.

III. The Modified PSO Algorithm

In the traditional PSO algorithm the n-dimensional search space (n is the number of parameters which are subject to optimization) is inhabited by a swarm of m agents-particles (elementary solutions). Position (location) of the i-th particle is determined by vector $x_i = (x_i^1, x_i^2, \ldots, x_i^n)$, which defines a set of values of optimization parameters. A herewith these parameters can be presented in an explicit form or even absent in analytical record of the objective function $f(x) = f(x^1, x^2, \ldots, x^n)$ of the optimization algorithm (for example, the optimum is the minimum which must be achieved).

The particles must be situated randomly in the search space during the process of initialization. A herewith each i-th particle ($i = \overline{1, m}$) has its own vector of speed $v_i \in R^n$ which influence i-th particle ($i = \overline{1, m}$) coordinates' values in every single moment of time corresponding to some iteration of the PSO algorithm.

The coordinates of the i-th particle ($i = \overline{1, m}$) in the n-dimensional search space uniquely determine the value of the objective function $f(x_i) = f(x_i^1, x_i^2, \ldots, x_i^n)$ which is a certain solution of the optimization problem [20] – [22].

For each position of the n-dimensional search space where the i-th particle ($i = \overline{1, m}$) was placed, the calculation of value of the objective function $f(x_i)$ is performed. A herewith each i-th particle remembers the best value of the objective function found personally as well as the coordinates of the position in the n-dimensional space corresponding to the value of the objective function. Moreover each i-th particle ($i = \overline{1, m}$) «knows» the best position (in terms of achieving the optimum of the objective function) among all positions that had been «explored» by particles (due to it the immediate exchange of information is replicated by all the particles). At each iteration particles correct their velocity to, on the one hand, move closer to the best position which was found by the particle independently and, on the other hand, to get closer to the position which is the best globally at the current moment. After a number of iterations particles must come close to the best position (globally the best for all iterations). However, it is possible that some particles will stay somewhere in the relatively good local optimum.

Convergence of the PSO algorithm depends on how velocity vector correction is performed. There are different approaches to implementation of velocity vector v_i correction for the i-th particle ($i = \overline{1, m}$) [20]. In the classical version of the PSO algorithm correction of each j-th coordinate of velocity vector ($j = \overline{1, n}$) of the i-th particle ($i = \overline{1, m}$) is made in accordance with formula [20]:

$$v_i^j = v_i^j + \hat{\varphi} \cdot \hat{r} \cdot (\hat{x}_i^j - x_i^j) + \tilde{\varphi} \cdot \tilde{r} \cdot (\tilde{x}^j - x_i^j), \quad (3)$$

where v_i^j is the j-th coordinate of velocity vector of the i-th particle; x_i^j is the j-th coordinate of vector x_i, defining the position of the i-th particle; $\hat{x}_i^j$ is the j-th coordinate of the best position vector found by the i-th particle during its existence; $\tilde{x}^j$ is the j -th coordinate of the globally best position within the particles swarm in which the objective function has the optimal value; $\hat{r}$ and $\tilde{r}$ are random numbers in interval (0, 1), which introduce an element of randomness in the search process; $\hat{\varphi}$ and $\tilde{\varphi}$ are personal and global coefficients for particle acceleration which are constant and determine behavior and effectiveness of the PSO algorithm in general.

With personal and global acceleration coefficients in (3) random numbers $\hat{r}$ and $\tilde{r}$ must be scaled; a herewith the global acceleration coefficient $\tilde{\varphi}$ operates by the impact of the global best position on the speeds of all particles and the personal acceleration coefficient $\hat{\varphi}$ operates by the impact of the personal best position on the velocity of some particle.

Currently different versions of the traditional PSO algorithm are known. In one of the most known canonical version of the PSO algorithm it is supposed to undertake the normalization of the acceleration coefficients $\hat{\varphi}$ and $\tilde{\varphi}$ to make the convergence of the algorithm not so much dependent on the choice of their values [20].

A herewith correction of each j -th coordinate of the velocity vector ($j=\overline{1,n}$) of the i -th particle ($i=\overline{1,m}$) is performed in accordance with formula:

$$v_i^j = \chi \cdot [v_i^j + \hat{\varphi} \cdot \hat{r} \cdot (\hat{x}_i^j - x_i^j) + \tilde{\varphi} \cdot \tilde{r} \cdot (\tilde{x}^j - x_i^j)], \quad (4)$$

where χ is a compression ratio;

$$\chi = 2 \cdot K \,/\, \left| 2 - \varphi - \sqrt{\varphi^2 - 4 \cdot \varphi} \right|; \quad (5)$$

$$\varphi = \hat{\varphi} + \tilde{\varphi} \ \ (\varphi > 4); \quad (6)$$

K is the some scaling coefficient, which takes values from the interval (0, 1).

When using formula (4) for correction of velocity vector the convergence of the PSO algorithm is guaranteed and there is no need to control the particle velocity explicitly [20].

Let the correction of velocity vector of the i -th particle ($i=\overline{1,m}$) is executed in accordance with one of the formulas (3) or (4). The correction of the j -th position coordinate of the i -th particle ($i=\overline{1,m}$) can be executed in accordance with the formula:

$$x_i^j = x_i^j + v_i^j . \quad (7)$$

Then for each i -th particle ($i=\overline{1,m}$) the new value of the objective function $f(x_i)$ can be calculated and the following check must be perfomed: whether a new position with coordinates vector x_i became the best among all positions in which the i -th particle has previously been placed. If new position of the i -th particle is recognized to be the best at the current moment the information about it must be stored in a vector $\hat{x}_i$ ($i=\overline{1,m}$).

A herewith value of the objective function $f(x_i)$ for this position must be remembered. Then among all new positions of the swarm particles the check of the globally best position must be carried out. If some new position is recognized as the best globally at the current moment, the information about it must be stored in a vector $\tilde{x}$. A herewith value of the objective function $f(x_i)$ for this position must be remembered.

In the case of the SVM classifier's development with the use of the PSO algorithm the swarm particles can be defined by vectors declaring their position in the search space and corded by the kernel function parameters and the regularization parameter: (x_i^1, x_i^2, C_i) , where i is a number of particle ($i=\overline{1,m}$); x_i^1, x_i^2 are the kernel function parameters of the i -th particle, [a herewith parameter x_i^1 is equal to the kernel function parameters d , σ or k_2 (depending on the kernel function type which corresponds to a swamp particle); parameter x_i^2 is equal to the kernel function parameter k_1 , if the swamp particle corresponds to the sigmoid type of the kernel function, otherwise this parameter is assumed to be zero]; C_i is the regularization parameter.

Then traditional approach to the application of the PSO algorithm in developing the SVM classifier must be concluded in numerous implementation of the PSO algorithm under the fixed kernel function type aiming to choose the optimal parameters values of the kernel function and value of the regularization parameter.

As result for each type T of the kernel function, participating in the search, the particle with the optimal combination of the parameters values ($\tilde{x}^1, \tilde{x}^2, \tilde{C}$) providing high quality of classification will be defined.

The best type and the best values of the required parameters get out by results of the comparative analysis of the best particles received at realization of the PSO algorithm with the fixed kernel function type.

Along with the traditional approach to the application of the PSO algorithm in the development of the SVM classifier there is a new approach that implements a simultaneous search for the best kernel function type $\tilde{T}$, parameters' values $\tilde{x}^1$ and $\tilde{x}^2$ of the kernel function and value of the regularization parameter $\tilde{C}$. At such approach each i -th particle in a swamp ($i=\overline{1,m}$) defined by a vector which describes particle's position in the search space: (T_i, x_i^1, x_i^2, C_i) , where T_i is the number of the kernel function type (for example, 1, 2, 3 – for polynomial, radial basis and sigmoid functions accordingly);

parameters x_i^1, x_i^2, C_i are defined as in the previous case. A herewith it is possible to «regenerate» particle through changing its coordinate T_i on number of that kernel function type, for which particles show the highest quality of classification. In the case of particles' «regeneration» the parameters' values change so that they corresponded to new type of the kernel function (taking into account ranges of change of their values). Particles which didn't undergo «regeneration», carry out the movement in own space of search of some dimension.

The number of particles taking part in «regeneration» must be determined before start of algorithm. This number must be equal to 15% – 25% of the initial number of particles. It will allow particles to investigate the space of search. A herewith they won't be located in it for a long time if their indicators of accuracy are the worst.

The offered modified PSO algorithm can be presented by the following consequence of steps.

Step 1. To determine parameters of the PSO algorithm: number m of particles in a swamp, velocity coefficient K, personal and global velocity coefficients $\hat{\varphi}$ and $\tilde{\varphi}$, maximum iterations number $N_{\max}$ of the PSO algorithm. To determine types T of kernel functions, which take part in the search ($T=1$ – polynomial function, $T=2$ – radial basis function, $T=3$ – sigmoid function) and ranges boundaries of the kernel function parameters and the regularization parameter C for the chosen kernel functions' types T : $x_{\min}^{1T}$, $x_{\max}^{1T}$, $x_{\min}^{2T}$, $x_{\max}^{2T}$, $C_{\min}^{T}$, $C_{\max}^{T}$ ($x_{\min}^{2T}=0$ and $x_{\max}^{2T}=0$ for $T=1$ and $T=2$). To determine the particles' «regeneration» percentage p .

Step 2. To define equal number of particles for each kernel type function T , included in search, to initialize coordinate T_i for each i -th particle ($i=\overline{1,m}$) (a herewith every kernel function type must be corresponded by equal number of particles), other coordinates of the i -th particle ($i=\overline{1,m}$) must be generated randomly from the corresponding ranges: $x_i^1 \in [x_{\min}^{1T},\ x_{\max}^{1T}]$, $x_i^2 \in [x_{\min}^{2T},\ x_{\max}^{2T}]$ ($x_i^2=0$ under $T=1$ and $T=2$), $C_i \in [C_{\min}^{T},\ C_{\max}^{T}]$. To initialize random velocity vector $v_i(v_i^1, v_i^2, v_i^3)$ of the i -th particle ($i=\overline{1,m}$) ($v_i^2=0$ under $T=1$ and $T=2$). To establish initial position of the i -th particle ($i=\overline{1,m}$) as its best known position $(\hat{T}_i, \hat{x}_i^1, \hat{x}_i^2, \hat{C}_i)$, to determine the best particle with coordinates' vector $(\tilde{T}, \tilde{x}^1, \tilde{x}^2, \tilde{C})$ from all the m particles, and to determine the best particle for each kernel function type T , including in a search, with coordinates' vector $(\overline{T}, \overline{x}^{1T}, \overline{x}^{2T}, \overline{C}^{T})$. Herewith number of executed iterations must be considered as 1.

Step 3. To execute while the number of iterations is less than the fixed number $N_{\max}$:

- «regeneration» of particles: to choose p % of particles which represent the lowest quality of classification from particles with coordinate $T_i \neq \tilde{T}$ ($i=\overline{1,m}$); to change coordinate T_i (with the kernel function type) on $\tilde{T}$; to change values of the parameters x_i^1, x_i^2, C_i of «regenerated» particles to let them correspond to a new kernel function type $\tilde{T}$ (within the scope of the corresponding ranges);
- correction of velocity vector $v_i(v_i^1, v_i^2, v_i^3)$ and position (x_i^1, x_i^2, C_i) of the i -th particle ($i=\overline{1,m}$) using formulas:

$$v_i^j = \begin{cases} \chi \cdot [v_i^j + \hat{\varphi} \cdot \hat{r} \cdot (\hat{x}_i^j - x_i^j) + \tilde{\varphi} \cdot \tilde{r} \cdot (\overline{x}^{jT} - x_i^j)], & j=1,\ 2, \\ \chi \cdot [v_i^j + \hat{\varphi} \cdot \hat{r} \cdot (\hat{C}_i - C_i) + \tilde{\varphi} \cdot \tilde{r} \cdot (\overline{C}^{T} - C_i)], & j=3, \end{cases} \tag{8}$$

$$x_i^j = x_i^j + v_i^j \text{ for } j=1,\ 2, \tag{9}$$

$$C_i = C_i + v_i^3, \tag{10}$$

where $\hat{r}$ and $\tilde{r}$ are random numbers in interval (0, 1), χ is a compression ratio calculated using the formula (5); a herewith formula (8) is the modification of formula (4): the coordinates' values $\overline{x}^{1T}, \overline{x}^{2T}, \overline{C}^{T}$ are used instead of the coordinates' values $\tilde{x}^1, \tilde{x}^2, \tilde{C}$ of the globally best particle;

- accuracy calculation of the SVM classifier with parameters' values (T_i, x_i^1, x_i^2, C_i) ($i=\overline{1,m}$) with aim to find the optimal combination $(\tilde{T}, \tilde{x}^1, \tilde{x}^2, \tilde{C})$, which will provide high quality of classification;
- increase of iterations number on 1.

The particle with the optimal combination of the parameters' values $(\tilde{T}, \tilde{x}^1, \tilde{x}^2, \tilde{C})$ which provides the highest quality of classification on chosen the function types will be defined after execution of the offered algorithm.

After executing of the modified PSO algorithm it can be found out that all particles will be situated in the search space which corresponds to the kernel function with the highest classification quality because some particles in the modified PSO algorithm changed their coordinate, which is responsible for number of the kernel function. A herewith all other search spaces will turn out to be empty because all particles will «regenerate» their coordinate with number of the kernel function type. In some cases (for small values of the iterations' number $N_{\max}$ and for small value of the particles' «regeneration» percentage p) some particles will not «regenerate» their kernel function type and will stay in their initial search space.

Using of this approach in the application of the PSO algorithm in the problem of the SVM classifier development allows reducing the time required to construct the desired SVM classifier.

Quality evaluation of the SVM classifier can be executed with the use of different classification quality indicators [3]. There are cross validation data indicator, accuracy indicator, classification completeness indicator and ROC curve analysis based indicator, etc.

IV. Experimental Studies

The feasibility of the modified PSO algorithm using for the SVM classifier development was approved by test and real data. In the experiment for a particular data set the traditional PSO algorithm and the modified PSO algorithm were carried out. Comparison between these algorithms was executed using the found optimal parameters values of the SVM algorithm, classification accuracy and spent time.

Actual data used in the experimental researches was taken from Statlog project and from UCI machine learning library. Particularly, we used two data sets for medical diagnostics and one data set for credit scoring:

- breast cancer data set of The Department of Surgery at the University of Wisconsin, in which the total number of instances is 569 including 212 cases with the diagnosed cancer (class 1) and 357 cases without such diagnosis (class 2); a herewith each patient is described by 30 characteristics ($q = 30$) and all information was obtained with the use of digital images (WDBC data set in the Table, the source is http://archive.ics.uci.edu/ml/; machine-learning-databases/breast-cancer-wisconsin/);
- heart disease data set, in which the total number of instances is 270 including 150 cases with the diagnosed heart disease (class 1) and 120 cases without such diagnosis (class 2); a herewith each patient is described by 13 characteristics ($q = 13$) (Heart data set in the Table, the source is http:// archive.ics.uci.edu/ml/machine-learning-databases/ statlog/heart/; a herewith desease was found for 150 patients (class 1) and desease was not found for 120 patients (class 2));
- Australian consumer credit data set, in which the total number of instances is 690 including 382 creditworthy cases (class 1) and 308 default cases (class 2); a herewith each applicant is described by 14 characteristics ($q = 14$) (Australian data set in the Table, the source is http://archive.ics.uci.edu/ml/ machine-learning-data bases/statlog/australian/).

Moreover two testing data sets were used in experimental researches: Test [11] and МОТП12 (the source is http://machinelearning.ru/wiki/images/b/b2/ MOTP12_svm_example.rar).

For all data sets binary classification was performed.

For development of the SVM classifier the traditional and the modified PSO algorithms were used; a herewith the choice of the optimal values of the SVM classifier parameters was realized. The kernels with polynomial, radial basis and sigmoid functions were included in the search and the identical values of the PSO algorithm parameters and the identical ranges of values' change of the required SVM classifier parameters were established.

The short description of characteristics of each data set is provided in the Table. Here search results of the optimal values of parameters of the SVM classifier with the application of the traditional PSO algorithm and the modified PSO algorithm are presented (in the identical ranges of parameters' change and at the identical PSO algorithm parameters), number of error made during the training and testing of the SVM classifier and search time. For example, for WDBC data set with the use of the traditional and the modified PSO algorithms the kernel with radial basis function (number 2) was determined as the optimal. For the traditional PSO algorithm the optimal values of the kernel parameter and the regularization parameter are equal to $\sigma = 6.81$ and $C = 4.93$ accordingly. For the modified PSO algorithm the optimal values of the kernel parameter and the regularization parameter are equal to $\sigma = 4.01$ and $C = 9.83$ accordingly.

The classification accuracy by the traditional PSO algorithm is equal to 99.12%, and the classification accuracy by the modified PSO algorithm is equal to 99.65%. A herewith the search time came to 10108 and 3250 seconds accordingly.

For Heart data set in the Figures 2 – 4 the examples of position of the particles swarm in the D-2 search spaces and in the D-3 search space during initialization, at the 3-rd iteration and at the 12-th iteration (with the use of the modified PSO algorithm) are shown.

The kernels with polynomial, radial basis and sigmoid functions were included in the search. A herewith the following change ranges of values' parameters were set: $3 \le d \le 8$, $d \in N$ (for polynomial function); $0.1 \le \sigma \le 10$ (for radial basis function); $-10 \le k_2 \le -0.1$ и $0.1 \le k_1 \le 10$ (for sigmoid function).

TABLE I. THE SEARCH RESULTS BY MEANS OF THE TRADITIONAL PSO ALGORITHM AND THE MODIFIED PSO ALGORITHM

Data set	Number of objects	Number of characteristics	PSO algorithm type	Found parameters				Errors		Number of support vectors	Accuracy (%)	Search time (sec.)
				Kernel number	C	x_1	x_2	During the training	During the testing			
WDBC	569	30	traditional	2	4.93	6.81	-	5 of 427	0 of 142	66	99.12	10108
			modified	2	9.83	4.01	-	1 of 427	1 of 142	80	99.65	3250
Heart	270	13	traditional	2	8.51	3.18	-	7 of 192	11 of 78	106	93.33	6558
			modified	2	5.92	2.89	-	4 of 192	10 of 78	99	94.81	2733
Australian	690	14	traditional	2	6.79	2.31	-	21 of 492	28 of 198	237	92.9	16018
			modified	2	7.64	2.38	-	21 of 492	26 of 198	225	93.19	8013
MOTП12	400	2	traditional	2	8.10	0.18	-	9 of 340	9 of 60	166	95.50	15697
			modified	2	6.37	0.26	-	10 of 340	8 of 60	107	95.50	9145
Test	300	2	traditional	1	8.85	3	-	0 of 240	0 of 60	7	100	3433
			modified	1	9.24	3	-	0 of 240	0 of 60	6	100	648

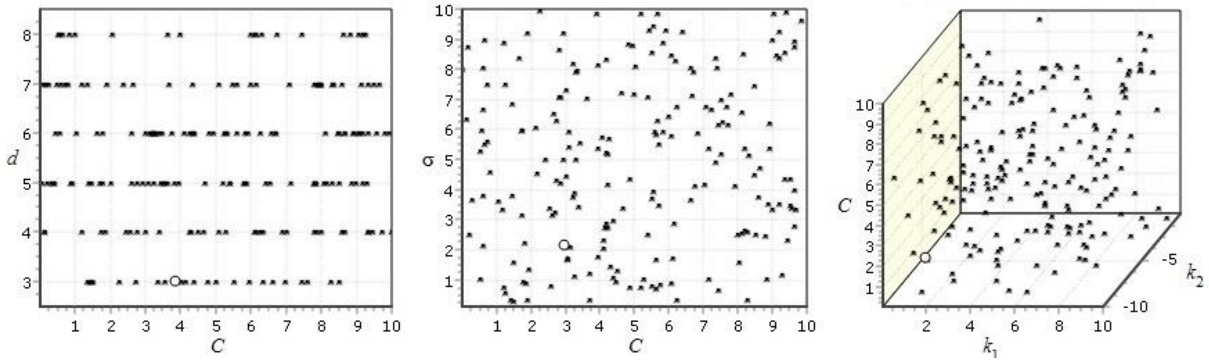


Fig. 2. Location of particles in a swamp during the initialization (polynomial kernel function is on the left, radial basis is in the middle, sigmoid is on the right)

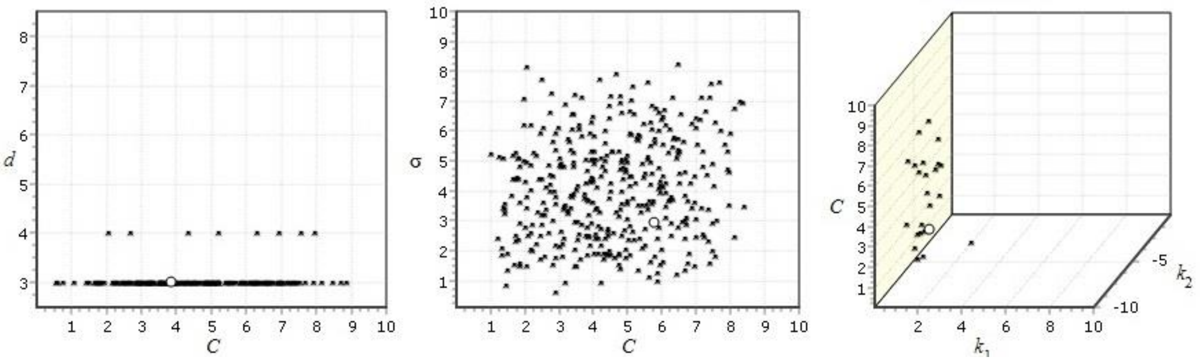


Fig. 3. Location of particles in a swamp during the 3-rd iteration (polynomial kernel function is on the left, radial basis is in the middle, sigmoid is on the right)

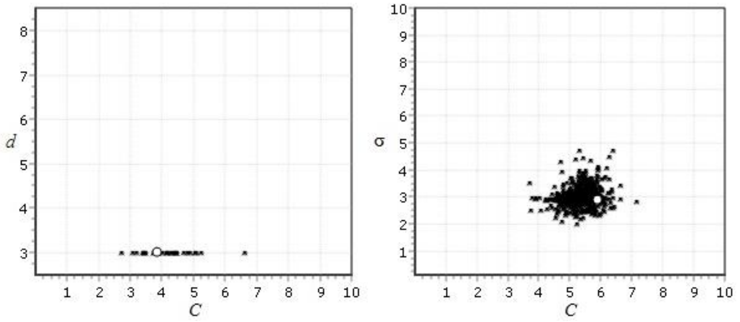


Fig. 4. Location of particles in a swamp during the 12-th iteration (radial basis kernel function is on the left and sigmoid kernel function is on the right)

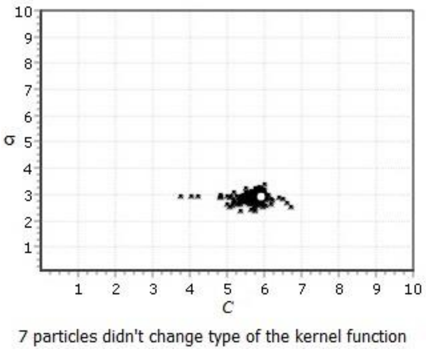


Fig. 5. Location of particles after the 20-th iteration

Change range for the regularization parameter C was determined as: $0.1 \le C \le 10$. Moreover, the following values of parameters of the PSO algorithm were set: number m of particles in a swarm equal to 600 (200 per each kernel function type); iterations' number $N_{\max} = 20$; personal and global velocity coefficients equal to $\hat{\varphi} = 2$ and $\tilde{\varphi} = 5$ accordingly; the scaling coefficient $K = 0.3$; «regeneration» coefficient of particles $p = 20\%$. Particles marked by asterisk bullets in the search spaces and the best position from the search space is marked by white round bullet. During realization of the modified PSO algorithm the swamp particles moves towards the best (optimal) position for the current iteration in the search space and demonstrate collective search of the optimal position. A herewith velocity and direction of each particle are corrected. Moreover «regeneration» of particles takes place: some particles change own search space to space, in which particles show the best quality of classification.

Thus, during realization of the modified PSO algorithm there is a change of the particles' coordinates, which are responsible for parameters of the kernel function $\kappa(z_i, z_\tau)$ and the regularization parameter C. Besides, the type of the kernel function also changes. As a result the particles moves towards the united search space (in this case – the space corresponding to radial basis kernel function) leaving the space where they were initialized.

In the reviewed example only 7 particles didn't change their kernel function type after 20 iterations. Other particles situated near the best position responsible for the optimal solution in the search space (Figure 5).

It is visible from the Table that that as a result of search for the reviewed data sets both algorithms determined identical kernel function type as the optimal, similar values of the kernel function parameter and the regularization parameter, and also similar accuracy values of training and testing of the SVM classifier.

But the modified PSO algorithm is more effective, because it took less (more than 2–3 times) time for search than traditional one.

V. Conclusion

The experimental results obtained on the base of the test data traditionally used to assess the classification quality, confirm the efficiency of the modified PSO algorithm. This algorithm allows choosing the best kernel function type, values of the kernel function parameters and value of the regularization parameter with the time expenditures which are significantly less, than in the case of the traditional PSO algorithm. A herewith high accuracy of classification is provided.

The obtained results had been reached thanks to «regeneration» of particles in the modified PSO algorithm. Particles which participate in the «regeneration» process change their kernel function type to the one which corresponds to the particle with the best value of the classification accuracy. Also, these articles change the accessory ranges of their parameters.

Further researches will have been devoted to the development of recommendations on the application of the modified PSO algorithm in the solution of the practical problems.

References

[1] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing Multiple Parameters for Support Vector Machine," Machine Learning, vol. 46, no. 1–3, pp. 131–159, 2002.

[2] V. Vapnik, Statistical Learning Theory. Wiley, New York, 1998.

[3] L. Yu, S. Wang, K. K. Lai and L. Zhou, Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines. Springer-Verlag, 2008.

[4] J.S. Raikwal and K. Saxena, "Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over Medical Data set," International Journal of Computer Applications, vol. 50, no. 14, pp. 35–39, 2012.

[5] Y. LeCun, L.D. Jackel, L. Bottou, C. Cortes at al., "Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition," in Neural Networks: The Statistical Mechanics Perspective, J. H. Oh, C. Kwon and S. Cho, Eds. World Scientific, 1995, pp. 261–276.

[6] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Lecture Notes in Computer Science, vol. 1398, pp. 137–142, 2005.

[7] Y. Li, K. Bontcheva and H. Cunningham, "SVM Based Learning System For Information Extraction," Lecture Notes in Computer Science, vol. 3635, pp. 319–339, 2005.

[8] M. Oren, C. Papageorgious, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian Detection Using Wavelet Templates," in 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 193–199.

[9] E. Osuna, R. Freund and F. Girosi, "Training Support Vector Machines: An Application to Face Detection," in 1997 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 1997, pp. 130–136.

[10] J.C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," in Advances in Kernel Methods. Support Vector Learning, 1998, pp. 185–208.

[11] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya and K.R.K. Murthy, "Improvements to the SMO Algorithm for SVM Regression," IEEE Trans. on Neural Networks, vol. 11, no. 5, pp. 1188–1193, 2000.

[12] E. Osuna, R. Freund and F. Girosi, "Improved Training Algorithm for Support Vector Machines," in 1997 IEEE Workshop Neural Networks for Signal Processing, 1997, pp. 24–26.

[13] S.V.N. Vishwanathan, A. Smola and N. Murty, “SSVM: a simple SVM algorithm,” Proceedings of the 2002 International Joint Conference on Neural Networks, vol. 3, pp. 2393-2398, 2002.

[14] S. Shalev-Shwartz, Y. Singer, N. Srebro and A. Cotter, “Pegasos: Primal Estimated sub-Gradient Solver for SVM,” Mathematical Programming, vol. 127, no. 1, pp. 3–30, 2011.

[15] L. Bottou and C.-J. Lin. Support Vector Machine Solvers, 2007.

[16] D.E. Goldberg, B. Korb and K. Deb, “Messy genetic algorithms: Motivation, analysis, and first results,” Complex Systems, vol. 3, no. 5, pp. 493–530, 1989.

[17] D.R. Eads, D. Hill, S. Davis, S.J. Perkins, J. Ma at al., “Genetic algorithms and support vector machines for time series classification,” in Proc. SPIE 4787 Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation, vol. 74, 2002, p. 74.

[18] S. Lessmann, R. Stahlbock and S.F. Crone, “Genetic algorithms for support vector machine model selection,” in 2006 International Joint Conference on Neural Networks, 2006, pp. 3063–3069.

[19] D. Karaboga and B. Basturk, “Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems,” in Proc. of the 12th Intern. Fuzzy Systems Association world congress on Foundations of Fuzzy Logic and Soft Computing, 2007, pp. 789–798.

[20] J. Sun, C.-H. Lai and X.-J. Wu, Particle Swarm Optimisation: Classical and Quantum Perspectives. CRC Press, 2011.

[21] R. Poli, J. Kennedy and T. Blackwell, “Particle swarm optimization,” Swarm Intelligence, vol. 1, no. 1, pp. 33–57, 2007.

[22] L. Demidova and Yu. Sokolova, “Modification Of Particle Swarm Algorithm For The Problem Of The SVM Classifier Development,” in 2015 International Conference "Stability and Control Processes" (SCP), 2015, pp. 623–627.

[23] L. Demidova, Yu. Sokolova and E. Nikulchev, “Use of Fuzzy Clustering Algorithms’ Ensemble for SVM classifier Development,” International Review on Modelling and Simulations, vol. 8, no. 4, pp. 446–457, 2015.

[24] L. Demidova and Yu. Sokolova, “SVM-Classifier Development With Use Of Fuzzy Clustering Algorithms' Ensemble On The Base Of Clusters' Tags' Vectors' Similarity Matrixes,” in 16th International Symposium on Advanced Intelligent Systems, 2015, pp. 889–906.

[25] L. Demidova, and Yu. Sokolova, “Training Set Forming For SVM Algorithm With Use Of The Fuzzy Clustering Algorithms Ensemble On Base Of Cluster Tags Vectors Similarity Matrices,” in 2015 International Conference Stability and Control Processes (SCP), pp. 619–622, 2015.

[26] A. Karatzoglou, D. Meyer and K. Hornik, Support vector machines in R. Research Report, WU Vienna, 2005.

3D Servicescape Model: Atmospheric Qualities of Virtual Reality Retailing

Aasim Munir Dad
The School of Business & Management
University of Gloucestershire
United Kingdom

Professor Barry Davies
The School of Business & Management
University of Gloucestershire
United Kingdom

Dr. Asma Abdul Rehman
Cardiff Metropolitan University
United Kingdom

***Abstract*—The purpose of this paper is to provide a 3D servicescape conceptual model which explores the potential effect of 3D virtual reality retail stores' environment on shoppers' behaviour. Extensive review of literature within two different domains, namely: servicescape models, and retail atmospherics, was carried out in order to propose a conceptual model. Further, eight detailed interviews were conducted to confirm the stimulus dimension of the conceptual model. A 3D servicescape conceptual model is offered on the basis of stimulus-organism-dimension, which proposes that a 3D virtual reality retail (VRR) store environment consists of physical, social, socially symbolic and natural dimensions. These dimensions are proposed to affect shoppers' behaviour through the mediating variables of emotions (pleasure and arousal). An interrelationship between pleasure and arousal, as mediating variables, is also proposed. This research opens a number of new avenues for further research through the proposed model of shoppers' behaviour in a VRR store environment. Further, a systematic taxonomy development of VRR store environment is attempted through this proposed model that may prove to be an important step in theory building. A comprehensive 3D service scape model along with a large number of propositions is made to define a 3D VRR store environment.**

Keywords*—*Virtual Reality Retailing (VRR); Servicescape; 3D Servicescape; Retail Atmospherics; Shoppers' behaviour

I. Introduction

Consumer behaviour has been known to be one of the most important areas of interest for marketers and academics for a long time (Dooley et al., 2012; Richarme, 2007). Organizations are always interested to know the effect of almost every single marketing activity on customers' behaviour and the reason for this is they want to satisfy their customers' needs such that they keep making a profit in the market (Schiffman et al., 2010; Young et al., 2010). Within the realm of consumer behaviour, the most popular research is the one in which researchers investigates the effect of the retail environment on shoppers' positive or negative shopping behaviour (Graves, 2013).

Retailing is one of the oldest models of selling goods; however, today the face of retail industry has been changed a lot because of new technological developments within it (Chang, 2014). The internet has made shopping far quicker, easier and more efficient; today shoppers can shop using their mobile devices (e.g. cell phones, tablets etc.) and desktop computers while comparing prices and quality at the same time. Self-check and contactless payments made the shopping process faster (Gibbs, 2015; Chang, 2014).

In the past, shoppers had two main channels of retailing: brick and mortar and web retailing (present on web 2.0 interface). However, today shoppers have another newly emerging channel called 3D virtual reality retailing, which is accessible in 3D virtual worlds. These 3D VRR stores are present in 3D virtual worlds, where users (known as residents) can buy and sell goods by exchanging virtual money (Vrechopoulos et al., 2009). As mentioned here and in detail in the literature review, there is a plethora of research into brick and mortar and online retail environments; however, research into virtual reality retail environments is still in its infancy (Hassouneh and Brengman, 2015; Krasonikolakis et al., 2011).

The current study provides a parallel research in the context of 3D virtual reality retail stores environment. Some researchers (Hassouneh and Brengman, 2015; Krasonikolakis et al., 2011; Vrechopoulos et al., 2009) have already started calling for more systematic research on the nature of the 3D virtual reality retail format by using established retailing and consumer behaviour theories. A large number of research questions have emerged in the domain of retail atmospherics; considering the impact of brick and mortar and web retail atmospheres on shoppers' behaviour; this prompts such questions as: what, if any, is the role of such atmospheric cues in the 3D virtual reality retail (VRR) shopping environment? What cues do virtual shoppers notice while shopping in 3D VRR stores? What is the effect of 3D VRR stores' environmental cues on shoppers' emotions and behaviours, for instance pleasure, arousal, dominance, satisfaction, and approach or avoidance behaviours? The purpose of this study is to provide a basic 3D VRR store atmospherics domain. The objective of this research is to define the 3D virtual reality retail store environment by establishing a list of stimuli that form a 3D VRR store environment.

II. Literature Review

In this section, a brief review of literature is presented to understand the relationship between the retail environment and shoppers' behaviour. The aim of this research is to explore 3D virtual reality retail stores' environments and their effect on virtual shoppers' behaviour; however, there is a paucity of research in this new electronic channel of retailing. Therefore, parallel literature is reviewed in terms of brick and mortar and traditional online (2 dimensional/ 2D / Web 2.0) retailing. A model is also conceptualized on the basis of a theoretical review of brick and mortar and online retail environment studies.

A. The Retail Environment and Shoppers' Behaviour:

According to The Statistical Portal, in the UK in 2014, £404.26 billion was generated in the retail industry as sales revenue (Statista, 2015) and it is estimated to be £443.25 billion in 2018. Moreover, in the United States retail sales generated a record sales revenue of $4.5 trillion in 2013, which is expected to rise to $5.552 trillion in 2018 (eMarketer, 2014). The retail industry has grown enormously and today there is huge competition among retailers and there are about 281,930 retailers on the UK's high streets (Centre of Retail Research, 2013). Baker et al. (1992) postulated that since creating a competitive advantage through product differentiation in this era is almost impossible, retailers can create differentiation through the retail store itself. Solomon et al. (2006) supported Baker's statement by arguing that in today's world of huge competition, many retail stores and airlines also try to create differentiation with their competitors by providing a unique environment to their customers. Graves (2013) confirmed all the previous research and determined that if retailers want to know about why shoppers buy what they buy or why they do not, then retailers need to understand the chemistry of the retail environment and how it influences shoppers' behaviour. According to Quartier (2011), research in this area is called 'retailology'. Researchers have been discussing the retail environment and its effect on shoppers since 1973; Kotler (1973) was the first to discuss the retail atmosphere and declare it as an important part of the 'Total Product'. Kotler (1973) further categorized the environment into two parts: Intended Environment and Perceived Environment. Intended Environment is created by the retailers through planning, whereas the Perceived Environment is the one which is judged or perceived by the consumers, which varies from consumer to consumer because of their individual personality characteristics (Kotler, 1973).

Furthermore, customers obtain satisfaction from their shopping when the entire shopping experience either meets or exceeds their expectations prior to shopping, and it can be achieved when they have ease in shopping, ease in fulfilling the transaction process, and customer satisfaction after purchasing and consuming goods or services (Dunne et al., 2002). Furthermore, it has already been proven that the store environment (store layouts and design: one of the important environmental cues of the brick and mortar and online retail environment) plays an important role in providing shoppers with a feeling of ease in shopping (Manganari et al., 2011; Vrechopoulos et al., 2004). Other researchers have also provided a plethora of research in which they have proven that the retail environment contributes a major role for making shopping easy and enjoyable (Varley and Rafiq, 2014; Quartier, 2011; Dunne et al., 2002; Eroglu et al., 2001; Donovan et al., 1994; Donovan and Rossiter, 1982). Graves (2013) claimed that to maximise sales it is vital that the retail environment is optimised. Moreover, the environment and its effect on people has long been discussed in the field of psychology (for example, by Stokols, 1978; Mehrabian, 1976; Craik, 1973) and Mehrabian and Russell (M-R, 1974) are credited with presenting a Stimulus – Organism - Response (S - O - R) model, which postulates that every environment affects the human beings present in it. This model further explains that every environment affects human emotions (Mehrabian and Russell posited that all types of human emotions could be categorised into three main emotions i.e. Pleasure, Arousal, and Dominance, also known as PAD), which lead individuals to show either approach or avoidance behaviour. This M-R model was later modified by Russell and Pratt (1980), who deleted the dominance dimension of the emotions because of its weak relationship with behaviour. Most retail environment studies (e.g. Wang et al., 2011; Davis et al., 2008; Baker et al., 1992; Donovan and Rossiter, 1982) that have adapted the M-R model did not include the dominance dimension, as suggested by Russell and Pratt (1980). In this research the dominance dimension is also not considered, as suggested by Russell and Pratt (1980). Their theory is general enough to be applicable in any kind of environment, even in a retail environment. However, environmental psychologists tended to focus on different environments (such as residential, entertainment, institutional, hospitals, schools and prisons) instead of retail environments before 1982 (Quartier, 2011; Donovan and Rossiter, 1982). Therefore, Donovan and Rossiter (1982) were pioneers in applying Mehrabian and Russell's S-O-R framework in the 'retail' context. This model of Mehrabian and Russell (1974) is discussed in detail later on in this study.

At this time a few studies on the retail environment and its effect on consumers' behaviour had been conducted, but the environment-behaviour relation was not conceptualized properly before Donovan and Rossiter's work in 1982. Though a few retailers had claimed that many environmental cues (e.g. layout and design etc.) affected consumers' shopping experience (Wysocki, 1979), their claim was not supported by any empirical evidence until the study of Donovan and Rossiter (1982). After the successful application of the Mehrabian and Russell (M-R) theory in the work of Donovan and Rossiter (1982), there were subsequently a number of studies in which researchers applied the M-R theory in the retail context (e.g. Hunter and Mukerji, 2011; Quartier, 2011; Kim and Lenon, 2010; Eroglu et al., 2003; Sherman et al., 1997; Donovan et al., 1994).

Since the development of conventional retail stores (Brick Environment), the retail atmosphere and its effect on shoppers has remained the primary interest of marketers and academic researchers (Turley and Milliman, 2000). Since 1982, as discussed above, researchers have done a lot in this area and explored many surprising effects of retail environments on shoppers' behaviour. Some of these researchers used a holistic approach to study the whole environment and its effect on shoppers (e.g. Donovan and Rossiter, 1982), while some others adopted a micro level approach to measure the effect of any/a few environmental cue(s), such as Music, Lighting and Colour, on different variables affecting consumer behaviour in the retail environment. For instance, Milliman (1982) measured the impact of music tempo on traffic pace, sales volume and music awareness; Bateson and Hui (1987) examined the effect of crowding on the approach/avoidance behaviour of shoppers; Lyer (1989) cited in Kalla and Arora, 2011) studied the relationship between a store's layout and impulse purchase behaviour of shoppers; Machleit et al. (1994) observed the relationship between crowding and consumer satisfaction; Reddy et al. (2011) studied the role of in-store lighting in contributing to consumer satisfaction.

The development in information technology is surprising; for example, Haeckel (1998) foresaw that these frequent developments in technology and the internet would definitely change the human cognition process. This development in technology made it possible for shoppers to purchase almost everything from their homes through the internet. The internet changed the concept of retailing, making online shopping possible and a simple process. Eroglu et al. (2001) said that online retail stores' environment (Click) affects the sequence of behaviours of shoppers through the webpage in the same way that a traditional retail store's environment (Brick) does. Eroglu et al. (2001) were pioneers who did research in the area of the online retail environment and later a huge number of researchers followed them and confirmed their propositions (Koo and Ju, 2010; Hunter and Mukerji, 2008; Chang and Chen, 2008; Wu et al., 2008; Price-Rankin, 2004; Eroglu et al., 2003). However, some of the environmental cues of a brick environment are not present in the click environment, e.g. olfactory cues. This development of web/online retail stores provided academic researchers and marketers with a wide field for new research (Manganari et al., 2009).

Although research on the impact of online retail stores' environments is not as old as that conducted for tradition retail stores, since it has only been practised for the past sixteen years; there is still a lot of empirical evidence showing that the environment of online retail stores has a significant impact on consumers' behaviour. Hunter and Mukerji (2011) argued that researchers applied a number of studies of brick environments into click environments and found that the online stores' environmental cues affect the shoppers' behaviour in the same way as the conventional stores' environmental cues do. Likewise, the M-R theory, which provides the S-O-R framework, has also been tested in many studies to prove the impact of an online store's environment on online shoppers. After the conventional stores studies, the M-R model has been applied successfully in a wide range of studies of online retail stores' environments (e.g. Manganari et al., 2011; Wu et al., 2008; Eroglu et al., 2003).

B. *The Background and History of Virtual Worlds:*

According to Nood and Attema (2006) the virtual world is not a novel concept as it was regarded in the past; in fact, they see it as old as "Dreaming". There have been two kinds of worlds from the beginning: a primary world and a secondary world (Auden, 1968). The primary world is the world in which a person can feel or see by using sensory organs, whilst the secondary world is the world of a person's imagination. As a species, Man always, consciously or unconsciously, imagines many things in his/her mind. Computerized virtual worlds (VWs) were initially played only as 3D video games. VWs are also known as Massively Multiplayer Online Games and different terms are used synonymously, e.g. Massively Multi-Player Online Role Playing Games (MMORPGs), Multi-User Online Virtual Environments (MUVEs), and Networked Virtual Environments (NVEs). Instead of using virtual worlds just for entertainment purposes, users are now using them for social interaction amongst one another in their daily routines (Wyld, 2010). The number of users of these worlds is increasing day by day and it seems that Gartner's (2007) prediction will become true that the number of VW users will reach one billion by 2018. According to DFC intelligence the world's gaming market was worth $67 billion in 2012 and it is expected to grow to $82 billion over the subsequent five years. In the past, video games were less interactive, single player oriented and users were of a young age. Nonetheless, now virtual worlds are highly interactive, multi-players can interact instantaneously and the users belong to groups of all ages (Adolph, 2011; ITU Telecommunication Standardization Bureau, 2011).

Today's virtual worlds are more developed, interactive and collaborative; therefore, there are many opportunities for different disciplines of life (Barnes and Mattsson, 2008). 3D virtual worlds are commonly categorized into two forms, game oriented and free form virtual worlds (Bainbridge, 2007). Game oriented virtual worlds, e.g. World of Warcraft, are only used for gaming purposes. Avatars (an electronic body to represent users in virtual worlds) are bound to wear some specific items to play within that environment. In game oriented virtual worlds users usually play with computer-controlled characters and try to win the levels of the game just for pleasure or entertainment. For this purpose users might have to purchase some virtual items to make them powerful in the game to win the level. The free form of virtual worlds (e.g. Second Life) are totally different in nature, as users are not there to play games but are free to perform most of their real life activities within the virtual world. Free form VWs are more similar to the real world; they are also known as open virtual worlds (Messinger et al., 2008). People buy and sell different kinds of applications, such as virtual apparels for their avatars, lands, islands, virtual vehicles and many more items. They can do many activities within these virtual worlds that are more simulative to the real world.

C. *Three Dimensional (3D) Retail Stores:*

Three dimensional (3D) retail stores, also called 3D virtual reality retail (VRR) stores, provide a new and innovative way of shopping, full of opportunities for both retailers and shoppers (Vrechopoulos et al., 2009). These 3D stores, or VRR stores, are available in 3D virtual worlds. Virtual reality retail stores have given a new concept to the retail industry. Virtual reality retail (VRR) stores are present within these 3D VWs, some of which offer virtual goods for free and some that are really costly. The virtual world has provided businesses with a new opportunity to market and sell their products. Virtual worlds (e.g. Second Life) are offering an alternative, improved, and quite potential medium to the consumers where they can shop for their virtual lives by paying with virtual money (e.g. Linden Dollars in Second Life). In the same way consumers can also buy for their real lives, but this is just at an introductory stage at the moment (Vrechopoulos et al., 2009).

Unlike web retail stores where many components of traditional retail environments are absent, e.g. Social Factor, the VRR stores that are made up of computer graphics that provide a real world simulated environment. VRR stores, like real world stores, are constructed with walls, colours, lighting, in-store music, floors, shelves, layout and design. One of the components of a retail store's environment is social presence, i.e. the presence of other customers or employees.

VWs are providing real world retailers with a unique opportunity to set up their retail business inside these virtual stores (Haenlein and Kaplan, 2009), but the concept of retailing in VWs is quite different to that of traditional web 2.0 technology. If traditional online stores are analysed, many discrepancies will be found; for instance, the images of products placed on traditional online stores are not true representations of the real product (Keeney, 1999). Moreover, whilst visiting traditional online stores there is a feeling of loneliness and inadequate interaction with other customers (Wang et al., 2007). This is not the case with VWs as they provide the customers with a simulated real environment. The retail stores are built within a 3D electronic environment where all the products are 3D electronic objects that closely resemble the real world product. Moreover, visitors of VRR stores can also interact with one another (Haenlein and Kaplan, 2009). Virtual reality retail stores are the most appropriate representation of real world stores, which could enhance a company's branding and advertising campaigns. It has also been proven through previous studies that 3D object placement in VWs has a very positive impact on users' intentions to purchase the same product in real life (Schlosser, 2006: 2003).

Kukreja and Humphreys (2014) argue that in traditional online stores, goods or services were shown on two dimensional flat interfaces where shoppers are not able to see a 3D view of the product. Moving inside traditional web stores is known as 'scrolling down or up'. Kukreja and Humphreys (2014) determined that 3D virtual reality retail stores are a substitute for 2D web stores. They are far better than web store and shoppers can move around the stores by walking, flying or running with the help of their avatars. There is none of the navigational difficulty in 3D VRR stores which shoppers face in traditional web retailing (Kukreja and Humphreys, 2014).

D. Research in 3D retail environment:

As aforementioned, today's world has another retailing channel in the form of 3D virtual reality retail stores, which are present in VWs. Not only do retailers have more options to market their products now, shoppers have multichannels to shop in too. Therefore, all these available shopping mediums should be examined from a consumer's perspective (McGoldrick and Collins, 2007).

There is a surfeit of research on traditional and web retail environments and how they affect a varied range of consumers' emotions and behaviours. However, in the case of VRR environments the research is still in its infancy (Hassouneh and Brenjman, 2015; Krasonikolakis et al., 2011; Vrechopoulos et al., 2009). These VWs have existed since 2003 (Shen and Eder, 2009), but remain ignored by retail environment researchers in terms of parallel research in VRR environments to find out how this environment affects the behaviour of virtual shoppers. To date, there are only three known studies in the context of 3D VRR stores atmospherics (Hassouneh and Brenjman, 2015; Krasonikolakis et al., 2011; Vrechopoulos et al., 2009); however, there is only one known research investigating the effect of the layout of 3D virtual reality retail stores on shoppers' behaviours, with a result that found no effect at all (Vrechopoulos et al., 2009). Vrechopoulos et al. (2009) opened the door for future exploration of VRR stores' environmental cues other than layout and design, such as crowding, sounds and store theatrics.

The objective of this research is to generate a list of stimuli which could be a part of 3D VRR store environments and affect shoppers' emotions and behaviour. There are two existing studies which made an attempt to generate a list of stimuli (Hassouneh and Brengman, 2015; Krasonikolakis et al., 2011). These two studies proved that there is a need to investigate the effect of 3D VRR store atmospherics on shoppers' behaviour. However, these studies defined the environment of 3D VRR stores but did not specify the different environmental cues. Moreover, Krasonikolakis et al. (2011) also did not mention the usage of these atmospherics in VRR stores. In this research a list of stimuli of 3D VRR stores will be generated, adapting Rosenbaum and Massiah's (2011) model into an expanded servicescape model.

This study is based on previous empirical work done in the context of traditional and web retail environments (Manganari et al., 2011; Ward et al., 2007; Eroglu et al., 2001; Donovan et al., 1994; Donovan and Rossiter, 1982), where researchers investigated how different retail environments affected shoppers' behaviour. This study is specifically parallel to the work of Eroglu et al. (2001) and aims at proposing a model to measure the effect of VRR stores' environmental cues on shoppers' behaviours.

A Stimulus-Organism-Response (SOR) model (Mehrabian and Russell, 1974) and an expanded servicescape model (Rosenbaum and Massiah, 2011) are adapted to provide a 3D servicescape model to measure 3D VRR environments. Although there is one known previous study (Vrechopoulos et al., 2009) in which researchers tried to investigate the effect of the layout of 3D VRR stores on shoppers' behaviours, they did not adapt Mehrabian and Russell's (1974) S-O-R model. Moreover, as argued by Lam (2001), a majority of the research investigating the effect of retail environments adapted an M-R (1974) model. Moreover, Vrechopoulos et al. (2009) also did micro-level research to investigate the effect of one environmental cue, such as layout and design, but the current research adapts a macro level approach to investigate the complete environment of 3D VRR stores and its effect on shoppers' behaviours, taking an S-O-R model as a basic framework.

This section has provided an extensive review of virtual worlds and virtual reality retailing. A gap in the research has also been discussed, which leads towards the development of a conceptual model for this research on the basis of an extensive theoretical review. The next chapter presents a conceptual model for this research, which is based on the up-to-date review of the existing servicescape models. Mehrabian and Russell's (1974) affect model is discussed along with Rosenbaum and Massiah's (2011) expanded servicescape model. The conceptual model was developed by adapting the M-R (1974) affect model, an expanded servicescape model and through interviews.

III. The Conceptual Model

A. *Up-to-date Review of Servicescape Models*

To develop the proposed conceptual model of this study all well-known servicescape studies within the retail environment were considered. An attempt is made in this study to offer a review of all known contemporary research within the field since 1980. Research between 1980 and 1990 within the field of service environment and its effect on human behaviour are few in number. Among these few, the majority of the studies were done within the North American region (e.g., Milliman, 1986; Bellizzi et al., 1983; Milliman, 1982; Donovan and Rossiter, 1982; Russell, 1980), and one in Australia (Amato and McInnes, 1983). During that time period research was conducted using different methods; for instance, some research was self-reported (Russell, 1980), by description (Donovan and Rossiter, 1982), and through field studies (Milliman, 1986, 1982; Amato and McInnes, 1983). There were some other studies between 1986 and 1991 that conducted research into the retail environment and its effect on shoppers' behaviours, for example Bateson and Hui (1987) studied the effect of crowding within the service environment, while Bawa et al. (1989) investigated the effect of store environments on brand loyalty, and Lyer (1989) focused on the store environment and its effect on unplanned purchasing. However, these studies ignored the Mehrabian and Russell (1974) model. Yalch and Spangenberg (1990, 1988) conducted the only two studies during this time period (1986 and 1991) that adapted the Mehrabian and Russell (1974) affect model. These two studies were also conducted in North America, and were field studies.

One of the factors that caused a lack of studies in this area before 1990 was that researchers did not know the importance of the retail environment and its effect on shoppers' behaviour (Yalch and Spangenberg, 1988). Hence, Yalch and Spangenberg (1990) pointed towards the retail environment and its great effect on consumers' behaviours, and called on other researchers to explore the environmental cues of retail environments other than music and crowding, which had already been investigated in the past (Yalch and Spangenberg, 1990, 1988; Milliman, 1986, 1982). This call for more research in this area and motivated the researchers. During 1991 and 1999, sixteen research papers could be found focusing on the retail environment and its effect on consumers' behaviours (Kearney et al., 2007). Among these sixteen studies, a majority of them were again from the North American Region (Spangenberg et al., 1996; Wakefield and Blodgett, 1996; Herrington and Capella, 1996; Dube et al., 1995; Gulas and Schewe, 1994; Wakefield and Blodgett, 1994; Yalch and Spangenberg, 1993; Baker et al., 1992; Bellizzi and Hite, 1992; Kellaris and Kent, 1992). Out of the rest of them, four studies were from Europe (Foxall and Greenley, 1999; Kenhove and Desrumaux, 1997; Spies et al., 1997; Hui and Bateson, 1991); one from Australia (Donovan et al., 1994) and one from Hong Kong (Tai and Fung, 1997). During this time period the trend to use field experiments increased and nine out of sixteen studies (Foxall and Greenley, 1999; Tai and Fung, 1997; Kenhove and Desrumaux, 1997;Spies et al., 1997; Herrington and Capella, 1996; Wakefield and Blodgett, 1996;Donovan et al., 1994; Gulas and Schewe, 1994; Yalch and Spangenberg, 1993), focused on field studies/ field experiments; however, the other seven studies were done in laboratories (Spangenberg et al., 1996; Dube et al., 1995; Wakefield and Blodgett, 1994; Kellaris and Kent, 1992; Baker et al., 1992; Bellizzi and Hite, 1992;Hui and Bateson, 1991) A majority of the researchers (Herrington and Capella, 1996; Dube et al., 1995; Gulas and Schewe, 1994; Yalch and Spangenberg, 1993; Kellaris and Kent, 1992) again followed the footsteps of previous researchers and investigated the same environmental cue, music, which had been studied many times even before 1991 (Yalch and Spangenberg, 1998, 1998; Milliman, 1986, 1982). There are only a few studies, during this time period, which investigated the effect of environmental cues other than music. For instance Bellizzi and Hitte's (1992) study on colours, Spangenberg et al.'s (1996) on olfaction, and Spies et al.'s (1997) on two environmental cues at the same time: lighting and colours.

Studies during this time period also adapted Mehrabian and Russell's (1974) affect model to investigate the effect of retail environments on shoppers' behaviours. However, Foxall and Greenley (1999) were the first researchers who considered dominance as a mediating variable between environment and shoppers' behaviours, but finding had a vague relationship between stimuli and behaviour. Later, this was confirmed by Gilboa and Rafeli (2003) as they indicated that dominance was the weakest element in the organism dimensions of the M-R (1974) model, and could be ignored in future studies. Between 2000 and 2007, nineteen more studies are known to have been done (Kearney et al., 2007). These nineteen only include those studies that were done in brick and mortar environments and not in online retail environments. Thirteen (majority) studies were again done in the North American region (Morin et al., 2007; Mattila and Wirtz, 2006; Spangenberg et al., 2006; Spangenberg et al., 2005; Babin et al., 2003; Chebat and Michon, 2003; Hightower et al., 2002; Chebat et al., 2001; Dube and Morin, 2001; Mattila and Wirtz, 2001; Summers and Herbert, 2001; Machleit et al., 2000, Yalch and Sapangenberg, 2000), three studies were done in Europe (Newman, 2007; Bigne et al., 2005; Reimer and Kuehn, 2004), one in Israel (Gilboa and Rafaeli, 2003), one in Singapore (Wirtz, Mattila, and Tan, 2000) and one in Australia (Sweeney and Wyber, 2002). Ten out of nineteen studies were done in field experiments, which is almost half of the total (Newman, 2007; Morin et al., 2007; Machleit et al., 2000; Spangenberg et al., 2006; Bigne et al., 2005; Reimer and Kuehn, 2004; Chebat and Michon, 2003; Hightower et al., 2002; Dube and Morin, 2001; Mattila and Wirtz, 2001). The rest of the studies were done within the laboratory environment. Though music remained the most frequently studied environmental cue even during this time period (e.g. Mattila and Wirtz, 2001; Yalch and Sapngenberg, 2000), olfaction (Spangenberg et al., 2005 and 2006; Chebat and Michon, 2003; Mattila and Wirtz, 2001), colours (Chebat and Morin, 2007), and lighting (Summers and Herbert, 2001) also got more attention compared to previous studies.

Until today, a majority of the research in retail environments has been conducted in North America and Europe. There are just a few studies done in other parts of the world, e.g., Malaysia, China and India.

B. Servicescape Frameworks

Serivcescape was first discussed by Kotler (1974) as a store atmospheric, and he argued that it is the store environment that is built to influence shoppers' behaviours so that sales can be increased. Bitner (1992) also defined servicescape as possessing all the physical factors of stores that are controlled by retailers in order to enhance or constrain customers' and employees' emotions and behaviours. Servicescape is also defined as possessing all the physical factors of stores that facilitate customers' shopping and communicating when in the store (Bitner and Zeithaml, 2003).

Servicescape is given a high level of importance in the literature on building customers' perceptions and expectations regarding the service being delivered (Grewal et al., 2003; Baker et al., 2002; Bitner, 1992). There are many frameworks to measure the service environments (Kearney et al., 2007) and these servicescape models are helpful in evaluating, assessing, measuring and understanding store environment and their atmospheres (Reimer and Kuehn, 2004; Turley and Milliman, 2000; Gulas and Schewe, 1994; Kellaris and Kent, 1992). Well known frameworks that measure the service environment are: the S-O-R model presented by Mehrabian and Russell (1974), cognitive theory by Lazarus (1991) and the servicescape model by Bitner (1992). Untill today, the most frequently adapted frameworks that measure the retail environment come from Mehrabian and Russell's (1974) S-O-R framework and Bitner's (1992) servicescape model.

The main difference between M-R and Lazarus' cognition theory is that an M-R model adapts emotions to a cognition approach, whereas Lazarus's (1991) cognition theory supports the application of cognition to an emotional approach (Bigne, et al., 2005; Chebat and Michon, 2003). For a long time researchers supported both sides (Lin, 2004). Bigne et al. (2005) supports Mehrabian and Russell's (1974) emotions to cognition approach, whilst Chebat and Michon (2003) support Lazarus's (1991) cognition to emotions approach. However, arguments from both sides have research limitations; hence more research was called for by Bigne et al. (2005).

This researcher is supporting Mehrabian and Russell's (1974) emotions-cognition approach, with a strong rationale outlined below.

C. The Mehrabian and Russell Affect Model

In 1974 Albert Mehrabian and James A. Russell proposed an affect model in their book titled 'An Approach to Environmental Psychology'. That affect model went on to become well respected in the field of environmental psychology and marketing. Researchers adapted it again and again in their research. Mehrabian and Russell's (1974) affect model proposed that every built or physical environment affects human behaviour through the intervening variables of emotions. This model contains three dimensions known as stimuli (S), which include the environment and all the environmental cues; organism (O), which here contains three emotions, which are pleasure, arousal and dominance; and a response (R) dimension, which comprises of human behaviour (approach and avoidance). That is why this M-R model is often called an S-O-R model. This approach from environmental psychology has been adapted in different studies to measure the specific environment and its effect on human behaviour (Harrell and Hutt, 1976; Lutz and Kakkar, 1975; Belk, 1974). However, it is not known if it was adapted as an overall framework in a retail setting until Donovan and Rossiter adapted the M-R affect model in 1982. An initial illustration of Mehrabian and Russell's (1974) model is given below in Figure 1.

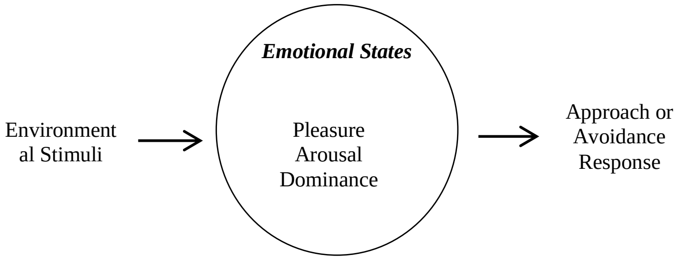


Fig. 1. Mehrabian and Russell affect model. Adapted from *An Approach to Environmental Psychology* (p. 8) by A. Mehrabian., & J. A. Russell, 1974, Cambridge, MA: MIT Press. Copyright 1974 by The Massachusetts Institute of Technology. Adapted with permission

The rationale behind adapting the Mehrabian and Russell (M-R) model is that not only does it provide an appropriate framework to base the measurement of the retail environment and shoppers' emotions and behaviours on (Quartier, 2011), but it also helps in measuring possible emotional responses by combining the basic PAD emotional dimensions (Graa and Dani-elKebir, 2011). It is also claimed that this model could be used to measure the effect of any built environment (Russell, 1980; Russell and Pratt, 1980; Mehrabian and Russell, 1974). Researchers have already adapted this model to measure both traditional retail environments: brick and mortar and click and mortar (Quartier, 2011; Kim and Lennon, 2010; Donovan and Rossiter, 1982). Therefore, in this research an M-R affect model is assumed to provide a base to measure 3D VRR environments and shoppers' behaviours within them.

The M-R model is adequate to measure emotions, but it is weak in the stimulus taxonomies area and researchers need to develop new stimulus taxonomies to measure relationships between intervening variables (Newman, 1997; Donovan and Rossiter, 1982). Mehrabian and Russell (1974) argued that different environments have different environmental cues and each environment is different from the other. Therefore, a stimuli section of the S-O-R framework was left vague as this dimension needs more experimental research to generate the taxonomy of the specific environment. Though researchers in the past developed different models to measure physical settings and generate stimulus taxonomies for different environments as discussed above, Bitner (1992) coined the term 'servicescape' for the first time to cover this area specifically. Later, other researchers developed other models to measure environments and the stimulus taxonomy for those specific environments (Williams and Dargel, 2004 etc.). Bitner (1992) separated environmental cues into three dimensions: ambient cues; spatial and functional cues; and signs, symbols and artifacts. However, Bitner's (1992) servicescape framework itself originated from ecological theory, which was presented in the early 1900s by Darwin and later, gave the foundation to Barker's (1968) study in the field of environmental psychology (Stokols, 1972).

Bitner's (1992) servicescape model has been adapted by many researchers; however, it has not been without its problems. Later, in 2011, Rosenbaum and Massiah (2011) conducted a contemporary review of all the existing servicescape models and proposed another servicescape framework, which is based on Bitner's (1992) original servicescape model. This model is also known as an Expanded Servicescape Model. Rosenbaum and Massiah's (2011) expanded their servicescape model to take into consideration the study of Proshansky (1978). They proposed that the service environment not only includes physical dimensions but also comprises social, socially symbolic and natural dimension, and all these dimensions influence shoppers' and employees' behaviour.

As a result, an 'expanded' servicescape model (Rosenbaum and Massiah, 2011; pp. 473) is adapted in this research in an attempt to define the stimulus dimension of an M-R (1974) model. This expanded servicescape model illustrates four environmental dimensions: physical, social, socially symbolic, and natural stimuli. Researchers (Rosenbaum and Massiah, 2011) illustrated in this model that merely objective, measureable and managerially controllable environmental cues are not the part of servicescapes. However, social, socially symbolic, and natural environmental cues, which are not controllable by retailers, are part of servicescapes. The aim is to adapt this model to fit a VRR environment and attempt to define VRR stores' environmental stimuli.

D. Developing a 3D Servicescape Model

Rosenbaum and Massiah's model was presented in the context of a brick and mortar retail environment, and although a 3D VRR environment has many similarities with a brick and mortar retail environment, it also differs in many ways. Shoppers can fly and teleport in VRR store environments, which is not possible in physical brick and mortar environments (Vrechopoulos et al., 2009). As mentioned earlier, to define the stimulus dimension of an M-R model in this research, Rosenbaum and Massiah's model will be adapted, although it is not completely compatible with a 3D retail environment because it is a simulated environment. A 3D retail environment has different features, and it is an electronic environment that a customer enter through your electronically simulated body, called an avatar, rather than with a customer's real bodies. Therefore, assuming all these differences, this research needs to alter Rosenbaum and Massiah's (2011) expanded servicescape model through detailed interviews with university students. Eight participants were invited to attend a session in a designated research office that was suitable for conducting interviews, and they were invited to attend at different times and on different days. Four female and four male participants took part in this research, their ages being between 21 and 33 years. Participants were, initially, informed about virtual worlds, their usage, benefits and how to use them. Later on they were requested to make their own accounts in Second Life. Then they were requested to log in by using 'Second Life Viewer', which was available to them on a computer in the research office.

The computer in the research office was tested for its ability to run Second Life and whether the internet speed was able to support its smooth running. If the internet speed had dropped at any time when a participant was using Second Life the session would have been closed; in this case the participant would have been excused and a different student invited in to participate in the experiment. If a participant encountered any difficulties whilst experiencing Second Life, they would have had a negative experience due to time wasting and being in an irritating situation. However, during these eight detailed sessions to define the retail environment of 3D VRR stores by adapting an expanding servicescape model as an initial framework, not a single problem was faced regarding internet speed. All the sessions were successful and raised much interest amongst those willing to participate. Each participant received a 10 to 15 minute explanation of virtual worlds, especially those in Second Life. They were also given a brief outline of the research. It was not explained to the participants that the purpose of these sessions was to define the environment because it was assumed that if the participants were informed of this, later on after experiencing the 3D VRR environment when they were asked about the environmental cues they experienced in Second Life they would have been more conscious of them. This may have resulted in results bias. The participants were allowed to enter into the environment freely, and later on they were asked different detailed questions about being in the environment. Participants were requested to experience 3D VRR stores for at least 20 minutes, but there was no maximum time limit. Participants were allowed to leave the session whenever they wanted to, but none of the eight participants left the session before it was completed. They were offered tea or coffee, biscuits and sandwiches at the end of the session.

Participants were initially asked about their overall experience within Second Life. Five out of eight participants said it was really fascinating and they had enjoyed it. Moreover, none of the participants said they had wasted their time as at least they had experienced something new. It was noticed that those three participants who did not find Second Life a fascinating simulated world were not technology oriented. It is therefore also possible that they did not find it fascinating because they were not able to enjoy and experience all the features of Second Life.

After this, all eight participants were individually asked about virtual reality retail stores. They were asked what they saw in the VRR stores and how they compared VRR stores to physical stores (Brick and Mortar). Most of the participants (6 out of 8) said they found VRR stores to be more interesting and less boring. During this question and answers session, participants named many cues that they had experienced in the VRR store environments. All the environmental cues that were mentioned in Rosenbaum and Massiah's (2011) servicescape model and that were experienced by participants were marked on the model by the researcher. Later on, all those cues that were experienced and mentioned by the participants but were not present in Rosenbaum and Massiah's

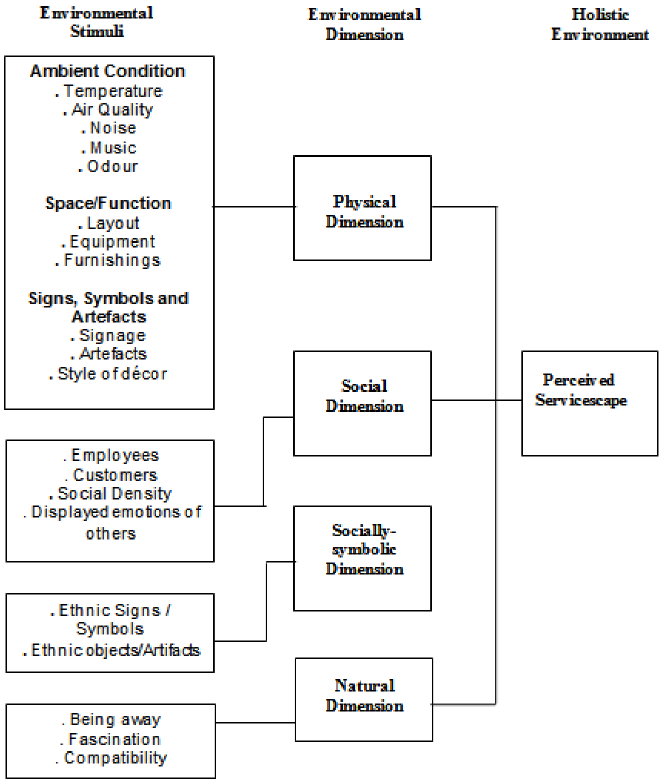


Fig. 2. An Expanded Servicescape Model for Understanding Four Environmental Dimensions of the Retail Environment. Adapted from "An expanded servicescape perspective" by M. S. Rosenbaum., & C. Massiah, 2011, *Journal of Service Management*, 22(4), p. 473. Copyright 2011 by Emerald Group Publishing Limited. Adapted with permission

model were added into the model, for example: being able to teleport, virtual air, flying etc. Details of these cues are given in the Stimulus section of this model. The purpose of these interviews was to confirm the stimuli dimension of the conceptual model. The finalized conceptual model is given in Figure 3. Participants agreed that the concept of an electronic virtual world and a virtual reality retail store was quite good. Although it is easy to use Second Life, participants still said they needed some practice before they became able to use all of the features that would make their shopping process easier.

- *Stimulus (S):*

According to M-R (1974), the stimulus dimension of the S-O-R framework is the one that affects human emotions, and the effect on emotions further leads to changes in behaviour. In the context of VRR stores in this conceptual model, it is assumed that stimulus is the sum total of all the environmental cues that are audible and visible to virtual shoppers. There are many environmental cues in VRR stores that are absent in traditional retail stores; for example, avatars in VRR stores can fly around the store (Vrechopoulos et al., 2009). There are some environmental cues in VRR stores that are similar to those in a traditional retail environment and absent in traditional online retailing. For example, Eroglu et al. (2001) determined that in contrast to the VRR environment an online retail environment lacks a visible presence of other shoppers and employees.

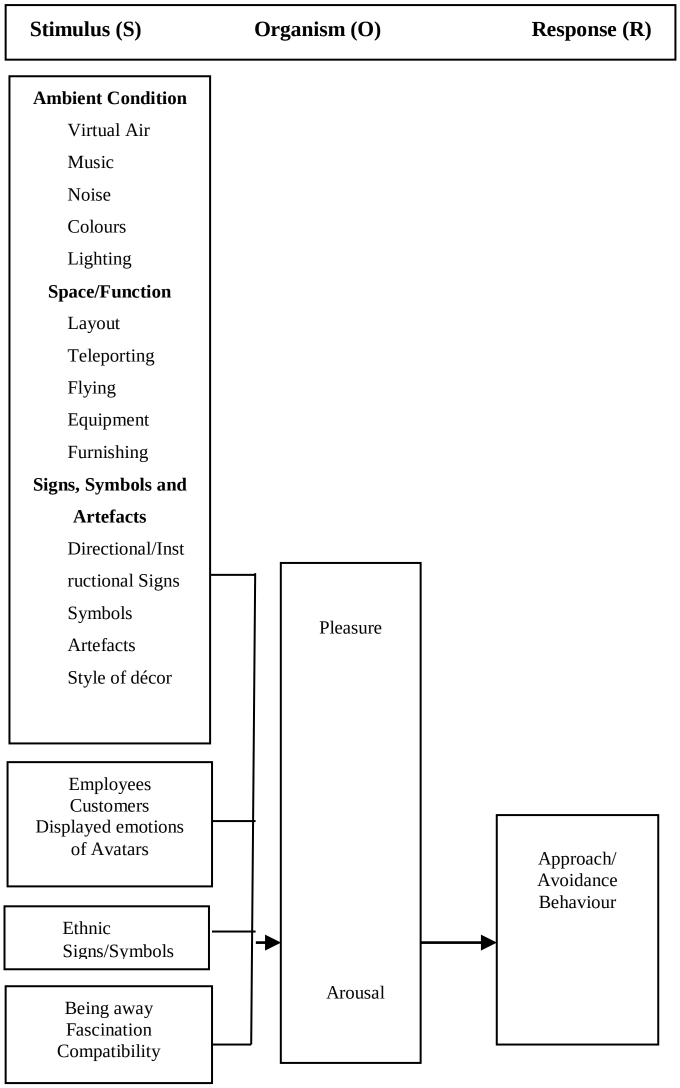


Fig. 3. The Proposed Conceptual Model of Shoppers' Behaviour in 3D VRR Store

It is obvious and easy to understand that a VRR store environment lacks some characteristics of a traditional retail store environment, such as three (sense of touch, sense of smell, sense of taste) of the five sensory appeals (Vrechopoulos et al., 2009). However, there is perhaps no doubt that a VRR store environment is far better than that of the traditional online retail stores because it more closely replicates the environment of a traditional brick and mortar retail environment (Vrechopoulos et al., 2009). Ostensibly, a VRR store's whole environment is limited to the computer screen, like that of traditional online retail stores (Eroglu et al., 2001), but in reality it is more than that. As explained earlier, in a VRR retail environment, shoppers, through their avatar, can walk around the store, can see, chat and communicate with other avatars, thus proving it to be a pure replication the environment of traditional brick and mortar retail stores (Vrechopoulos et al., 2009). Hence, the classification of the traditional online retail store environment is not applicable to that of VRR stores, for which an alternative taxonomy is necessary. Since the VRR store has more in common with the traditional retail store, so the taxonomy development should also be parallel to that of the traditional retail environment. Hence, an expanded servicescape model presented by Rosenbaum and Massiah (2011) is adapted in this research. Bitner (1992) proposed three dimensions of servicescape, which were physical, social and natural stimuli. However, the expanded servicescape model of Rosenbaum and Massiah (2011) consists of socially symbolic dimensions other than physical, social and natural stimuli.

Physical Dimension: Physical dimensions consist of three main environmental cues, which are ambience, space/function, signs, symbols and artifacts. Ambient conditions are temperature, air quality, noise, music and odour. In the past, there were studies in brick and mortar retail environment where researchers explored the effect of temperature and odour on shoppers' shopping behaviour (Ward et al., 2007; Spangenberg et al., 2006; Spangenberg et al., 1996; Michon et al., 2005; Spangenberg et al., 2005; Chebat and Michon, 2003; Mattila and Wirtz, 2001; Yalch and Spangenberg, 2000). However, there is no concept of temperature and odour in online web stores and hence there is no study found in this context. The same is the case with virtual reality retail store environment, which are 3D but computer based and hence users cannot experience any smell or temperature. Therefore, these two cues of ambient conditions were eliminated in the process of adapting this model in a VRR environment context. However, in virtual worlds (e.g., Second Life) there is a presence of virtual air and artificial simulated weather (summer or winter, wind or storm etc.), therefore the air quality cue is replaced by virtual air.

Space/Function consists of layout, equipment and furnishing. There are a few well known studies in brick and mortar (Nath, 2009; Ryu and Jang, 2008; Countryman and Jang, 2006; Li, 2004; Yalch and Spangenberg, 2000; Wakefield and Blodgett, 1996) and click and mortar (Manganari et al., 2011; Vrechopoulos et al., 2004) retail environments that have discussed stores' layout and its effect on shoppers' behaviour. This concept of store layout is present in 3D virtual reality retail stores as well, and one contemporary study has been found in this area too. Vrechopoulos et al. (2009) explored the different form of store layouts in 3D VRR stores environment and their effect on shoppers' behaviour. Krasonikolakis et al. (2011) have also considered layout as an important environmental cue of VRR stores. However, Vrechopoulos et al. (2009) argued that layout has no effect on consumers' behaviour in Second Life (SL) because of the shoppers' ability to fly and teleport in SL. Thus, teleport and flying is added to the section of Space/Function. MacKenzie et al. (2013) described the function of teleporting as moving an avatar from one part in the virtual world to another. Users' flying ability is enabled in Second Life and they can experience flying inside the store too. Vrechopoulos et al. (2009) argue in the results of their study that the layout of a VRR store might not have an effect because of the ability to fly. It suggests that this ability could affect shoppers' emotions and behaviour; therefore, flying is added in the conceptual model of this research to investigate its effect further.

Rosenbaum and Massiah (2011), while explaining their work, argued that retailers use signs to deliver general messages to customers, such as symbols and artefacts for communicating and decoration purposes in a retail environment. These environmental cues are also adapted in the conceptual model of this research as signs, symbols, and artefacts are seen and experienced by participants at the time of the model confirmation process. Hence, parallel to the set of propositions from Eroglu et al. (2001) and Bitner (1992) concerning the 3D servicescape, this research offers a number of propositions about 3D virtual reality retail environments. It is postulated that:

P1a: A 3D virtual reality retail store consists of physical dimensions

P1b: The physical dimensions of a 3D VRR store affect shoppers' approach/avoidance behaviour through the mediating variables of emotions (pleasure and arousal).

Social Dimension: The social dimension of an expanded servicescape model contains employees, customers, social density and the displayed emotions of others. Since there is no visible presence of employees and customers in traditional online retail environments, as explained by Eroglu et al. (2001), there is therefore no social density. That is why researchers focusing on an online retailing context have ignored the social dimension and its effect on shoppers' emotions and behaviours. However, in brick and mortar retail environments there is a wide concept of social factor and a large amount of research has been done in this area already (Lin and Liang, 2011; Dion, 2004; Wang, 2003; Baker et al., 2002; Sherman et al., 1997; Baker et al., 1994; Baker et al., 1992). Likewise, shoppers can see other avatars in virtual worlds while shopping in 3D VRR stores, they can also chat (by text) and Voice Over Internet Protocol (VOIP) with each other as well.

There is a concept of social factors in virtual worlds and it has been appreciated by the participants in past research (for example, Hassouneh and Brengman, 2011). Shopping in virtual worlds is perceived as more fun, with no crowds or rude people. However, none of the seven participants in this study experienced any employees inside the VRR stores, and the same was the case with Hassouneh and Brengman (2011). Hence, it was decided that researchers would act as employees in VRR stores during the main study, and other participants in the main study would be acting as customers to each other.

It is assumed that the presence of a social factor in virtual reality retail stores should have some effect on shoppers' emotions and behaviour. Additionally, research claims that there is an obvious display of emotions among avatars as they show their anger or pleasure through their faces (Lee et al., 2013). Although all these eight participants, in this research, did not notice any avatar's emotions, despite this Avatars' displayed emotions, has been adopted in the social dimension of this conceptual model by following Lee et al.'s (2013) research, As a result, the social dimension of the conceptual model includes customers, employees and displayed emotions of other avatars. Hence it is posited that:

P2a: A 3D virtual reality retail store consists of social dimensions

P2b: The social dimensions of a 3D VRR store affects shoppers' approach/avoidance behaviour through the mediating variables of emotions (pleasure and arousal).

Socially symbolic dimension: The socially symbolic dimension of the model consists of ethnic signs/symbols and ethnic objects/artefacts. These environmental cues of socially symbolic dimensions were considered as integral cues of the servicescape dimension in previous studies (Bitner, 1992). Bitner (1992) and Rosenbaum (2005) argued that managers try to influence shoppers' approach or avoidance behaviour through ethnic signs and symbols because these ethnic signs and symbols reflect affiliation with the shoppers. They could have either a positive or negative (if perceived negatively by customers) effect on shoppers' emotions and, consequently, on their subsequent behaviour. Since most of the participants noticed ethnic signs and symbols in VRR stores, ethnic signs and symbols were added in this conceptual model within the socially symbolic section (as shown above in figure 3). Therefore, it is posited that:

P3a: A 3D virtual reality retail store consists of socially symbolic dimensions.

P3b: The socially symbolic dimensions of 3D VRR store affect shoppers' approach/ avoidance behaviour through the mediating variables of emotions (pleasure and arousal).

Natural Dimension: The natural dimension of an expanded servicescape model consists of three stimuli: being away, fascination, and compatibility. Arguing about the first stimuli of this dimension, Rosenbaum and Massiah (2011) posited that being away does not really mean being physically away from one's place of existence. Rather it is the feeling of being away from one's work routine and experiencing a sense of relaxation and enjoyment. Jin and Bolebruch (2009), Wyld (2010) and Melancon (2011) are all of the opinion that virtual worlds are immersive enough to make a user forget about his or her real identity through the adoption of a virtual identity (Avatar). Therefore, it is assumed that virtual worlds give users the feeling of being away. This stimulus is also included in this model in the VRR store context. Since Virtual worlds also provide a fascinating environment (Melancon, 2011) so this stimulus is presumed to have an effect on shoppers' emotions and behaviours. VWs are a gift from technology and predominantly technology-oriented people are able to use them in the home environment. Compatibility is, therefore, an important stimulus to be investigated in the context of research into VRR stores' environments. That is why this third and last stimulus is also included in the natural dimension of this proposed model of shoppers' behaviours in VRR store environments. Hence, it is postulated that:

P4a: A 3D virtual reality retail store consists of natural dimensions.

P4b: The natural dimensions of 3D VRR stores affect shoppers' approach/avoidance behaviour through the mediating variables of emotions (pleasure and arousal).

- *Organism (O)*

In an S-O-R model, the organism dimension shows the human emotional state (M-R, 1974). Mehrabian and Russell (1974) posited that all human emotions could be placed into three categories, which they called Pleasure, Arousal, and Dominance, and are often referred to as the PAD dimensions. It is also hypothesised that each environment, including a retail environment, affects an individual's state of pleasure, arousal and dominance (Mehrabian and Russell, 1974).

This PAD dimension is orthogonal:

Pleasure - Displeasure

Arousal - No arousal

Dominance – Submissiveness

Later, in 1980, Russell and Pratt modified this S-O-R model and deleted the dominance dimension from the organism section of the model. They argued that dominance requires interpretation by the individual and could not be included in a situation where effective responses are required. Although initially the dominance dimension was included in much retail environment research (traditional and online), it was later excluded at the time of final studies (Quartier, 2011; Eroglu et al., 2001; Donovan and Rossiter, 1982). Hence, in this research the dominance element is also excluded from organism dimensions of the model. The rationale behind ignoring the dominance dimension in this research is Mehrabian and Russell's (1974) study in which they themselves did not pay much attention to dominance. Later, other researchers also supported it by ignoring and declaring dominance as a less beneficial dimension (Quartier, 2011; Sweeney and Wyber, 2002; Eroglu et al., 2001; Dube et al., 1995; Donovan et al., 1994; Amato and McInnes, 1983; Donovan and Rossiter, 1982; Russell, 1980).

In a majority of research results have supported both pleasure and arousal as an important predictor of behaviour; however, some research has also contradicted this (e.g. Kearney et al., 2007). Moreover, Donovan et al. (1994) found arousal insignificant in a pleasant retail environment, which contradicts Donovan and Rossiter's (1982) study of the retail environment. Along with many other reasons for this contradiction, it is also possible that it is due to a difference in sample size and methodological approach. Donovan and Rossiter (1982) used a small sample size, whilst Donovan et al. (1994) used a larger sample size alongside a field study. Kenhove and Desrumaux (1997) supported previous research (Donovan et al., 1994; Donovan and Rossiter, 1982) and argued that indeed Mehrabian and Russell's (1974) model fits well in retail settings. They confirmed that an organism plays an important role of mediator between the environment and shoppers' behaviour. However, previous studies (Donovan et al., 1994; Donovan and Rossiter, 1982) only stressed the importance of pleasure, while Kenhove and Desrumaux (1997) confirmed the importance of arousal as well. This result was later confirmed by Tai and Fung (1997) who found a partial relationship between pleasure and arousal in their study. Hence it is posited that:

P5: The pleasure and arousal state of virtual shoppers mediates the relationship between 3D VRR store environments and shoppers' behaviour (approach/avoidance).

P6: There is an interrelationship between pleasure and arousal as mediating variables when predicting the effect of a 3D VRR environment on shoppers' behaviour.

- *Response (R)*

Mehrabian and Russell's (1974) S-O-R model consists of a response dimension, which in a VRR store environment context shows how the shoppers' final response is affected by the VRR store environment through PAD (Pleasure-Arousal-Dominance) dimensions as an intervening variable (Quartier, 2011; Eroglue et al., 2001 & 2003; Sherman et al., 1997; Donovan and Rossiter, 1982). The response dimension consists of approach and avoidance behaviour. Approach behaviour reflects all the positive behaviours towards any particular retail environment in contrast to avoidance behaviour, which represents all the negative intentions/actions towards that particular retail environment. Therefore:

P7: Virtual shoppers' positive emotional states lead to approach behaviour in a 3D VRR store environment such as revisiting the store, spending more money, staying for longer etc.

P8: Virtual shoppers' negative emotional states lead to avoidance behaviour in a 3D VRR store environment such as leaving the store as soon as possible, never visiting it again, lower spending etc.

IV. CONCLUSION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. *Research Implications*

This research opens a number of new avenues for further investigation through the proposed model of shoppers' behaviour in a VRR store environment. As aforementioned, there is one known research (Vrechopoulos et al., 2009), which attempted to explore the VRR store environment and its effect on shoppers' behaviour. This research was at a micro level, where the researchers tried to explore the effect of layout on shoppers' behaviour in VRR store environments. However, the present research represents the first step to explore the VRR store environment through holistic (molar approach according to Quartier, 2011) approach where all the environmental cues of a VRR store environment are considered. Further, a systematic taxonomy development of VRR store environment is attempted through this proposed model that may prove to be the main step in theory building. It is thus assumed that this research opens ample opportunities in future for theoretical and empirical research into VRR store environments. This model, through its propositions and each of the relationship between variables, offers an avenue of further in-depth research to future researchers.

As described earlier a number of researchers (Quartier, 2011; Eroglu et al., 2001 and 2003; Donovan and Rossiter, 1982) in the retail environment area, who adapted the S-O-R model of Mehrabian and Russell (1974), argued that this (S-O-R) model is weak in its Stimulus (S) dimension and there is a dire need to focus on the taxonomy development stage. As it is clear that a VRR store environment provides a more simulating environment to the traditional retail environment (Brick and Mortar) than that of traditional online retail stores; hence taxonomy developed by traditional online retail environment researchers is not fully applicable to the VRR environment. Therefore, in this proposed model an expanded servicescape model (Rosenbaum and Massiah, 2011) is adapted in the Stimulus dimension of S-O-R model.

In a nutshell, this conceptual model proposes as a first step to theorise the environmental cues and shoppers' responses in a VRR store context holistically. As mentioned above, both VWs and virtual users are growing in number very rapidly; therefore, it is assumed that VRR stores will emerge as a new channel of retailing and attain greater attention from academics and practitioners. This vibrant area is open to be explored by researchers for their theoretical and methodological contributions.

B. *Future Studies*

This research filled the gap in the knowledge and provided a 3D servicescape conceptual model which defines the 3D VRR store environment in depth. However, the environment was defined through eight detailed interviews and this limits the extent to which it can be generalised. It is possible that there is no effect of some environmental cues on shoppers' emotions and behaviour. It is proposed that the adapted Stimulus dimension (Rosenbaum and Massiah, 2011) would be appropriate to measure a VRR store environment; however, further detailed research is needed in this area. Other moderating and intervening variables (e.g. involvement and atmospheric responsiveness, from Eroglu et al., 2001) can also be considered in future research. Further empirical research is necessary to confirm the effect of 21 environmental cues of the 3D servicescape conceptual model on shoppers' behaviour through the mediating variables of emotions. This research calls for quantitative research in a controlled lab environment where the propositions made in this paper should be tested.

REFERENCES

[1] Adolph, M. (2011). Trends in Video Games and Gaming. Retrieved from ITU-T Technology Watch Report. URLhttp://ocw.metu.edu.tr/pluginfile.php/10647/mod_resource/content/1/T23010000140002PDFE.pdf (accessed 12 August, 2015)

[2] Amato, P. and McInnes, I. (1983), ‘‘Affiliative Behaviour in Diverse Environments: A consideration of Pleasantness, Information Rate, and the Arousal-Eliciting Quality of Settings’’, *Basic and Applied Social Psychology,* Vol. 4, No. 2, pp. 109-122.

[3] Auden, W. H. (1968), *Secondary Worlds: Essays*. Random House.

[4] Babin, B. J., Hardesty, D. M. and Suter, T. A. (2003), "Color and Shopping intentions: The intervening effect of price fairness and perceived affect", *Journal of Business Research,*Vol. 56, pp. 541-551.

[5] *Bateson, J.E.G., and Hui, M.K.M. (1987), "A model for crowding in the service experience: empirical findings", The Services Challenge: Integrating for Competitive Advantage, pp. 85-89.*

[6] Bainbridge, W.S. (2007), "The Scientific Research Potential of Virtual Worlds", *Science,* Vol. 317, No. 5837, pp. 472-476

[7] Baker, J., Levy, M., and Greval, M. (1992), "An experimental approach to making retail store environmental decisions", *Journal of Retailing,*Vol. 68, No. 4, pp. 445-46.

[8] Baker, J., Parasuraman, A., Grewal, D. and Voss, G. (2002), "The Influence of multiple store environment cues on perceived merchandise value and patronage intentions", *Journal of Marketing,* Vol. 66, pp. 120-141

[9] Barnes, S. and Mattsson, J. (2008), "Brand Value in Virtual Worlds: An Axiological Approach", *Journal of Electronic Commerce Research*, 9 (3), 195-207.

[10] Barker, R. G. (1968), "Explorations in ecological psychology", *American Psychologist,* Vol. 20, pp. 1-14

[11] Bawa, K., Landwehr, J.T. and Krishna, A.A. (1989), "Consumer response to retailers' marketing environments: an analysis of coffee purchase data", *Journal of Retailing*, Vol. 65, No. 4, pp. 471-495

[12] Bellizzi, J. A., Crowley, E.A. and Hasty, R.W. (1983). "The effects of Color in Store Design", *Journal of Retailing,* Vol. 59, No. 1, pp. 21-45.

[13] Bellizzi, J.A. and Hite, R.E. (1992), "Environmental Color, Consumer Feelings, and Purchase Likelihood", *Journal of Psychology and Marketing,* Vol. 9, No. 5, pp. 347-63.

[14] Belk, R.W. (1974), "Application and Analysis of the Behavioral Differential Inventory for Assessing Situational Effects in Buyer Behavior", In *NA – Advances in Consumer Research, 01*, pp. 370-380

[15] Bigne, J., Andreu, L. and Gnoth, J. (2005), "The Theme park experience: An analysis of pleasure, arousal and satisfaction" *Tourism Management,* Vol. 26, pp. 833-844.

[16] Bitner, M. J. (1992), "Servicescapes: The impact of physical surroundings on customers and employees", *Journal of Marketing,* Vol. 56, pp. 57-71.

[17] Bitner, M. J. and Zeithaml, V. A. (2003), *Services Marketing integrating customers focus across the firm (*3rd ed), McGraw Hill, New York

[18] Centre for retail research (2013). "Retail in 2018 - Shop numbers, Online and the High Street", accessed from www.retailresearch.org/retail2018.php (accessed on 15 August, 2015)

[19] Chang, G. (2014), "Top Innovations that Changed the Worlds of Retail", *RIS Retail Info Systems NEWS,* accessed from http://risnews.edgl.com/retail-news/Top-Innovations-that-Changed-the-World-of-Retail94709 (accessed on 23 August 2015)

[20] Chebat, J.C., Chebat, C.G. and Vaillant, D. (2001), "Environmental Background music and in-store selling", *Journal of Business Research,* Vol. 54, pp. 115-123.

[21] Chebat, J. C. and Michon, R. (2003), "Impact of ambient odors on mall shoppers' emotions, cognition, and spending: A Test of Competitive causal theories", *Journal of Business Research,* Vol. 56, pp. 529-39.

[22] Countryman, C.C., and Jang, S. (2006), "The effects of atmospheric elements on customer impression: the case of hotel lobbies", *International Journal of Contemporary Hospitality Management,* Vol. 18, No. 7, pp. 534-545

[23] Craik, K. H. (1973). "Environmental Psychology", *Annual Review of Psychology*. Vol. 23, pp. 403-422

[24] Davis, L., Wang, S. and Lindridge, A. (2008), "Culture Influences on emotional responses to on-line store atmospheric cues", *Journal of Business Research,* Vol. 61, pp. 806-812.

[25] Dion, D. (2004). "Personal control and coping with retail crowding", *International Journal of Service Industry Management.* Vol. 15, No. 3, pp. 2250-263.

[26] Donovan, R.J., and Rossiter, J.R. (1982), "Store atmosphere: an environmental psychology approach", *Journal of Retailing,* Vol. 58, No. 1, pp. 34-57.

[27] Donovan, R.J., Rossiter, J.R., Marcoolyn, G., & Nesdale, A. (1994), "Store atmosphere and purchasing behaviour", *Journal of Retailing,* Vol. 70, No. 3, pp. 283-294.

[28] Dooley, J.A., Jones, S.C. and Iverson, D. (2012), "Web 2.0: an assessment of social marketing principles", *Journal of Social Marketing*, Vol. 2, No. 3, pp. 207-221

[29] Dube, L., Chebat, J.C. & Morin, S. (1995), "The Effects of Background Music on Consumers' Desire to Affiliate in buyer-Seller Interactions", *Psychology and Marketing,* Vol. 12, No. 4, pp. 305-319.

[30] Dube, L. & Morin, S. (2001), "Background music pleasure and store evaluation Intensity effects and psychology mechanisms", *Journal of business Research,* Vol. 54, pp. 107-113.

[31] Dunne, P.M., Lusch, R.F. and Griffith, D.A. (2002), *Retailing* (4th ed.), South-Western Educational Publishing, Mishawaka, IN, USA

[32] eMarketer (2014), "Total US Retail Sales Top $4.5 Trillion in 2013, Outpace GDP Growth", Retrieved from www.emarketer.com/Article/Total-US-Retail-Sales-Top-3645-Trillion-2013-Outpace-GDP-Growth/1010756 (accessed on 26 November 2015).

[33] Eroglu, S.A., Machleit, K.A., and Davis, L.M. (2001), "Atmospheric qualities of online retailing: a conceptual model and implications", *Journal of Business Research,* Vol. 54, pp. 177-184.

[34] Eroglu, S.A., Machleit, K.A., and Davis, L.M. (2003), "Empirical testing of a model of online store atmospherics and shopper response", *Psychology and Marketing,* Vol. 20, No. 2, pp. 139-50.

[35] Foxall, G. & Greenley, G. (1999), "Consumers' Emotional Responses to Service Environments", *Journal of Business Research,* Vol. 46, pp. 149-158.

[36] Gartner. (2007), "Gartner Says 80 Percent of Active Internet Users Will Have A "Second Life" in the Virtual World by the End of 2011", *Technology Research Gartner Inc.* Available from www.gartner.com/it/page.jsp?id=503861 (accessed 11 November, 2015)

[37] Graa, A., and Dani-elKebir, M. (2011), "Situational factors influencing impulse buying behavior of algerian consumer" *RRM,* Vol. 2, pp. 52-59.

[38] Gibbs, S. (2015), "Apple Pay launches in the UK: here's how to use it", *Theguardian*, Retrieved November, 2015 from www.theguardian.com/technology/2015/jul/14/apple-pay-launches-uk-how-to-use (accessed 13 November, 2015).

[39] Gilboa, S. and Rafaeli, A. (2003), "Store environment, emotions and approach behaviour: applying environmental aesthetics to retailing", *International Review of Retail, Distribution and Consumer Research,* Vol. 13, pp. 195-211.

[40] Graves, P. (2013), *Consumer.ology*. Nicholas Brealey Publishing, London

[41] Gulas, C. S. and Schewe, C. D. (1994), "Atmospheric segmentation: Managing Store Image with Background Music", In R. Acrol and A. Mithcell (Eds.) *Enhancing Knowledge Development in Marketing (*pp. 325-330). American Marketing Association, Chicago, IL

[42] Haeckel, S. H. (1998), "About the nature and future of interactive marketing", *Journal of Interactive Marketing*, Vol. 12, No. 1, pp. 63–71.

[43] Haenlein, M., and Kaplan, A.M. (2009), "Flagship brand stores within virtual worlds: the impact of virtual store exposure on real-life attitude toward the brand and purchase intent", *Researche et Applications en Marketing,* Vol. 24, No. 3), pp. 57-79.

[44] Hassouneh, D. and Brengman, M. (2011), "Shopping in Virtual Worlds: Perceptions, Motivations, and Behavior", *Journal of Electronic Commerce Research,* Vol. 12, No. 4, pp. 320-335

[45] Hassouneh, D., and Brengman, M. (2015), "Retailing in social virtual worlds: developing a typolog of virtual store atmospherics", *Journal of Electronic Commerce Research*, Vol. 16, No. 3, pp. 218-241.

[46] Harrell, G.D. and Hutt, M.D. (1976), "Crowding in Retail Stores", *M.S.U. Business Topics,* pp. 33-39.

[47] Herrington, J.D. and Capella, L.M. (1996), "Effects of Music in Service Environments: A Field Study", *Journal of Services Marketing*, Vol. 10, No. 2, pp. 26-41

[48] Hightower, R., Brady, M.K. and Baker, T.L. (2002), "Investigating the role of the physical environment in hedonic service consumption: an exploratory study of sporting events", *Journal of Business Research,* Vol. 55, pp. 697-707.

[49] Hunter, R., and Mukerji, D.B. (2011), "The role of atmospherics in influencing consumer behaviour in the online environment", *International Journal of Business and Social Science,* Vol. 2, No. 9, pp. 118-125.

[50] Hui, M.K. and Bateson, J.G. (1991), "Perceived control and the effects of crowding and consumer choice on the service experience", *Journal of Consumer Research,* Vol. 18, pp. 174-184.

[51] Jin, S.A., and Bolebruch, J. (2009), "Avatar-based advertising in second life: the role of presence and attractiveness of virtual spokespersons", *Journal of Interactive Advertising,* Vol. 10, No. 1, pp. 51-60.

[52] Kalla, S.M., and Arora, A. (2011), "Impulse Buying: A literature Review", *Global Business Review,* Vol. 12, No. 1, pp. 145-157.

[53] Kearney, T., Kennedy, A. Coughlan, J. (2007), "Servicescapes: A review of contemporary empirical research" Annual Frontiers in Service Conference, San Francisco, CA, pp. 58-88.

[54] Keeney, R.L. (1999), "The value of internet commerce to the consumer", *Management Science*, Vol. 45, No. 4, pp. 533-542

[55] Kellaris, J.J. and Kent, R,J. (1992), "The influence of Music on Consumers' Temporal Perceptions: Does Time Fly When You're Having Fun?", *Journal of Consumer Psychology,* Vol. 1, No. 4, pp. 365-376.

[56] Kenhove, P. and Desrumaux, P. (1997), "The Relationship between Emotional States and Approach or Avoidance Response in a Retail Environment", *The International Review of Retail, Distribution and Consumer Research,* Vol. 7, pp. 351-68.

[57] Kim, H., and Lennon, S.J. (2010), "E-atmosphere, emotional, cognitive, and behavioral responses", *Journal of Fashion Marketing and Management,* Vol. 14, No. 3, pp. 412-428.

[58] Kotler, P. (1973), "Atmosphere as a marketing tool", *Journal of Retailing,* Vol. 49, No. 4, pp. 48-64.

[59] Krasonikolakis, I. G., Vrechopoulos, A. P. and Pouloudi, A. (2011), "Defining, Applying and Customizing Store Atmosphere in Virtual Reality Commerce: Back to Basics?" *International Journal of E-Services and Mobile Applications,* Vol. 3, No. 2, pp. 59-72

[60] Kukreja, V.I. and Humphreys, D.W. (2014), "3D Virtual Store", *Google Patents*, Vol. 14, No. 169

[61] Lam, S.Y. (2001), "The Effects of Store Environment on Shopping Behaviors: a Critical Review'', *in NA - Advances in Consumer Research,* Vol. 28, eds. Mary C. Gilly and Joan Meyers-Levy, Valdosta, GA: Association for Consumer Research, pp. 190-197.

[62] Lazarus, R. S. (1991), *Emotion and Adaptation,* Oxford University Press, New York

[63] Lee, M., Kim, M., ad Peng, W. (2013), "Consumer reviews: Reviewer avatar facial expression and review valence", *Internet Research*, Vol. 23, pp. 116–132.

[64] Lin, I. (2004). "Evaluating a servicescape: the effect of cognition and emotion", *Hospitality Management,* Vol. 23, No. 2, pp. 163-178.

[65] Lin, J.C. and Liang, H. (2011), "The influence of service environments on customer emotions and service outcomes", *Manage Service Quality,* Vol. 21, No. 4, pp. 350-372.

[66] Li, J. (2004), *The effects of store physical environment on perceived crowding and shopping behavior*, Unpublished Doctoral Dissertation, Auburn University.

[67] Lutz, R. J. and Kakkar, P. (1975), "*The Psychological Situation As a Determinant of Consumer Behavior",* In NA – Advances in Consumer Research, 2, eds. Mary Jane Schlinger, Ann Abor, MI: Association for Consumer Research, pp. 439 -454.

[68] Lyer, E. S. (1989), "Unplanned purchasing: knowledge of shopping environment and time pressure" *Journal of Retailing,* Vol. 65, No. 1, pp. 40-57.

[69] MacKenzie, K., Buckby, S. and Irvine, H. (2013), "Business research in virtual worlds: possibilities and practicalities", *Accounting, Auditing and Accountability Journal,* Vol. 26, No. 3, pp. 352-373

[70] Machleit, K.A., Kellaris, J.J. and Eroglu, S.A. (1994), "Human versus Spatial Dimensions of Crowding Perceptions in Retail Environments: A Note on Their Measurement and Effect on Shopper Satisfaction", *Marketing Letters,* Vol. 5, No. 2, pp. 183-194.

[71] Machleit, K.A., Eroglu, S.A. and Mantel, P.S. (2000), "Perceived Retail Crowding and Shopping Satisfaction: What Modifies This Relationship?", *Journal of Consumer Psychology,* Vol. 9, No. 1, pp. 29-42.

[72] Manganari, E.E., Siomkos, G.J., and Vrechopoulos, A.P. (2009), "Store atmosphere in web retailing", *European Journal of Marketing,* Vol. 43, No. 9/10, pp. 1140-1153.

[73] Manganari, E.E., Siomkos, G.J., Rigopoulou, I.D. and Vrechopoulos, A.P. (2011), "Virtual Store Layout effects on Consumer Behaviour", *Internet Research,* Vol. 21, No. 3, pp. 326-346.

[74] Mattila, A.S. and Wirtz, J, (2001), "Congruency of scent and music as a driver of in-store evaluations and behaviour", *Journal of Retailing,* Vol. 77, pp. 273-89.

[75] Mattila, A.S. and Wirtz, J. (2006), "Arousal expectations and service evaluations", *International Journal of Service Industry Management,* Vol. 17, No. 3, pp. 229-244.

[76] McGoldrick, P.J., and Collins, N. (2007), "Multichannel retailing: Profiling the multichannel Shopper", *The International Review of Retail, Distribution and Consumer Research,* Vol. 17 No. 2, pp. 139–58.

[77] Melancon, J.P. (2011), "Consumer profiles in reality vs fantasy-based virtual worlds: implications for brand entry", *Journal of Research in Interactive Marketing,* Vol. 5, No. 4, 298-312.

[78] Mehrabian, A. (1976), *Public Spaces and Private Spaces: The Psychology of Work, Play and Living Environments.* Basic Books Inc, New York

[79] Messinger, P.R., Eleni, S., Lyons, K., Bone, M., Niu, R., Smirnov, K. and Perelgut, S. (2008), "Virtual worlds - past, present, and future: New directions in social computing", *Decision Support Systems*, Vol. 47, No. 3, pp. 204-228

[80] Mehrabian, A., and Russell, J.A. (1974), *An Approach to Environmental Psychology,* MA: MIT Press, Cambridge

[81] *Milliman, R.E. (1982), "Using background music to affect the behavior of supermarket shoppers", Journal of Marketing, Vol. 46, pp. 86–91.*

[82] Milliman, R.E. (1986), "The influence of background music on the behavior of restaurant patrons", *Journal of Consumer Research,* Vol. 13, pp. 286-289.

[83] Morin, S., Dube, L. and Chebat, J. (2007), "The role of pleasant music in servicescapes: A test of the dual model of environmental perception", *Journal of Retailing,* Vol. 83, pp. 115-130.

[84] Nath, C.K. (2009), "Behaviour of Customers in Retail Store Environment- An Empirical Study", *Journal of Management,* pp. 63-74.

[85] Newman, A.J. (1997), *Consumption and the inanimate environment: The airport setting* (Doctoral dissertation). Manchester Metropolitan University, Manchester

[86] Newman, A.J. (2007), "Uncovering Dimensionality in the Servicescape: Towards Legibility", *The Services Industries Journal,* Vol. 27, No.1, pp. 15-28.

[87] Nood, D.D. and Attema, J. (2006), "Second Life, the Second Life of Virtual Reality", *The Hague: EPN - Electronic Highway Platform*.

[88] Price-Rankin, K. (2004). *Online Atmospherics: An investigation of feeling and Internet purchase intention.* Unpublished Doctoral Dissertation, The University of Tennessee, Knoxville.

[89] Quartier, K. (2011), *Retail design: lighting as a design tool for the retail environment* (Doctoral dissertation). Retrieved from ProQuest database.

[90] Reddy, N.R.V.R., Reddy, T.N., and Azeem, B.A. (2011), "Role of in-store lighting in store satisfaction", *International Journal of Business and Management Tomorrow,* Vol. 1, No. 3, pp. 1-8.

[91] Reimer, A. and Kuehn, R. (2004), "The impact of servicescape on quality perception", *European Journal of Marketing,* Vol. 39, No. 7/8, pp. 785-808.

[92] Richarme, M. (2007), "*Consumer Decision-Making Models, Strategies, and Theories, Oh My!* (Decision Analyst)", Available from the Data Analyst website: http://www.decisionanalyst.com/Downloads/ConsumerDecisionMaking.pdf (accessed on 27 December, 2014).

[93] Rosenbaum, M.S. (2005), "The symbolic servicescape: your kind is welcomed here", *Journal of Consumer Behaviour,* Vol. 4, No. 4, pp. 257-67.

[94] Rosenbaum, M.S., and Massiah, C. (2011), "An expanded servicescape perspective", *Journal of Service Management,* Vol. 22, No. 4, pp. 471-490.

[95] Russell, J, (1980), "A Circumplex Model of Affect", *Journal of Personality and Social Psychology,* Vol. 39, No. 6, pp. 1161-1178.

[96] Russell, J. and Pratt, G. (1980), "A description of the affective quality attributed to environments*", Journal of personality and social psychology*, Vol. 38, pp. 311-346

[97] Ryu, K., and Jang, S. (2008), "DINESCAPE: A scale for customers' perception of dining environments", *Journal of Foodservice Business Research,* Vol. 11, No. 1, pp. 2-22.

[98] Schiffman, L.G., Kanuk, L.L. and Kumar, S.R. (2010), *Consumer Behaivor* (10th ed.). Dorling Kindersley Pvt. Ltd, UP, India

[99] Schlosser, A.E. (2003), "Experiencing products in a virtual world: The role of goals and imagery in influencing attitudes versus intentions", *Journal of Consumer Research*, Vol. 30, pp. 184-198

[100] Schlosser, A.E. (2006), "Learning Through Virtual Product Experience: The Role of Imagery on True and False Memories", *Journal of Consumer Research,* Vol. 33, pp. 377-383.

[101] Sherman, E., Mathur, A., and Smith, R.B. (1997), "Store environment and consumer purchase behavior: mediating role of consumer emotions", *Psychology & Marketing,* Vol. 14, No. 4, pp. 361-378.

[102] Solomon, M., Bamossy, G., Askegaard, S. and Hogg, M.K. (2006), *Consumer Behaviour: A European Perspective* (3rd ed.). Pearson Education Limited, New Jersey

[103] Spangenberg, E.R., Crowley, A.E. and Henderson, P.W. (1996), "Improving the store environment: Do olfactory cues affect evaluations and behaviors?", *Journal of Marketing,* Vol. 60, pp. 67-80.

[104] Spangenberg, E., Grohmann, B. and Sprott, D. (2005), "It's beginning to smell and (sound) a lot like Christmas: the interactive effects of ambient scent and music in a retail setting", *Journal of Business Research,* Vol. 58, pp. 1583-1589.

[105] Spangenberg, E., Sprott, D., Grohmann, B. and Tracy, D. (2006), "Gender-congruent ambient scent influences on approach and avoidance behaviors in a retail store", *Journal of Business Research,* Vol. 59, pp. 1281-1287.

[106] Spies, K., Hasse, F. and Loesch, K. (1997), "Store atmosphere, mood and purchasing behaviour", *International Journal of Research in Marketing,* Vol. 14, pp. 1-17.

[107] Statista (2015), "Retail sales revenue in the United Kingdom (UK) from 2012 to 2018 (in billion GBP)*",* Available from www.statista.com/statistics/285971/retail-sales-forecast-in-the-united-kingdom-uk-2012-2017/ (accessed December, 2015)

[108] Stokols, D. (1972), On the distinction between density and crowding. *Journal of American Institute of Planners,* Vol. 38, pp. 72-83.

[109] Stokols, D. (1978), "Environmental Psychology", *Annual Review of Psychology*, Vol. 29, pp. 253-295

[110] Summers, T.A. and Hebert, P.R. (2001), "Shedding some light on store atmospherics: Influence of illumination on consumer behaviour", *Journal of Business Research*, Vol. 54, No. 2, pp. 145-150.

[111] Sweeney, J.C. and Wyber, F. (2002), "The role of cognitions and emotions in the music approach avoidance behaviour relationship", *Journal of Service Marketing,* Vol. 16, No.1, pp. 51-69.

[112] Tai, S.H.C. and Fung, A.M.C. (1997), "Application of an environmental psychology model to instore buying behaviour", *The International Review of Retail, Distribution and Consumer Research,* Vol. 7, No. 4, pp. 311-337.

[113] Turley, L.W. and Milliman, R.E. (2000), "Atmospheric effects on shopping behavior: a review of the experimental evidence", *Journal of Business Research*, Vol. 49, No. 2, pp. 193-211.

[114] Varley, R. and Rafiq, M. (2014), *Principles of Retailing* (2nd ed.). Palgrave Macmillan, Basingstoke: England

[115] Vrechopoulos, A.P., O'Keefe, R.M., Doukidis, G.I. and Siomkos, G.J. (2004), "Virtual store layout: an experimental comparison in the context of grocery retail", *Journal of Retailing,* Vol. 80, pp. 13-22.

[116] Vrechopoulos, A., Apostolou, K., and Koutsiouris, V. (2009), "Virtual reality retailing on the web: emerging consumer behavioural patterns", *The International Review of Retail, Distribution and Consumer Research,* Vol. 19, No. 5, pp. 469–482.

[117] Wakefield, L.K. and Blodgett, J.G. (1994), "The importance of Servicescape in leisure Service Settings", *The Journal of Services Marketing,* Vol. 8, No. 3, pp. 66-76.

[118] Wakefield, K.L. and Blodgett, J.G. (1996), "The effect of servicescape on customers' behavioral intentions in leisure service settings", *Journal of Services Marketing*, Vol. 10, No. 6, pp. 45-61

[119] Wang, L.C., Baker, J., Wagner, J.A. and Wakefield, K. (2007), "Can a Retail Web Site be Social?", *Journal of Marketing,* Vol. 71, No. 3, pp. 143 - 157

[120] Wang, Y.J., Minor, M.S. and Wei, J. (2011), "Aesthetics and the online shopping environment: Understanding consumer responses" *Journal of Retailing*, Vol. 87, No. 1, pp. 46-58

[121] Ward, P., Davies, B.J. and Kooijman, D. (2007), "Olfaction and the retail environment: examining the influence of ambient scent", *Service Business,* Vol. 1, pp. 295-316.

[122] Williams, R. and Dargel, M. (2004), "From servicescape to 'cyberscape'" *Marketing Intelligence & Planning*, Vol. 22, No. 3, pp. 310-320.

[123] Wirtz, J., Mattila, A. and Tan, R. (2000), "The Moderating Role of Target-Arousal on the Impact of Affect on Satisfaction-An examination in the Context of Service Experiences", *Journal of Retailing,* Vol. 76, No. 3, pp. 347-365.

[124] Wu, C.S., Cheng, F.F., and Yen, D.C. (2008), "The atmospheric factors of online storefront environment design: An empirical experiment in Taiwan", *Information & Management,* Vol. 45, pp. 493–498.

[125] Wysocki, B. (1979), "Sight, Smell, Sound: They're All Arms in Retailers's Arsenal", *The Wall Street Journal,* Vol. 17, pp. 1-35.

[126] Wyld, D. C. (2010), "A Second Life for organizations?: managing in the new, virtual world", *Management Research Review,* Vol. 33, No. 6, pp. 529 - 562.

[127] Yalch, R. and Spangenberg, E. (1988), *An Environmental Psychological Study of Foreground and Background Music as Retail Atmospheric Factors*. American Marketing Association, Chicago, IL, pp. 106-110.

[128] Yalch, R. and Spangenberg, E. (1990), "Effects of Store Music on Shopping Behavior", *Journal of Consumer Marketing,* Vol. 7, pp. 55-63.

[129] Yalch, R. and Spangenberg, E. (1993), "Using store music for retail zoning", In L. McAlister and M. Rothschild (Eds.), *Advances in consumer research*, (pp. 632-636). Provo, UT: Association for consumer research

[130] Yalch, R. and Spangenberg, E. (2000), "The Effect of Music in a Retail Setting on Real and Perceived Shopping Times", *Journal of Business Research,* Vol. 49, No. 2, pp. 139-147

[131] Young, W., Hwang, K., Mcdonald, S. and Oates, C.J. (2010), "Sustainable consumption: green consumer behaviour when purchasing products", *Sustainable Development*, Vol. 18 No. 1, pp. 20-31

Hidden Markov Models (HMMs) and Security Applications

Rubayyi Alghamdi
Concordia University
Al-Baha University
Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

***Abstract*—The Hidden Markov models (HMMs) are statistical models used in various communities and applications. Such applications include speech recognition, mental task classification, biological analysis, and anomaly detection. In hidden Markov models, there are two states: one is a hidden state and the other is an observation state. The purpose of this survey paper is to further the understanding of hidden Markov models, as well as the solutions to the three central problems: evaluation problem, decoding problem and learning problem. In addition, applying HMMs in real world applications such as security and engineering will improve the classification and accuracy for the whole field.**

Keywords—Markov model; Hidden Markov model; HMM, Markove model; Forward algorithm; Viterbi Algorithm; Baum-Welch algorithm

I. Introduction

The Hidden Markov models (HMMs) were originally introduced in the statistics literature in 1957. They remain one of the most popular models used for evaluating sequential and temporal data due to their efficiency in estimating parameter and doing inferences. Moreover, HMMs are also rich enough to handle real world application. In an engineering field, such as speech processing, source coding and in a security field such as credit card fraud and cloud computing, HMM can be particularly useful [1].

Section two of this survey paper will explain the Markov model and how it relates to the hidden Markov model, but in particular how it generates patterns and hidden patterns. Moreover, the limitations of the Markov process are explained in a weather example. Section three will then address the most important algorithms that HMM uses which are: the Forward Algorithm, Viterbi Algorithm, and Forward-Backward Algorithm. In section four, the paper will show how HMMs apply in real world applications with some solid examples, followed by the last section for conclusion.

II. Background

A. Markov Model

Markov model is mathematical model that make it possible to study complex systems. It is canonical and probabilistic, but specifically for temporal and sequential data. The model are named for Andrey Markov, a Russian mathematician who did some work on stochastic processes in the early nineteen century. The basic idea behind a Markov model is establishing a state of the system and then moving to a new state depending only on the values of the current state not on the previous history of the system. In other words, the future is depending only on the present [2]. The Markov model can be used to model an extraordinarily large number of applications such as weather, economic data, music and more.

B. Markov Model definition

Random valuables $\{X_n\}$ (X1,X2,X3,……,Xn) are considered a Markov chain if their joint distribution respects the following three conditions:

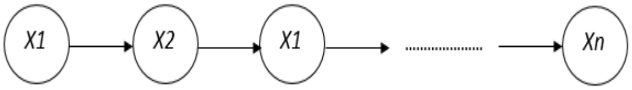


Fig. 1. Chain of observation state

1) The state-space of this process is independent (Discrete space).

2) The time of this process is independent (Discrete time)

3) This process meets the Markov property:

$$P(X_{n+1} = j \mid X_n = i,\ X_{n-1} = i_{n-1},..., X_1 = i_1) \\ = P(X_{n+1} = j \mid X_n = i) \tag{1}$$

$$P(X_1 X_2,..., X_n) = P(X_1)\prod_{n=2}^{N} P(X_n \mid X_{n-1}) \tag{2}$$

Thus, the chain $\{X_n\}$ is a Markov process because the value of the random variable X_{n+1} relies solely on the value X_n and it is not affected by $X_1, X_2,..., X_{n-1}$ variable values. Moreover, the parameter space (time) must be discrete and the state-space shall be discrete (finite) or countable. Furthermore, these processes, in which the right hand side of above the equation is independent of time, thereby leads to the set of state transition probabilities aij :

$$a_{ij} = P(X_{n+1} = j \mid X_n = i),\ 1 \le i,\ j \le N \tag{3}$$

With the state transition matrix coefficients having the properties:

Al- Baha University

$$a_{ij} \geq 0 \forall j, i \quad \text{And} \quad \sum_{j=1}^{N} aji = 1 \forall i$$

Where (j) is the current state and (i) is the previous state. So, all transitions probability aij is positive and each row must sum up to one, since each row represents the probability of jumping from or staying in the same state [2].

C. Weather example:

The figure below shows all possible first order transitions between the states of the weather example.

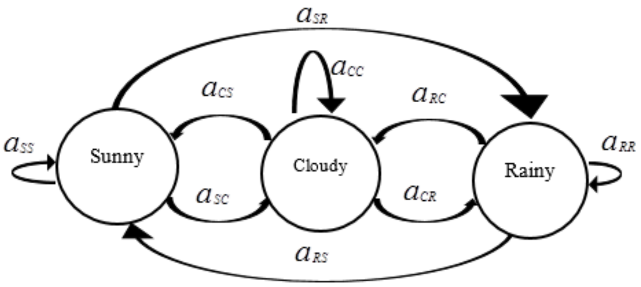


Fig. 2. 1st order transition Diagram

For a first order process with S states, there are S^2 transitions between the states as it is possible for any to follow another or return in the same state. Indeed, the probability of moving from one state to another is aptly called state transition probability. These S^2 probabilities may be collected together into a state transition matrix and do not vary in time, which is an important assumption. The state transitions matrix below shows possible transition probabilities for the weather example:

$$A = \{aij\} \quad i \quad \begin{array}{c|ccc} & \multicolumn{3}{c}{j} \\ & \text{Sunny} & \text{Cloudy} & \text{Rainy} \\ \text{Sunny} & a11 & a12 & a13 \\ \text{Cloudy} & a21 & a22 & a23 \\ \text{Rainy} & a31 & a32 & a33 \end{array}$$

Limitation of a Markov process

In some cases, a Markov process may not be able to describe the events. Consider, in the weather example, that a person who has no ability to see the weather outside somehow acquires a piece of seaweed. This seaweed is affected by the weather, making it perhaps in turn soggy, damp, and dry. In this way, some information leads one to observe the situation of the current state of the weather, forming a link between the seaweed and weather. The pattern should therefore break up into two parts: the observable and the hidden. A more realistic problem presents itself in speech-to-text processing. This application needs words to interact with the systems and these words are effected by some factors such as the vocal chords, size of throat, position of tongue, and several other things. In both examples, the number of hidden states and the number of observed states can be different. This motivation leads to the hidden Markov model [3].

III. Extension to Hidden Markov Model

Hidden Markov Models model time series data. They are used in a huge number of applications such as speech recognition, pattern recognition and data accuracy. The key difference is that a hidden Markov model is a traditional Markov model that assumes the process is modeled with hidden states [4]. Regularly, the state would be visible to the observer making the state transition probabilities the only parameters. However, in a hidden Markov model the observer cannot see the state, but the output is visible depending on the state [5]. Hidden Markov models also have two types of states:

1) The first type of state is (wj). This set of states are the hidden states, and cannot be observed.

2) The second type of state is (vk). This set of states is actually visible, allowing each one to be associated with a state (wj). A hidden Markov model can thereby have a number of hidden or visible states.

3) An initial state π_i

$$P(v_1, v_N, ...w_1, w_n) = P(\pi_1) \prod_{i=2}^{N} P(w_n | w_{n-1}) \prod_{n=1}^{N} P(v_n | w_n) \qquad (4)$$

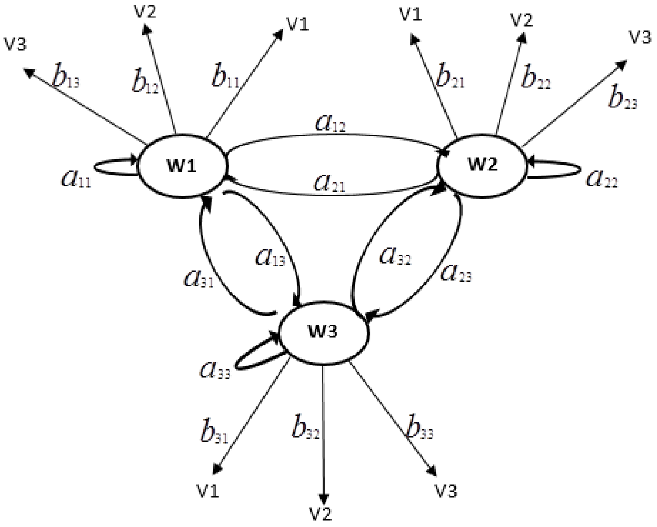


Fig. 3. HMM Transition Diagram

In figure (3), there are three states w={ $w1,w2,w3$} and the transition from state wi at (t-1) to wj at time (t) is given by transition probability matrix aij as in Markov process. The transition then takes the submission of aij because there are transitions between any hidden state wj . In addition, each hidden state can has one or more visible states (v) where bjk gives the probability of the emission matrix. i.e. the process then emits a symbol vk according to the output of the probability of bjk in the current state wj :

$$bjk = P(V_k \mid W_j)$$

The model can be mathematically defined as set: $\lambda = (\pi, w, v, A, B)$. Where λ denotes the HMM, and the vector $\pi = \{\pi_j\}$ is the initial state of probability distribution, A= aij transition probability matrix, and B= bjk emission probability matrix. By taking the submission of bjk because there are always transitions from any hidden state wj to visible states, vk and the submission must be equal to one:

$$\sum_{k=1}^{M} bjk = 1 \forall j$$

Given that, in such a hidden Markov model there are three important issues to be addressed.

A. *The Three Basic Problems for HMMs*

1) *Evaluation problem*

When the hidden Markov module is specified, it will contain: the number of hidden states, the number of visible states, and transition and emission probabilities. In such cases, many numbers of sequences of HMMs ($\lambda_1, \lambda_2, \lambda_3, \cdots \lambda_c$) have these parameters { π, *w*, *v*, aij, bjk }. With observation sequence (v^T) and model (λ), what is the probability that (λ) generated (v^T)? The solution to this valuation problem must choose the model that best matches the observations.

$$P(v^T \mid \lambda)$$

2) *Decoding problem*

Given the observation sequence (v^T) and the model (λ), there must be an optimal corresponding state sequence (w^T). The solution is to find which sequence (w^T) generated sequence (v^T). This problem attempts to uncover the hidden part of the model. In most cases, there is no correct solution, but usually optimal criteria are used to find best possible one. With an HMM model (λ) the first step is trying to find out all possible sequences of the hidden states w^T. The second step is trying to find out for each hidden state w^T the probability that a particular w^T has generated this visible sequence state v^T [6] by:

$$P(v^T \mid \lambda) = \sum_{r=1}^{r\max} p(v^T \mid w_r^T)\, p(w_r^T) \quad (5)$$

Where w_r^T = {*w(1), w(2),…, w(T)*} means all possible sequence of hidden states of length (*T*). The index (*r*) indicates one of the possible sequences of w^T and (*rmax)* is the number of possible sequences that can be generated. With *N* numbers of hidden states, that means $r\max = N^T$ number of possible sequences of hidden states of length (*T*).

To find particular w_r^T by applying the product of the transition probabilities from (*t= 1 to T*) as:

$$p(w_r^T) = \prod_{t=1}^{T} p(w_t \mid w_{t-1}) \quad (6)$$

To compute the visible symbols v^T by taking the product of probability of the emission:

$$p(v^T \mid w_r^T) = \prod_{t=1}^{T} p(v_t \mid w_t) \quad (7)$$

The expression below is to find that given hidden Markov model λ generated the visible symbols v^T :

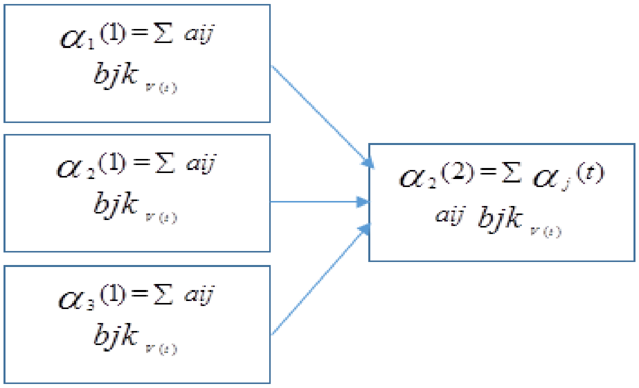

$\alpha_1(1) = \Sigma\ aij$
$bjk_{v(t)}$
$\alpha_2(1) = \Sigma\ aij$
$bjk_{v(t)}$
$\alpha_3(1) = \Sigma\ aij$
$bjk_{v(t)}$
$\alpha_2(2) = \Sigma\ \alpha_j(t)$
$aij\ bjk_{v(t)}$

Fig. 4. trellis Diagram (B)

$$P(v^T \mid \lambda) = \sum_{r=1}^{r\max} \prod_{t=1}^{T} p(v_t \mid w_t) \cdot p(w_t \mid w_{t-1}) \quad (8)$$

This expression's *O* = (N^T *.T*) complexity is substantial. Calculating the probability in this manner is therefore computationally expensive, in particular with large models or long sequences. The coming forward algorithm will solve this problem forward algorithm [6].

3) *Learning problem*

Giving a sequence of hidden sates (w^T) and a sequence of visible states (v^T) infers a number of transition probabilities. In other words, the transition probabilities of (aij) and (bjk) must be estimated, leading to each model working well. This 'learning problem' is one of the most important problems, not only in case of HMM, but for designing any classification [6].

IV. The Algorithms are Used in HMM

A. Forward algorithm (to solve the problem of evaluation)

The evaluation problem explains using recursive algorithms to reduce the complexity of direct expression. Given a sequence of visible symbols v^T, what is the probability that the hidden Markov model will be in a particular state at a particular time w_r^T ? Consider the particular state is $w_{r=2}^{t=2}$ so, what are the possible ways that the process can come in this state? As trellis below:

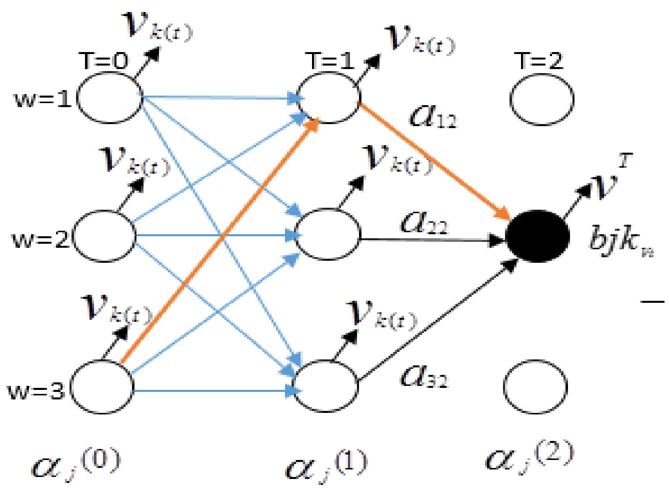


Figure4 trellis Diagram (A)

It is possible to compute $\alpha_j(t)$ By by knowing the HMM's parameters *A*, *B*, π; $\alpha_j(t)$ is also the probability for reaching an intermediate state in the trellis.

To calculate $\alpha_j(t)$ probability for reaching an intermediate state, one must take the sum of all possible paths to that state (every path can be computed as: multiply aij by $bjk_{V(t)}$ and by $\alpha_j(t)$ as shows in figure (4 -B)). This more or less means using $\alpha_j(1)$ to calculate $\alpha_j(2)$, using $\alpha_j(2)$ to calculate $\alpha_j(3)$, and so on until *t* = *T*. The $\alpha_j(t)$. [3] [6].

$$\alpha_j(t) = \begin{cases} 0 & t = 0 \text{ and } j \text{ not initial state} \\ 1 & t = 0 \text{ and } j \text{ is initial state} \\ \left[\sum_i \alpha_j(t-1)\, aij\right] bjk_{V(t)} & \end{cases}$$

To wire formal steps for forward algorithm [6]:

```
Initialize: t =0, $bjk_{V(t)}$, $aij$, $v^T$, $\alpha_j(0)$.
  For   t = t+1
    $\alpha_j(t) = bjk_{V(t)} \cdot \sum_{i=1}^{N} \alpha_i(t-1)aij$
      Until t = T
  Return $P(v^T|\lambda)$ = $\alpha_0(T)$   for final state
  End
```

B. Viterbi Algorithm to solve problem of decoding

This algorithm has ability finds the probable sequence of a hidden state through which the process has made the transition while generating v^T's sequences. Simply, at every step time *T*, the algorithm considers the most probability state among w_r^T. At the end of the process there will be a number of the most probability sequence of hidden states w_r^T that generating v^T's sequences. Figure (4-A) considers the most probability state on step *t=0* is state w_3^0 but only after generating the first symbol of sequences v^T, in step *t=1* is state w_1^1 after generating the first two symbols of sequences v^T. This goes on until having reached the final state in the sequence, the observing state w_2^2. As a result, the path

($\longrightarrow$) is: $w^T = \{w_3^0, w_1^1, w_2^2\}$.

To wire a formal step for the Viterbi algorithm [6]:

```
Initialize:     path= { } t = 0, j = 0;
      For   t = t+1, j = j+1;
         For  j = j+1;
                $\alpha_j(t) = bjk_{V(t)} \cdot \sum_{i=1}^{N} \alpha_i(t-1)aij$ ;
         Until  j =N;
      j' = arg max $\alpha_j(t)$
    Append $w_{j'}$ to path
  Until t= T
 Return Path
End
```

Then, applying bays' rule to classified the sequence of the symbols v^T that has been generated by model λ [2]

$$P(\lambda|v^T) = \frac{p(v^T|\lambda) \cdot p(\lambda)}{p(v^T)} \quad (9)$$

This quantity $P(\lambda|v^T)$ can be used for classification. For example, by computing two models (λ_i) and (λ_j) in the same way, the result is $\lambda_i > \lambda_j$ then obviously the v^T belongs to class λ_i.

$$P(\lambda_i|v^T) \;>\; P(\lambda_j|v^T) \therefore v^T \in \lambda_i$$

C. Forward-Backward or Baum-Welch algorithm (to solve learning problems)

The learning or training process of a hidden Markov model is actually supervised learning. Indeed, the number of sequences of visible symbols and hidden states are already known. Therefore, the goal is to estimate the transition probability matrix aij and bjk with an algorithm similar to forward algorithm, called backward algorithm. The idea of backward algorithm is finding the probability that the process will be in particular state $w_i(t)$ and thereby generate the remaining part of the sequence of visible symbols v^T. The definition of this algorithm is [6]:

$$\beta_i(t) = \begin{cases} 0 \quad w_i(t) \neq w_0 \text{ and } \boldsymbol{t=T} \text{ where } w_0 \text{ is final state} \\ 1 \quad w_i(t) = w_0 \text{ and } \boldsymbol{t=T} \text{ initial state} \\ \sum_j \beta_j(t+1)aij.b_{jkv(t+1)}\; bjk_{v(t)} \end{cases}$$

To wire formal steps for backward algorithm:

```
Initialize: $\beta_i(T)$ , $bjk$ , $aij$ , $v^T$ , , t = T.
 For    t = t-1
  $\beta_i(t) = \sum_j \beta_j(t+1)aij.b_{jkv(t)}$
    Until t = 1
  Return $P(v^T)$ = $\beta_i(0)$ for the known initial state.
End
```

Using $\beta_i(t)$ and $\alpha_j(t)$ to find the probability transition is not correct because the exact value of transition aij and bjk is not known. This means using γ_{ij} to define the probability of transition from state $w_i(t-1)$ to state $w_j(t)$ for a particular sequence v^T by initially using a random value of aij and bjk to estimate $\beta_i(t)$ and $\alpha_j(t)$ as below:

$$\gamma_{ij}(t) = \frac{\alpha_i(t-1)a_{ij}b_{jk}\cdot\beta_j(t)}{p(v^T|\lambda)} \qquad (10)$$

After defining γ_{ij} one must then find the expected number of transitions from $w_i(t-1)$ to $w_j(t)$ at any time in the sequence v^T by:

$$\sum_{t=1}^{T}\gamma_{ij}(t). \qquad (11)$$

After that, finding the total expected number of transitions from state w_i to any state by:

$$\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t). \qquad (12)$$

There are two quantities: the total expected number of transitions from state w_i to state w_j in the sequence v^T and the total number of expected transitions from state w_i to another state. The next step is finding the transition probability aij [6][4]:

$$\hat{aji} = \frac{\sum_{t=1}^{T}\gamma ij(t)}{\sum_{t=1}^{T}\sum_{k}\gamma_{ik}(t)} \qquad (13)$$

One can also find transition probability bjk in the same way by applying:

$$\hat{bji} = \frac{\begin{array}{c}\sum_{t=1}^{T}\sum_{l}\gamma_{il}(t) \\ v(t)=v(k)\end{array}}{\sum_{t=1}^{T}\sum_{l}\gamma_{il}(t)} \qquad (14)$$

V. Real-World Applications

A. Security

1) Credit Card Fraud Detection

Demand for access to online shopping is increasing rapidly with the most popular method of payment being the credit card. Indeed, more than 350 million transactions per year are made in USA including online and regular purchases. As the number of credit card users increases, opportunities for attackers to steal credit card details also increase [8]. There are two types of credit card purchases:

1- Physical credit card: This entails physically using the card in-store to purchase items. To make a fraudulent transaction in such case the attacker need to physically steal the credit card.

2-Virtual credit card: This is use of the credit card without it being physically present; some sensitive information about a credit card is needed here in order for a thief to capitalize on someone else's card (card number, expiration date, secure code) [8].

Virtual credit card fraud can be detected based on the analysis of existing purchase data usually found through studying card behaviors. Card behaviors are stored on the cardholder file. By analyzing purchase patterns, it is possible to figure out the abnormal patterns, then detect the fraud. If human behavior is modeled in a perfect way, any abnormal behavior will be detected because the attacker does not have the same behavior of user.

The hidden Markov model has the power to model much more complicated stochastic processes than Markov model. The key idea of this paper is to build a model with multiple layers of behaviors based on HMM and enumerating methods for normal detection. The bank who issued the credit card has an FDS system detailing when any transaction is made, so that any detection will first send go to FDS for verification purpose. The DFS then receives card details and transaction details such as value of purchase, though the types of items bought are hidden. If the transaction is unknown to the FDS, then it will try to find any abnormal information in the transaction based on the spending profile of the cardholder. This might include shipping address, billing address, etc. If the FDS confirms the transaction as fraud, it raises an alarm, whereby issuing bank will reject the transaction [7] [8]. As the cardholder is using his credit card and making various purchases with different amounts over time, this sequence of types can be used it to create an HMM model. To map the credit card transactions to the HMM model there are a few steps:

1) Define the range of price for example M = {high, medium, low} and this range is considered as K-means in the clustering algorithm.

2) Define the observation symbols Vk, k= {1, 2, 3,….., M} where v is the price of the item. By applying the K-means algorithm on the sequence of observation symbols, the v can be clustered in one of the categories v = {high, medium, low}.

3) The cardholder purchasing is dependent on his need, and every transaction amount is dependent on the corresponding type of purchase, making it so that the transaction amount can be considered as state on the model.

4) Each type of purchase (groceries, electronics items, miscellaneous, etc.) links to a line of business with a corresponding merchant. The information about the merchant's line of business is not known to the bank running the FDS. On the other hand, because this information is required on the stage of registration of the merchant, the merchant is known to the acquiring bank. In addition, some stores have a wide range of varying items. In such cases, the store will be considered miscellaneous; there is no need to determine the actual type of items because any assumption about the information will not be accepted from the bank or from the FDS anyway.

5) The last step is to determine the probability matrix A, B and the initial state – a very important aspect for the Baum-Welch algorithm; that algorithm will then determine these parameters.

1.2. Dynamic Generation of Observation Symbols:

This paper applies the clustering algorithm (K-means) found on each past cardholder transition to find the observation symbols. In fact, the database of the bank, which has several types of information, stores these transitions. However, in this model the amount of the transition is used. In the example mentioned in this paper (table 2) the K-means equals observation symbols (M=3) so k= { c_l , c_m , c_h } [9]. These means are responsible for clustering the amount of the new arriving transaction. FDS generates the observation symbol with:

$$O_x = V_{\arg\min|x-ci|}$$

Where x is the amount the new transaction shows having been spent.

1.3. Spending Profile of Cardholders:

A spending profile for normal behavior of cardholders builds over time. As in the example there are three categories {high, medium, low} and cardholders will assign to corresponding categories based on spending habit. For example corresponding (1) spends high amount when he uses his credit card, then he can be in group {high}. The clustering finishes the profile. Consider that p_i is the percentage of the total number of cardholders' transitions that belong to the mean c_i then the spending profile for cardholder (u) is:

$$SP(u) = \arg\max(p_i)$$

The spending profile (SP) of the cardholder thus determines that the most transactions are in group {high, medium, low}. This paper uses the Baum-Welch algorithm to estimate the HMM parameters (probability matrix A, B and the initial state) for each cardholder. The basic idea is that a cardholder's already existing spending profile can make the initial estimate more accurate. In the three categories {high, medium, low} and base on the cardholder spending profile, the initial state can be chosen. In the phase of training HMM model, there are three steps:

1) Initialization of HMM parameters,

2) Forward procedure

3) Backward procedure.

This stage will eventually create an HMM for each cardholder. After learning HMM, the second step is trying to detect fraudulent transactions. This will be done by taking the symbols from a cardholder's training data and forming an initial sequence of symbols. Let ($O_1, O_2, \ldots, O_R$) be one such sequence up to time *T*, using HMM to compute the probability acceptance of this sequence as follows:

$$\alpha_1 = p(O_1, O_2, O_3, \cdots O_R \mid \lambda)$$

Where α_i the probability and *R* is the length of the sequence. Let O_{R+1} represent a new transaction at a certain time *T+1*. To compute the acceptance probability, fires drop the O_1 from the previous sequence and add O_{R+1} to this sequence. One can then re-enter it to HMM as a new sequence as follows:

$$\alpha_2 = p(O_2, O_3, O_4 \cdots O_{R+1} \mid \lambda)$$

$$\Delta\alpha = \alpha_1 - \alpha_2$$

If $\Delta\alpha > 0$, the HMM has therefore accepted the sequence, indicating a likelihood of fraudulent activity on the credit card. To determine if the new transaction is fraudulent or not, one can simply evaluate if the percentage change in the probability is above a determined threshold.

$$\Delta\alpha \; / \; \alpha_1 \geq threshold$$

Figure	**credit card fraud detection using HMM**	**credit card fraud detection technique**
Figure 5 (a)	-TP is close to credit card the fraud detection technique. - FP is the same for both models.	
Figure 5 (b)	- Two models have accuracy and an average of TP-FP spread. - same exhibit with variation in μ	
Figure 6 (a)	- TP rate is low but increasing - FP rate is low	- TP rate is low - FP rate is obviously high
Figure 6 (b)	- accuracy is 80%	- accuracy is 60%
Figure 7 (a)	- TP rate is low but it is still increasing - FP rate is low	- TP rate is low - FP rate is obviously high
Figure 7 (b)	- accuracy is 80%	- accuracy is 80%
Figure 8 (a)	- TP rate is low - FP rate has changed little, still low	- FP rate is obviously high and higher than TP
Figure 8 (b)	- accuracy is still around 80% - negative value with μ = 2.5	- accuracy is below 40% - negative value with μ = 0.5

For example, if transaction O_{R+1} is fraudulent the bank will decline the transaction and the FDS will reject the symbol. If it is not, the symbol will be added to the normal behavior used to capture any change in spending behavior of a cardholder in a future transaction [8].

1.4. Comparative Performance:

Measuring the performance of HMM models needs a substantial amount of real world data, but because this information is sensitive, banks will never open the information up to researchers. The paper tries to evaluate the model and compares it with a credit card fraud detection technique proposed by Stolfo et al [9]. The methods used in this experiment are:

- Metrics True Positive (TP): to detect and identify fraudulent transactions correctly.

- (FP): means the genuine transactions identified as fraud transactions.

- TP-FP: find the difference between two values to measure the effectiveness of the systems.

- Accuracy: the total number of transactions detected correctly, both fraudulent and genuine.

The experiments are based on a mixed sequence of fraudulent transactions with a sequence of genuine transactions. The cardholder’s profile captures these transactions. Using three profiles { high, medium, low} and the value of the three profiles as P1(55,35,10), P2(70,20,10), P3(95,3,2). The experiment figure results can be found in [9].

As a result, using the HMM for fraudulent activity detection is more accurate than other techniques. Also, the model does not need to have prior knowledge about the fraud transactions to build the learning of the model; this is important because banks consider such information sensitive. In addition, this model can validate the transaction offline so it could not possibly affect credit card transaction processing performance, which needs an online response [9].

2) Predicting Multistage Attacks in Cloud Systems

The Internet provides people with a lot of services to make life easier. One of these services is cloud computing. While this is mostly a helpful and advantageous service, attackers have more opportunity to exploit vulnerabilities with malicious code. Indeed, most security technologies have no ability to raise an early alert about such attacks. This paper is provides a prediction technique based on the hidden Markov model so as to foresee multi-staged cloud attacks. Cloud computing is more difficult in the field of security than in a traditional platform due to shared resources between unknown users and the lack of control of data location. Intrusion Detection Systems (IDS) cannot handle such challenges because they do not incorporate risk assessment and prediction models [7]. This paper therefore implements an Autonomous Cloud Intrusion Detection Framework (ACIDF) so as to focus on such an issue. The purpose of this approach is to evaluate system vulnerability overall, providing control over the system for protection and the ability to respond when threats are predicted or detected.

a) Description of ACIDF:

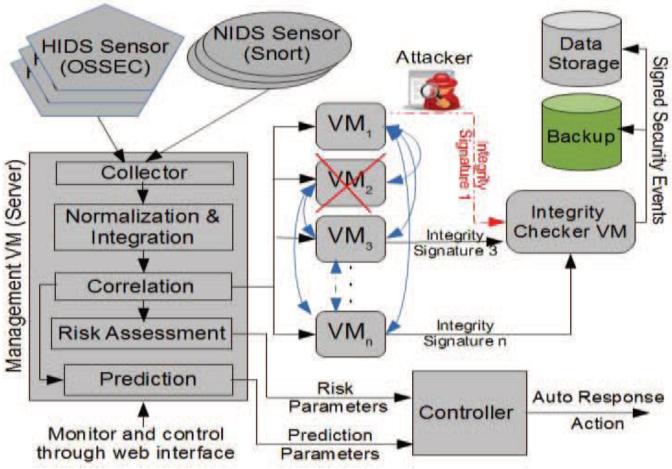


Fig. 5. overview of ACIDF diagram

There are six processes in the model as follows:

1- **Collection process**: this process has three fundamental functions: collecting logs, monitoring network packets, and scanning hosts. The collection process is responsible for collecting all logs and events from sensors such as Host Intrusion Detection Systems (HIDS) and Network Intrusion Detection Systems (NIDS), then redirecting them to the integration process.

2-Integration process: the purpose of this process is to integrate and normalize all collected data from the previous process into the Intrusion Detection Message Exchange Format (IDMEF) protocol to identify their correlation.

3- **Correlation process:** this process creates a correlation between a huge number of normalized events and logs from several sensors and finds the suspected ones by comparing each event with rules of attack. This at the same time reduces false positives.

4- **Risk Assessment process:** the process uses the formula below to evaluate the risk of every group of alerts and then decides if the risk is larger than or equal to one. If it is, an alarm will be fired:

*RISK= (AssetValue * AlertPriority * DetectionReliability)/ NF (1)*

5-Prediction process: works with the correlation process. In case of an attack the prediction component will respond.

6-Auto response process: in this process, the controller uses a fuzzy logic approach to choose the suitable response to protect hosts and the network from attacks.

The outputs of IDS are a huge number of disordered alerts and changes considerably. IDS need a model to deal with this challenge. The best way is using a hidden Markov model. This model works with streaming inputs and can predict future threats in IDS. This paper condones using HMM to trace and evaluate the attacker's actions while proceeding. The hidden state in this model is the sequence of event states that matches the rule of the attacks. However, the observation states are the name of the attack.

b) Explanation of the prediction model

The paper divides the perdition model into 10 elements:

1- **States:** there are four states**:** Hale (*H*) the system is normal with no perceived threat, Investigate (*I*) there is malicious code attempting to compromise the system, Attack (*A*): the malicious code has been executed, and Penetrate (*P*): the malicious code successfully compromised the system.

2- The system is in one of these states. Moreover, like in HMM, states can transition freely between themselves. That will lead to detecting single attacks and predicting multistage attacks.

3- **Observations**: $O = o1, . . . , oK$, represents the alerts which come from sensors. There are four levels for these states depending on the risk of each alert Low, Medium, High, and Very high or (*L, M, H, V*). This section later describes the alert severity function.

4- **State Transition Probability Matrix (*P*):** explains the probability of moving among states. To create states and calculate the transition possibilities there are three steps:

a) Define each attack in a sequence of states, depending on the signature.

b) Determine possible signatures possibly being used by more than one attack and create sequence states, then minimalize the states as much as possible.

c) Calculate the transition probability between states using the Forward-Back algorithm.

5- **Observation Transition Probability Matrix (*Q*):** Calculate the transition probability matrix for observation states.

6- Initial State Distribution Vector (*π)*: for indicating the initial state.

7**- Alert Observation Probability Matrix (*Å*):** if a state has a particular alert, this matrix will detect it. (*Å*) is built based on the training data in the attack dataset.

8- **Assets Cost Matrix (*C*):** using cost vector to find out possibility of consequences for each state in question format.

8- **The Output or emission probability Matrix (*Y*):** showing the output of each sequence of attack states

9**- Alert Severity Function:** This function explains why the severity of each alert is at a particular state. It can be computed using Eq.1, Eq.2 and 3. The result is then mapped to one of the four observation stats (*L, M, H, V*) to clarify the current state of the system.

$$AR^{S} = (AC^{S} * AP * DR^{S}) / NF^{S} \quad (2)$$

$$= (AC^{S} * \mathrm{C}_{\mathrm{Severity}} * \mathrm{N}_{\mathrm{Occurance}} * A_{\mathrm{Frequency}} * DR^{S}) / NF^{S} \quad (3)$$

10**- HMM Prediction Algorithm:** The algorithm is for computing the alert risk and then mapping this risk to an observation states (*L, M, H, V*).

ACIDF predicts against attacks based on the hidden Markov Model by considering signatures of attacks as a sequence of hidden states. Any future attack can be stopped with this model, based on collecting alerts. Furthermore, “the model uses a training algorithm to find the transition, output probabilities, and other prediction parameters.” [10].

B. Engineering

1) Improving time series classification using HMM

This paper proposes a method called a multiple classifier to improve the accuracy of time series by adding an HMM model. This will take into account the temporality of the data, and execute a second stage classification. Moreover, a single classifier system is less powerful than multiple classifier systems. This method assumes that the machine generating the data works under a number of hidden states. That assumption works in many time series datasets, such as the sensors of gas drilling systems.

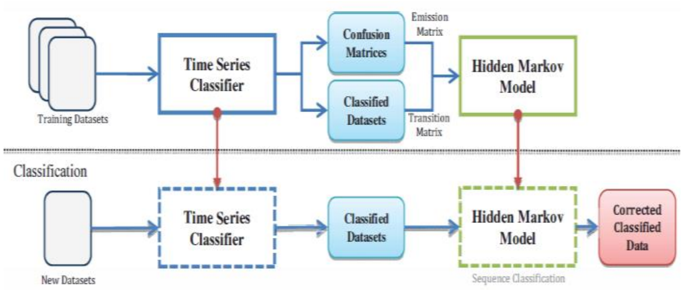


Fig. 6. overview of the model [11]

An overview of the model is in Figure (6). This model has two stages:

1-Training stage: the main purpose of this stage is training the model. The output of the classifier will generate two components: confusion matrices and classified datasets. These components will train the model.

2-Classification stage: this stage is concerned with reclassifying the sequence of classified samples by using the trained model. If there are mistakes on some classified samples, the HMM will correct them.

This model looks at the classification process itself with hidden states and observation states, making HMM suitable for that concept. HMM classifies the data based on varying temporal relations; this is the main difference between HMM and other traditional machine learning classifiers, such as SVM (support vector machines) and NN (Nearest Neighbors algorithm). In order to classify a given sequence of observations states, the most important thing is to find the best sequence of states that generated these observations. This will be done by using the Viterbi algorithm [11]. This paper does some experiments on the model by using real time datasets from different drilling scenarios. The result show that the accuracy of the classification is improved than using single classification.

VI. Conclusion

This survey paper addressed the Hidden Markov model. It explained the traditional Markov model in a simple case and shows that in some states, an observed sequence is probabilistically related to a hidden state. In addition, any real HMM has three central problems: 1) Evaluation problem: given observations sequence and hidden model, what is the probability that observations sequence was generated by that model? The forward algorithm solved this problem. 2) Decoding problem: what sequence of hidden states most probably generated a given sequence of observations? The Viterbi algorithm solved this problem. 3) Learning problem: By giving a sequence of hidden sates and a sequence of visible states, what are the transition probabilities? The forward-backward algorithm solved this problem. Also, this paper shows three real world applications for HMM: 1) Credit Card Fraud Detection; 2) Predicting multistage attacks in cloud systems; and 3) Improving time series classification. The result of these applications has shown that using HMMs provide more advantages such as better accuracy, more reliable, improve the classification, and predicts the future events for the systems.

References

[1] L. R. Rabiner, and B. H. Juang “*An Introduction to Hidden Markov Models*”, IEEE ASSP MAGAZINE, Vol. 3, no.1, 1986, pp.4 – 16.

[2] Fink, Gernot A. “*Foundations of Mathematical Statistics*” Markov Models for Pattern Recognition: From Theory to Applications.2nd ed, London, Springer, 2014, pp.32-49

[3] University of Leeds, “*Hidden Markov Model*” http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.htm. Nov 1st, 2015.

[4] L. R. Rabiner, "*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*" Proceedings of the IEEE, Vol.77, No.2, 1989, pp.257-286.

[5] Alpaydin, E. “*Hidden Markov Models*” Introduction to Machine Learning, 3ed ed. The MIT Press Cambridge, Massachusetts London, England, pp.363 – 385, 2010.

[6] Richard O. Duda, Peter E. Hart and David G. Stork. “*Maximum likelihood and Bayesian estimation*” Pattern Classification 2nd ed., Wiley, New York, 2010, pp 40-55.

[7] Afroza Sultana, Abdelwahab Hamou-Lhadj, and Mario Couture, "*An Improved Hidden Markov Model for Anomaly Detection Using Frequent Common Patterns* ", IEEE ICC Communication and Information Systems Security Symposium, IEEE, Ottawa, ON, 10-15 June, 2012, pp 1113 - 1117.

[8] Ayushi Agrawal, Shiv Kumar, and Amit Kumar Mishra “*Credit Card Fraud Detection: A Case Study*” 2nd International Conference on Computing for Sustainable Global Development, IEEE, New Delhi, 11-13 March, 2015.pp 5 – 7.

[9] Divya.Iyer,Arti Mohanpurkar,Sneha Janardhan,Dhanashree Rathod,and Amruta Sardeshmukh ”*Credit card fraud detection using Hidden Markov Model*” Information and Communication Technologies (WICT), IEEE , Mumbai, India ,11-14 Dec , 2011, pp 1062 – 1066.

[10] Kholidy, Erradi, A, Abdelwahed, S, and Azab, A.”*A Finite State Hidden Markov Model for Predicting Multistage Attacks in Cloud Systems*”, IEEE 12th International Conference on Dependable, Autonomic and Secure Computing, Dalian, 24-27 Aug, 2014, pp 14-19.

[11] Esmael, B. Arnaout, A. Fruhwirth, R.K. Thonhauser, G.” *Improving Time Series Classification U sing Hidden Markov Models”* 12th International Conference on Hybrid Intelligent Systems (HIS), Pune , 4-7 Dec, 2012, pp. 502 – 507.

Toward Secure Web Application Design: Comparative Analysis of Major Languages and Framework Choices

Stephen J. Tipton
College of Arts & Sciences
Regent University
Virginia Beach, Virginia, U.S.A.

Young B. Choi
College of Arts & Sciences
Regent University
Virginia Beach, Virginia, U.S.A.

***Abstract*—We will examine the benefits and drawbacks in the selection of various software development languages and web application frameworks. In particular, we will consider five of the ten threats outlined in the Open Web Application Security Project (OWASP) Top 10 list of the most critical Web application security flaws [12], and examine the role of three popular Web application frameworks (Ruby on Rails (Ruby), Play Framework (Scala), and Zend Framework 2 (PHP)) in addressing a selection of these major threats. In addition, we will compare the strengths and weaknesses of each Web application framework as it pertains to the implementation of strong security measures. Furthermore, for each framework examined, assess how an organization should address these security threats in their software design utilizing their framework of choice. We will suggest the direction in which an organization facing such a decision ought to head; moreover, facilitate such a decision by assessing the benefits and drawbacks of each, based on the findings; and encourage one to decide what works best for the organization's technical direction.**

Keywords—Web; security; framework; application; authentication; ruby; ruby on rails; play framework; Scala; PHP; Zend Framework 2; SQL injection; threats

I. Introduction

In October 2014, Drupal, the popular PHP-based open source content management platform, reported experiencing multiple exploits of vulnerability within its database abstraction API involving carefully crafted requests that resulted in the execution of arbitrary SQL statements [18]. Despite the overarching purpose of the database abstraction API in preventing such exploits, the Drupal Security Team advised site administrators utilizing Drupal 7.x to upgrade to Drupal core 7.32. Administrators who were unable to upgrade were advised to apply a patch to the database.inc file. In a subsequent announcement from the Drupal Security Team, the importance of upgrading to Drupal 7.32 was further outlined and promulgated that simply upgrading would not remove the potential for backdoors in the database, code, or various other locations [5].

A SQL injection attack is such that takes advantage of holes in web services and other web applications by inserting, or "injecting" arbitrary SQL statements "via the input data from the client to the application" [12, 20]. Such vulnerability raises the extreme potential for reading sensitive data from the database; the modification of data via Insert, Update, and/or Delete statements; and the execution of administration operations on the database [20].

WhiteHat Security's 2014 Website Security Statistics Report [25] notes that as a language, "PHP stood out from the pack when looking at SQL Injection, with the languages instances of the vulnerability exhibiting the lowest average number of days at 6.8." Java fell with a much larger gap from PHP at an average of 64.8 days [25]. It is further noted that from the perspective of the Ruby language, statistics were much too minute to include in WhiteHat's report.

While SQL injection, or injection in general, leads the OWASP Top 10 list of web security threats, Web security considerations are not limited to this vulnerability. The scope of this assessment addresses five of the leading threats listed in the Top 10, with SQL injection rounding out this list. Additionally discussed are the threats involving broken authentication and session management; cross-site scripting; insecure direct object references; and security misconfiguration. In particular, a selection of these threats are addressed in relation to three Web application frameworks: Ruby on Rails (Ruby) addressing SQL injection; Play Framework (Scala) addressing Security Misconfiguration; and Zend Framework 2 (PHP) addressing Broken Authentication and Session Management. Addressing these top threats in relation to these three frameworks, and assessing their strengths and weaknesses may facilitate an organization facing the technical decision of choosing an appropriate software stack.

II. Web Security Threat Considerations

The OWASP Top 10 list of Web security threats is rounded out by five of the most critical threats noted within the previous year. Leading the list, as previously cited, are injection attacks (e.g., SQL injection), which were outlined in the case of Drupal's vulnerability in their database abstraction API.

This section considers the leading five threats from the Top 10 list: SQL injection; broken authentication and session management; cross-site scripting; insecure direct object references; and security misconfiguration. Each threat is detailed in its nature, with the primary objective to outline the threats in relation to the scope of this research.

The secure design and implementation of software applications are critically bound to the firm understanding of the threats in which software is designed against. It is imperative that these five threats are considered in detail to provide the understanding necessary for selecting the appropriate software stack to be leveraged in the implementation of the organization's web applications. The following considerations will describe each of the five threats, and the nature imposed upon software applications. The Network Defense Security and Vulnerability Assessment, Volume 5 of the Network Security Administrator Certification [19], echoes this critical aspect due to the increasing importance of Web sites to commercial businesses.

A. SQL Injection

SQL injection attacks exploit vulnerabilities in APIs, as well as other Web applications through the insertion, or injection of arbitrary SQL commands by way of inputting data through the gateway that links the client to the application [12, 20]. Patil and Bamnote [13] cite repercussions from injection attacks including the "unauthorized access to private or confidential information stored ... [via] authentication bypassing, [and] leaking of private information." The Network Defense Security and Vulnerability Assessment [19] parallels this illustration by noting that Web applications are extremely vulnerable due to the ability to receive input data in numerous ways. In general, input data should be analyzed and effectively wrapped by a server-side validation mechanism.

B. Broken Authentication and Session Management

According to the Top 10, authentication and session management are often incorrectly implemented, leaving vulnerable web applications in a broken state in which attackers may potentially compromise user-created passwords, API keys, or session tokens; vulnerabilities left unaccounted for may also "exploit other implementation flaws to assume other users' identities." Web service authentication is not a feature that comes built-in to various Web application frameworks [6]; rather, it is the expectation of developers to implement authentication. Furthermore, this is primarily the case due to the many flavors of adding authentication to HTTP-based web services, including basic authentication, token-based authentication, and session-based authentication.

C. Cross-Site Scripting

Cross-site scripting, or XSS, is the result of "insufficient data validation, sanitization, or escaping" [9] within web applications that present an opportunity for an attacker to execute malicious browser-side code, such as JavaScript. The exploitation of this vulnerability may consummate in the "complete ... compromise of the victim's session," cites Kern. Similar to SQL injection, the Network Defense Security and Vulnerability Assessment [19] asserts that all input data should be thoroughly validated. In XSS vulnerabilities, this threat relates to the browser-side; therefore, XSS can occur when proper validation or escaping on the browser-side is non-existent. According to the Top 10, the malicious execution of scripts can result in hijacked user sessions, defaced web sites, or redirection to phishing sites.

D. Insecure Direct Object References

The Top 10 defines insecure direct object references as "a reference to an internal implementation object, such as a file, directory, or database key" that lacks necessary access controls or other protective measures. For example, web applications are frequently known to use the actual name or key of an object when generating Web pages, without verifying the authorization to access that particular object [21]. The technical impact of such flaws includes the potential for compromising the data associated with the key. To expand upon this example, one may consider a RESTful Web service's URI structure as "intuitive and guessable" [7]. To counteract this, the MVC-pattern featured in many Web frameworks establishes the role of a controller intermediary between the route (the URI structure) and the model layer.

E. Security Misconfiguration

Efficient security requires the existence of secure configuration that is both defined and deployed for the Web application, its framework(s), its server and other related servers (e.g., web, database, etc.), and its platform, according to the Top 10. Furthermore, settings should constantly be maintained. The utilization of what is referred to as "patch management," which is "the administration and supervision of the processes and technology for keeping systems updated with the latest security software defenses," goes hand-in-hand with maintaining good security configurations, and is considered a "basic security must-have" [4]. Configuration defaults are also known to be insecure. For example, the Play framework default configuration includes a generated value for the application's secret key [6]. This is also the case for the Ruby on Rails framework [3]. Furthermore, it is also common to require configuration values to be stored within environment variables, and then referenced in configuration files [6].

III. COMPARATIVE ANALYSIS OF POPULAR WEB APPLIACTION FRAMEWORKS

At some point, an organization will be facing a technical decision involving the selection of a software development stack to accomplish a project that will ultimately enhance or increase business value. The importance of selecting the appropriate tool for the job is drastically increased when weighing the threats outlined in the OWASP Top 10 list. Having previously addressed in detail the five threats that round out the Top 10 list, the next measure to consider is analyzing the comparisons between three web application frameworks across three different software development languages: Ruby on Rails (Ruby); Play Framework (Scala); and Zend Framework 2 (PHP).

Each framework addressed will offer a high-level overview of the framework's features and typical use cases. In a comparative analysis, strengths and weaknesses of each framework will be weighed; the objective is to understand what each framework may or may not offer "out-of-the box," and how each framework will assist developers in designing and implementing secure web services, modular components, or full-blown web applications. From a business angle, such an understanding will facilitate a technical decision.

It ought to be understood, however, that neither of these frameworks are not in itself “more secure than another” [17]; rather, it is the functional features that reside within each framework that assist developers with the tools necessary to secure web applications.

To round out the comparative analysis of these three frameworks, each will include a real-world example within a summarized case study, demonstrating how organizations have utilized that framework of choice to deliver a secure software application. In these short studies, the scope will be limited to a single selection from the five threats that round out the OWASP Top 10 list. It is the objective of this discussion to encourage technical leadership in an organization to make a sound decision when selecting a software development stack.

A. Ruby on Rails (Ruby)

The 10,000-foot level. The overall purpose of a Web application framework is to provide a toolset to developers that facilitate the implementation of Web-based software applications. From a security standpoint, no one framework is going to outweigh another in its own security [17]. The challenge in securing web-based software is raised when developers are faced with implementing secure code. The good news is, Web frameworks provide a set of tools that make this simple for developers to achieve. Ruby on Rails is an example of a Web application framework that achieves this function. In short, the Rails framework "makes it easier to develop, deploy, and maintain web applications" [16].

The leading threat according to the OWASP Top 10 is the exploitation of API vulnerabilities using SQL injection. By virtue of "clever methods," [17] most Rails applications are nearly immune to this threat. However, this is not to assert SQL injection is impossible in Rails applications. If not utilized properly, these "clever methods" will serve no other purpose than to sit unused, leaving a Rails application open to this vulnerability. Ruby on Rails utilizes an Object Relational Mapper (ORM) called Active Record which exposes methods facilitating safe database transactions by properly escaping SQL, which in itself "is immune from SQL injection attacks" [16].

Strengths and weaknesses of the framework. In addressing SQL injection vulnerabilities within Rails applications, it is the responsibility of developers to take advantage of the toolset provided by the Rails framework. As noted previously, Rails exposes "clever methods" that facilitate a near-immunity against SQL injection. While these methods do exist, holes are occasionally uncovered that expose vulnerabilities within the internal method. For example, in January 2013, such a vulnerability was found in dynamic finder methods (e.g., find_by_foo(params[:foo])). The scenario was verified when applications were using the third-party authentication library Authlogic, and the secret session token was known [10].

[16] describe the functionality of Active Record and how it handles the prevention of SQL injection as follows: When multiple parameters are passed into the where method call— a method call that corresponds to the SQL where clause— the first parameter is effectively utilized as a template for generated SQL. Strengthening this feature is the utilization of placeholders, which are replaced with the values from the remaining parts of the array at runtime. Additionally, named placeholders may have their values passed in as a hash of key-value pairs (e.g., {pay_type: pay_type, ...}). Furthermore, these key-value pairs can be passed in as a direct hash reference (e.g. params[:order]) as a single argument to the where method (e.g., Order.where(params[:order])). This latter form is cautioned, however, as it takes in every key-value pair residing within the hash. An even more secure method would essentially white list the key-value pairs that are needed for the Active Record query (e.g., Order.where(name: params[:name], ...)).

Case study: Object Injection and Rails’ Dependency on YAML. William (B.J.) Snow Orvis is a software programmer with Artemis Internet and iSec Partners, and has frequented the Ruby community presenting talks on addressing security issues in Ruby on Rails development. In Orvis' *Secure Development on Rails* presentation [11], he covered an object injection vulnerability (similar to SQL injection) that was discovered by Rails contributor Aaron Patterson [14]. This vulnerability affected all versions of the Rails framework, and entailed "multiple weaknesses in the parameter parsing code ... which allow[ed] attackers to bypass authentication systems, inject arbitrary SQL, inject and execute arbitrary code, or perform a DoS attack on a Rails application." It is noted that the parameter parsing code provides applications the ability to automatically typecast strings to certain data types. The caveat uncovered revealed that certain conversions, in particular the creation of symbols and parsing YAML— a highly utilized dependency in Rails— were supported in the parsing code. "These unsuitable conversions can be used by an attacker to compromise a Rails application," warned Patterson.

The previous scenario outlined by Patterson [14] varied depending on which version of Rails was being used, and whether or not the Web application depended upon support for XML parameters. Mitigating the issue followed a two-fold approach. Primarily, users who did not rely upon XML parameter support were advised to disable XML parsing entirely by deleting Mime::XML from ActionDispatch::ParamsParser::DEFAULT_PARSERS (e.g., ActionDispatch::ParamsParser::DEFAULT_PARSERS.delete(Mime::XML) in Rails 3.x). Alternatively, developers of applications that relied heavily upon XML parsing were advised to disable the YAML and symbol type conversion from the XML parser by deleting Mime::YAML from ActionDispatch::ParamsParser::DEFAULT_PARSERS (e.g., ActionDispatch::ParamsParser::DEFAULT_PARSERS.delete(Mime::YAML) in Rails 3.x). Additionally, this latter approach was further advised to be in parallel with reducing the value of REXML::Document.entity_expansion_limit to limit the risk of entity explosion attacks. Orvis' talk on Secure Development on Rails covered many aspects of Web security, and is recommended as a supplement to this composition.

B. Play Framework (Scala)

The 10,000-foot level. As previously discussed, Ruby on Rails experienced a vulnerability involving parameter parsing, which automatically typecasts strings to certain data types. In Play for Scala, data types are cast statically at compile time, rather than dynamically at runtime. Furthermore, it is this "increased type safety" that garners an immediate benefit throughout the development lifecycle [6]. Play is not constrained to type safety benefits, either. It offers a declarative application URL scheme configuration; it features an HTML5-embraced architecture; it silently reloads on code changes; and more importantly, it is a full-stack framework providing persistence, security, and internationalization [6].

The OWASP Top 10 listed security misconfiguration as the fifth-most critical Web security threat in 2013. Adequate security relies on the definition and deployment of secure configuration for the web application and its numerous components. In addition, the maintenance of these configurations are of equal importance. Cyber Security [4] stresses patch management, along with good security configuration maintenance as a "basic security must-have." Expanding upon this, security misconfiguration is classified by OWASP as easily exploitable. An attacker may access default accounts, unused pages, unpatched flaws, unprotected files and directories, etc. for the primary purpose of obtaining unauthorized access to or knowledge of the system.

Strengths and weaknesses of the framework. Hilton, Bakker, and Canedo [6] confidently assert that simply creating a Play application requires no configuration. This is true as well with Ruby on Rails, which boasts of its convention over configuration. Play initializes a configuration file automatically, with almost all of the parameters being optional. However, with optional parameters, values must sensibly be defaulted. Configuration defaults, in production, are susceptible to insecurity. For example, Play adds a default configuration value for the application's secret key. As expected, these values are able to be overridden, or referenced with environment variables. Moreover, it is required to utilize environment variable references for OS-independent, machine-specific configuration; likewise, it is encouraged to use environment variable references— primarily in production environments— for sensitive configurations, such as database credentials and secret keys.

During development, there is only the need for a single configuration file (e.g., conf/application.conf). However, when deploying to production, different configuration settings will be necessary. Hilton, Bakker, and Canedo [6] note that due to the application being packaged within a JAR file, simply deploying the application, and then manually editing the configuration is inefficient. Consequently, this practice is known to be error-prone and automation-unfriendly. It is highly advised to not make the mistake of sharing identical settings for all environments (e.g., development, test, and production), to shortcut the need for separate configurations. It is likely that at some point, a developer who has shortcut this necessary step could potentially wipe out an entire production asset, such as a database, simply by mistaking which environment was currently being utilized.

It is encouraged to have a "safe" default configuration that is easily overridable by other environments, such as the test environment [6]. Play allows configuration overriding by specifying the override function on a given configuration (e.g., mail.override.address = "info@example.org"). Following any overrides, the developer would then specify the inclusion of a separate configuration file (e.g., development.conf), which would override the default configuration.

Case study: Secure Network Configuration using the Typesafe Reactive Platform and the Play Framework. Auvik Networks "is a hybrid cloud, software-as-a-service (SaaS) application that provides IT professionals with a better way to monitor, configure and automate their network" [24]. The company created a cloud-managed network automation platform to simplify enterprise networking, which has the potential of being highly complex. To deliver this business value, Auvik utilized the Typesafe Reactive Platform to provide a reliable and scalable solution, allowing a continual value add to the business [2].

Utilizing Akka, which facilitates the building of "highly concurrent, distributed, and resilient message-driven applications on the JVM" [1], Auvik leveraged the scalability, clustering, and load balancing to build and deploy their hybrid cloud configuration. Auvik delivered a cloud-based UI that allows a customer to sign up, manage, monitor, and configure their network environment— all via a Web application built on the Play framework. By using Play, and deploying onto the Typesafe Reactive Platform, Auvik was able to take advantage of developer productivity, a modern web application experience, minimal resource consumption, and a high-performing, highly scalable application.

To read more about Auvik Networks use of the Typesafe Reactive Platform and Play framework, the *Auvik Networks simplifies enterprise networking* [2] case study is recommended.

C. Zend Framework 2 (PHP)

The 10,000-foot level. Broken authentication and session management appear in the OWASP Top 10 list second to SQL injection. The vulnerability does not reside within the framework itself; rather, it is in the incorrect implementation that leaves web applications in a vulnerable state potentially allowing attackers to compromise passwords, API keys, or session tokens. Hilton, Bakker, and Canedo [6] echo this fact by disclosing against the misconception that frameworks ship with built-in authentication handling. Because of the many attributes of HTTP-based web services (e.g., basic authentication, token-based authentication, and session-based authentication), the responsibility of handling an authentication mechanism is left to developers; and since developers are the sole proprietors of enabling a secure authentication implementation, it is imperative that entry-points into a web application are efficiently secure.

In some cases, developers are encouraged to utilize open-source libraries to leverage authentication functionality. Rather than reinvent the wheel, frameworks such as Ruby on Rails, in collaboration with the rich Ruby community, foster the utilization of libraries such as Devise or OmniAuth; of course, developers may roll their own authentication implementation as well [15]. The Play framework likewise does not ship with authentication functionality built-in. In fact, rolling one's own authentication implementation in Play is a straightforward process. Hilton, Bakker, and Canedo [6] state that authentication may be performed alongside every HTTP request, prior to an appropriate HTTP response. This allows the existence of a stateless application that requires valid credentials on every HTTP request.

When addressing user-created passwords, the obligation of encryption is introduced. Where libraries such as Devise facilitate Rails developers in integrating a robust, encrypted authentication solution, frameworks such as Zend Framework 2 (ZF2) for PHP ship with encryption components ready to deal with symmetric or asymmetric algorithms; additionally, cryptographic fingerprints [8] further protect authenticated sensitive data. When considering security benefits in ZF2, it is valuable to note that "all the cryptographic and secure coding tools you need to do things right" are readily available out-of-the-box [8].

Strengths and weaknesses of the framework. Karadzhov (2013) outlines the steps and code involved in securing a valuable authentication mechanism in ZF2 applications. One of the many components available to achieve this is Zend\Authentication\Adapter. This component receives user credentials such as username and password; however, it may also be an International Mobile Equipment Identity (IMEI) key unique to mobile devices. If authentication is verified, the identity information is stored to alleviate the need for the user to provide credentials repeatedly. Subsequent requests utilize the stored identity to check accessibility to a given controller and action in the MVC pattern. Coupled with an authentication adapter, a connection the system involved in credential verification is established. For example, when using MD5 for password hashing, an instance of a database adapter such as Zend\Db\Adapter\Adapter would be utilized along with database table information relating to storing the username and password. However, this approach is no longer considered secure [8].

As previously discussed, ZF2 ships with encryption components that ease the challenges of implementing properly secured authentication. As storing passwords hashed with the MD5 algorithm are no longer considered secure, according to Karadzhov [8], ZF2 features the Zend\Crypt\Password component that more efficiently and securely stores passwords. Furthermore, it is advised to use the *Bcrypt* algorithm in replacement of any use of MD5. Enrico Zimuel, creator of Zend\Crypt, states that Bcrypt is considered secure due to the slow computational time of a single hash; therefore, a brute force or dictionary attack would require a much larger amount of time to complete [8]. The Bcrypt algorithm is implemented via the Zend\Crypt\Password\Bcrypt class, in which an instance of this class would create a 60-character hashed string given a plain-text string:

```
use Zend\Crypt\Password\Bcrypt;

$bcrypt = new Bcrypt();

$password = $bcrypt->create('password');

#=>$2a$14$yuD/3v/ldbdOZ0pfljUyJ.a0Q4Ue0UTAoES2BlgK0Op1Z6IF9.aTS
```

Case study: Brute-force Password Cracking. Compounding the threat of compromising passwords is a brute-force method in which bots are used to submit multiple string combinations to authentication forms. While brute-force attacks are more difficult to be successful when employing encryption algorithms such as Bcrypt in ZF2 web applications, it is still a considerable vulnerability to address. Vikram Vaswani, founder of Bombay-based web design company Melonfire, has outlined in a very robust how-to article [23] the mitigation of various security scenarios when developing web applications in the ZF2 architecture. As previously discussed, web applications are vulnerable to attacks including, but not limited to, SQL injection, XSS, CSRF, spam, and brute-force password hacking. Also outlined is the ease in protecting against such vulnerabilities when developing a PHP web application in ZF2. In Vaswani's article, he addresses countermeasures developers can take in mitigating form-based brute-force attacks.

The simplest measure to take to counteract bot interaction via web application forms is to implement a CAPTCHA [23]. ZF2 includes a component that implement this functionality: Zend\Captcha. This component can add FIGlet— ASCII-generated text banners made up of many typefaces— or an image CAPTCHA to the Web form. It also supports the third party web service reCAPTCHA, which integrates remote-generated CAPTCHAs. A caveat to the integration of reCAPTCHA lies in the requirement that the dependency would need to be specified in the Composer configuration. Aside from this, ZF2 essentially ships with many components necessary to secure Web applications.

Vaswani [23] illustrates the setup of a simple contact form, with inputs for name, email address, and CAPTCHA verification. ZF2 provides the Zend\Captcha\Image component, which accepts a number of configuration options (e.g., length of CAPTCHA word, font, directory to store the CAPTCHA, etc.) to generate the CAPTCHA. It is further noted that this component utilizes PHP's GD extension to generate the CAPTCHA image. Once the CAPTCHA is in place, validators are automatically set up and available to the controller and action via the Zend\Captcha component.

To understand more of how ZF2 can assist in securing web applications, Vaswani’s thorough article, *Improve web application security with Zend Framework 2* [23], is recommended.

IV. Conclusion

As outlined in the OWASP Top 10, there is much more to securing Web applications than addressing three of the more common threats in relation to three corresponding web application frameworks. It must be restated as well that no single web application framework is going to be more secure than the other. However, there are features that prove beneficial to developers; while there are features that may not be of much assistance aside from providing necessary tools for developers. It is important to recall that most frameworks do not ship with authentication functionality, or any other fully implemented security threat mitigation. Therefore, the onus is on developers to understand the threats facing web applications. Because these threats are constantly evolving, it is important to remain engaged in current threat assessments in the industry.

We examined the benefits and drawbacks in selecting software stacks comprised of Ruby and the Ruby on Rails framework; Scala and the Play framework; and PHP and Zend Framework 2. It has further considered the leading five threats from the OWASP Top 10, and compared the three frameworks in mitigating a subset of the five threats. In exemplifying such mitigation, we covered three scenarios in which a given framework was utilized in counteracting an exploited vulnerability.

The determination of which software stack works best for a given organization's technical needs must now rely upon the technical focus of the organization. If an organization is seeking to build a robust, scalable, and easily configurable web service, along with a modern user interface, then perhaps the choice for the organization may lead to developing on the JVM using Scala and the Play framework. Companies such as Twitter, LinkedIn, DirecTV, WhitePages, and The Huffington Post have all made this decision to migrate away from their original architectures to the Reactive Platform offered by Typesafe [22].

It may be in the business' interest to quickly deliver a robust application with security-minded authentication functionality, and common threat mitigation approaches— all while not being in possession of a large, knowledgeable team of developers that would be able to roll their own approach. If this is the scenario, perhaps utilizing the Ruby on Rails framework would be the choice, with its rich community of developers and open source libraries that are able to be seamlessly integrated into a complete application.

However, it is noted that one framework is not more secure than the other; likewise, it is noted that most frameworks leave it to developers to implement security measures in Web applications, while being provided the tools necessary for it to be achieved. In retrospect, the single framework considered in this research that demonstrates the most robust set of tools is arguably Zend Framework 2. With components available to achieve more secure encrypted password functionality, ZF2 may be the choice for an organization warranting such a complete toolset.

The decision, however, is up to the organization's technical leadership. It is also highly encouraged to not only understand the threats facing today's Web technologies, but to understand what those threats mean to one's organization. By understanding these threats, and how these threats may affect one's organization, the determination of an appropriate software stack may be decided upon. We only provided a handful of tools; like many Web application frameworks, the responsibility is now up to developers. Likewise, the responsibility is now in the hands of technical leadership.

References

[1] Akka. (2014). Retrieved from http://akka.io/.

[2] Auvik Networks simplifies enterprise networking. (2013). *Typesafe Case Studies & Stories*. Retrieved from http://downloads.typesafe.com/website/casestudies/Auvik-Case-Study.pdf?_ga=1.61301934.324464605.1417574558.

[3] Configuring Rails Applications. (n.d.). *Rails Guides*. Retrieved from http://guides.rubyonrails.org/configuring.html#initializers.

[4] Cyber Security: Doing the Right Things. (2013). Securing our connected world. *TMForum Security and Defense Publication*. Retrieved from http://www.tmforum.org/ResearchPublications/7097/home.html#TRCPublications/Link51039.

[5] Drupal Core - Highly Critical - Public Service announcement - PSA-2014-003. (2014, Oct. 29). *Drupal Security Advisories*. Retrieved from https://www.drupal.org/PSA-2014-003.

[6] Hilton, P., Bakker, E., and Canedo, F. (2014). Play for Scala. Covers Play 2. Shelter Island, NY: Manning Publications Co.

[7] Insecure Direct Object Reference or Forceful Browsing. (2014). *OWASP*. Retrieved from https://www.owasp.org/index.php/Ruby_on_Rails_Cheatsheet#Insecure_Direct_Object_Reference_or_Forceful_Browsing.

[8] Karadzhov, S. (2013). Learn ZF2 Zend Framework 2: Learning by Example. Slavey Karadzhov.

[9] Kern, C. (2014). Securing the Tangled Web: Preventing script injection vulnerabilities through software design. *Communications Of The ACM, 57*(9), 38-47. doi:10.1145/2643134.

[10] Lai, H. (2013). Rails SQL injection vulnerability: hold your horses, here are the facts. *Phusion Corporate Blog*. Retrieved from http://blog.phusion.nl/2013/01/03/rails-sql-injection-vulnerability-hold-your-horses-here-are-the-facts/.

[11] Orvis, W. S. (2013). Secure Development on Rails. *Pivotal Labs*. Retrieved from http://pivotallabs.com/bj-orvis-rails-security/.

[12] OWASP. (2013). OWASP Top 10 - 2013. The Ten Most Critical Web Application Security Risks. *The Open Web Application Security Project*. Retrieved from http://owasptop10.googlecode.com/files/OWASP%20Top%2010%20-%202013.pdf.

[13] Patil, V. S., and Bamnote, Dr. G. R. (2014). An Overview to SQL Injection Attacks and its Countermeasures. *International Journal of Innovative Research & Development, Vol. 3, Issue 1*. Retrieved from http://ojms.cloudapp.net/index.php/ijird/article/view/45590/36927.

[14] Patterson, A. (2013). Multiple vulnerabilities in parameter parsing in Action Pack (CVE-2013-0156). Retrieved from https://groups.google.com/forum/?fromgroups=#!topic/rubyonrails-security/61bkgvnSGTQ.

[15] Rolling Your Own Auth. (n.d.). *Sessions, Cookies, and Authentication. The Odin Project*. Retrieved from http://www.theodinproject.com/ruby-on-rails/sessions-cookies-and-authentication#sts=Rolling Your Own Auth.

[16] Ruby, S., Thomas, D., and Hansson, D. H. (2011). Agile Web Development with Rails. 4th ed. Raleigh, NC; Dallas, TX: The Pragmatic Bookshelf.

[17] Ruby on Rails Security Guide. (n.d.). *Rails Guides*. Retrieved from http://guides.rubyonrails.org/security.html.

[18] SA-CORE-2014-005 - Drupal core - SQL injection. (2014, Oct. 15). *Drupal Security Advisories*. Retrieved from https://www.drupal.org/SA-CORE-2014-005.

[19] Security and Vulnerability Assessment. (2011). *Network Security Administrator Certification, Vol. 5*. EC-Council.

[20] SQL Injection. (2014, Aug. 14). *OWASP*. Retrieved from https://www.owasp.org/index.php/SQL_Injection.

[21] Top 10 2013-A4-Insecure Direct Object References. (2013). *OWASP*. Retrieved from https://www.owasp.org/index.php/Top_10_2013-A4-Insecure_Direct_Object_References.

[22] Typesafe Clients. Retrieved from https://typesafe.com/.

[23] Vaswani, V. (2014). Improve web application security with Zend Framework 2. Retrieved from http://www.ibm.com/developerworks/library/se-zend-security/index.html.

[24] What is Auvik. (2014). Auvik Networks. Retrieved from https://www.auvik.com/about/.

[25] WhiteHat Security. (2014). 2014 Website Security Statistics Report. Retrieved from http://info.whitehatsec.com/rs/whitehatsecurity/images/statsreport2014-20140410.pdf.

Pattern Visualization Through Detection Plane Generation for Macroscopic Imagery

Hanan Hassan Ali Adlan[1,2]
Dept. of Computer Science[1,2],
Faculty of Mathematical Science, U.of K.
Khartoum, Sudan[1]
Faculty of Computer and Information Science, PNU
Riyadh, Saudi Arabia[2]

***Abstract*—Macroscopic images are kind of environments in which complex patterns are present. Satellite images are one of these classes where many patterns are present. This fact reflects the challenges in detecting patterns present in this kind of environments. SPOT1b satellite images provide valuable information. These images are affordable and can be applicable in wide applications. This paper demonstrates an approach to generate detection plane that visualize patterns present in the satellite image. The detection plane uses rough neural network to provide optimal representation in backpropagation architecture. Rough set theory combined with multilayer perceptron constitutes the rough neural network. Reduction in the feature dimensionality via the rough module improves the recognition ability of the neural network. It is found that the rough module provides the neural network with optimal features. The ability of the neural network to efficiently detect and visualize the pattern stems from a developed extraction algorithm. The result of the hybrid architecture provides the plane with the best features that visualize the phenomena under investigation. Together with the novel extraction algorithm, the developed system provides a tool to visualize patterns present in SPOT1b Satellite image.**

***Keywords*—pattern detection; hybrid architecture; backpropagation networks; rough set; image patterns**

I. Introduction

Hybrid architectures are novel techniques to intelligent and powerful systems. The architectures usually present combination of different techniques. Recently rough set theory in combination with neural networks enriches the literature with architectures that enhances the recognition ability of neural networks. Hybrid architectures are formed to overcome limitations of individual system. The power in such systems stem from their capabilities to exhibit multiple information processing [1, 2, 3, 12]

Recent developments in the area of image recognition involve methods for extraction, classification, and selections [4, 13, 14]. An open problem in this area is to find the best features that enable success recognition in classification processes.

The best set that can represent features is generally goal, data, and classification design dependent. Complex application problems such as in remote sensing, medical imaging ,…etc are likely to present large numbers of features so a method for reduction is highly desirable [5,11, 14,15].

This paper demonstrates an approach to automatic detection plane generation for certain patterns present in SPOT1b satellite image. The paper also provides an automatic feature extraction pseudo code that improves the extraction process in such big data environment. A highlight on the architecture used is provided. The approach is theoretically robust, and found efficient to such complicated image.

II. The RNN Architecture

RNN composed of two phases. A feature extraction phase (FE), and a feedforeward neural network phase. The FE accepts an input signal which is two dimensional input image, generates frames for feature computations. The rough module within the FE phase filtered the features. Outputs of the FE are the inputs to the second phase. Neurons in the input layer accept distilled features from the FE phase.

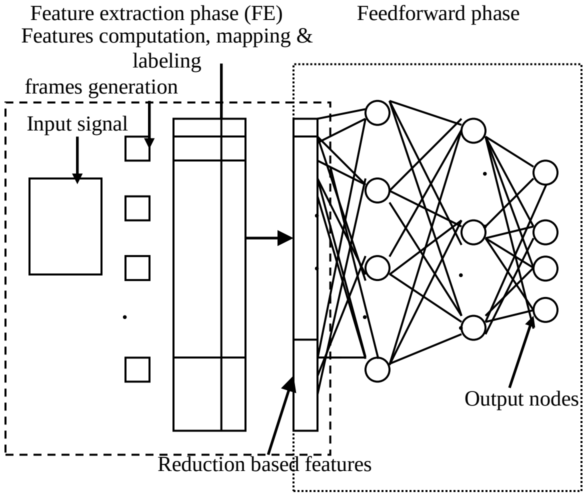


Fig. 1. Rough Neural Network Architecture

Two sets are to be composed from the image, one for training and the other for testing. The network is a backpropagation network. Figure 1 gives configuration of the architecture. [11]

III. Atomatic Feature Extraction

Features extracted from the satellite image assigned labels that distinguish them. For efficient extraction a pseudo code to automatically assign the labels to the extracted features has been developed. The extraction process forms frames from the satellite image, and subsequently extract features from these frames. Association of the frame label is done by a 3x3 filter. Whenever the frame matched a filter kernel, a label is created to identify the frame.

The following pseudo code (pseudo code I) illustrates automatic extraction and labeling

Pseudo code I: Automatic Extraction and Labeling

```
1.  read image file.
2.  generate n kernels.
3.  get next frame.
4.    set i = 1
5.   set sub = ( current frame-kernel(i))
     d = mean(abs(sub))
6 . If d <threshold
7.        set label=i
8.      else
9.              if i<n  , i = i+1; go to step 5
10.             else
11.             set label = n+1, go to step 3
12.     if more frames go to step 3
13.     else end
```

IV. Detection Plane Generation

The detection plane is the plane produced by the network as a result of the present of a pattern in the satellite image. It is considered visualization for the detected phenomena, object,…etc. In this work, we span the image for three categories. The first are water bodies, the second is buildings and networks, and the third is vegetation and swaps. The trained architecture is used in the plane generation. RNN was trained and tested over a portion of 512x512 from an original satellite image (Figure 2). The original image is 2997x4139 scene over the area of Sepang, Kajang, Bangi in Malaysia.

In order to creat the detection plane, a source image is supplied to the RNN. Starting at the top left corner, the trained RNN scanned the image horizontally 3x3 frame at a time. Then it is moved down three rows and moved horizontally again. The process continued until the network traverse the entire image. A 3x3 destination frame is created each time the network moves to a new position. At any position a distance measure is formed from the original frame vector to the network vector, then threshold. The result of the threshold yields the detection plane frame corresponding to the current position. The following pseudo code (Pseudo code II) illustrates the process.

Pseudo code II: Detection plane generation

```
1 .load net
2.  read image
3.      generate detection plane
4.        scan image from upper left
5.        set  f = frame;
6.          fv = extract(f);
7.        red = reduct(fv);
8.  c = sim (net, red);
9.  feature detected?
10.             yes; set corresponding detection
        plane frame = 0
12.             no; set corresponding detection
         plane frame = 1
```

Experiments

The approach described in the previous section illustrated with the SPOT1b satellite image shown in Fig 2. The image is portion of the SPOT1b of the area of Sepang, Bangi, Kajang, in Malayisa. The objective is to recognize three patterns present in the image. The image is 512 × 512 pixels of surface from the original scene of 2997 × 4139 pixels.

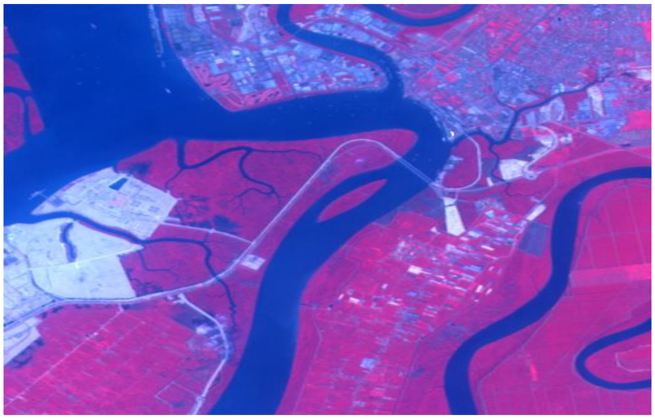


Fig. 2. Portion of spot1b 2997 × 4139 scene over the area of Sepang, Kajang, Bangi (Malaysia). The image is 512 × 512 pixels

The RNN trained and tested using the 512x512 image. Features used in the recognition system isolated from measures of intensities and intensity variations of the pixels, texture features based on spatial gray level dependence matrices, and moment’s invariants. More discussion on this can be found in [10].

Labeling is performed through the mean, standard deviation, and the color. These statistics are computed for each band of the image; in addition to the color, which plays a major factor. Experiments are carried extensively to find the appropriate frame size that best fit in this application. Starting from 16x16, 8x8, 5x5, and ends with 3x3 which is found to be the best frame size. The 5x5 frame produce very low recognition rate, the best of is 35%. The others fail to represent any of the features. This can be attributed to the pixel resolution of the SPOT sensor. 24 feature detectors representation of the satellite image patterns were extracted from the 3×3 frames composed from the original scene.

V. Disscussion

The architecture is examined for the recognition of the three categories water bodies, buildings and networks, and vegetations. 24 feature detectors representation of the satellite image patterns were extracted from 3×3 frames composed from the original scene.

The rough module generated the information system based on the distilled features. The best algorithm suited for this application is found to be the recursive minimal entropy algorithm [9]. 8900 frames were extracted from the image. 5000 of them were used to form the training set. The whole image is used for testing.

Reduction based on Johnson's reducer is given in Table 1. The table displays nine reducts that are considered the filtered features for the recognition process.

TABLE I. JOHNSON REDUCER

	Reduct	Support	Length
1	{ dv, fme, mom 0° , entr 45° , con 0° , con 45° , con 0° , con 90° , cr 135° }	100	9

Different architectures are experimented. Hidden neurons of 200, 150, 100, 50, 10 are examined together with varying learning rates of 0.8, 0.5, 0.1, 0.05, 0.07, 0.01.

The best architecture found composed of 9 input neurons, resulted from the Johnson algorithm. Neurons in the hidden units are 10, and 4 output neurons. The fourth output neuron accommodates the mixed frames, which the system fails to classify as any of the three categories. This can be attributed to the present of more than one pattern or may represent different pattern. The best architecture's parameters are given in Table 2.

TABLE II. NETWORK PARAMETERS FOR THE BEST ARCHITECTURE IN THE ROUGH NEURAL NETWORKS

Network	Parameters
Hybrid rough backpropagation network	Performance function: MSE Goal: 0.01 Learning rate: 0.5 Epochs: 15904 Momentum: 0.95

Table 3 displays the best recognition rates obtained.

TABLE III. ROUGH NEURAL NETWORKS RECOGNITION RATES

Network	Recognition Rates
Training	98.55%
Generalization	96.615%

Previous work [11], experimented a backpropagation network of the same parameters, excluding the rough module resulted in 89.076% recognition rate.

1) Network Scannin

Network scanning mode is the process of generating the detection plane. The scanning mode results in a visualization plane for certain patterns in the image.

The network scanning mode results in Figures 3 to 5. The figures visualize the three patterns under investigation, and found to be present in the satellite image.

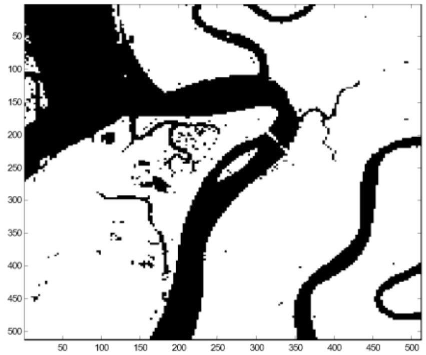


Fig. 3. Detection Plane Representation for the Water bodies

Figure 3 gives visualization of the water bodies. The figure is a result of traversing the image with the neural network in a scan mode. The water bodies appear in black color.

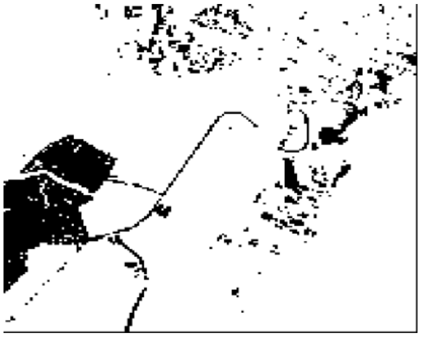


Fig. 4. Detection plane representation for the buildings and networks

The buildings and networks are shown in Figure 4 in black.

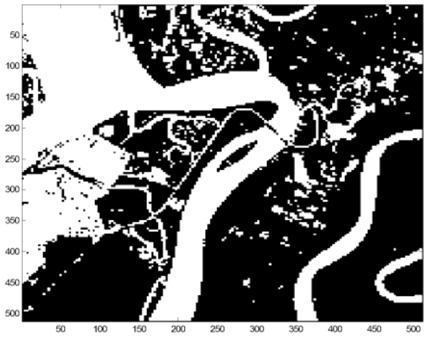


Fig. 5. Detection plane representation for the vegetation and swamps

Figure 5 detect the vegetation and swamps present in the image. The feature appears in black.

VI. CONCLUSIONS

The work in this paper represents visualization of patterns present in SPOT1b satellite image. Each pattern is visualized in a separate plane.RNN which is found to generalize and deal with uncertainty and vagueness present in the SPOT1bsatellite image is developed to produce a system capable of visualizing the patterns present in the satellite image.

REFERENCES

[1] LingasPawan 1998 “Comparison of Neofuzzy and rough neural networks” *Information Sciences* 110, 207-215.

[2] Yasdi R. 1995 “Combining Rough Sets Learning and Neural Learning-method to deal with uncertain and imprecise information”, *Neurocomputing*7 , 61-84.

[3] Ruppert George S. and Mathias Schardt 1997, ‘A Hybrid Classifier for Remote Sensing Applications’. *International Journal of Neural Systems*, Vol. 8, No. 1, 63-68.

[4] Hussain B. and Kabuka M. R. 1994 “A Novel Feature Recognition Neural Network and its Application to Character Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No.1.

[5] Swiniarski Roman W., Hargis Larry, 2001 “Rough sets as a front end of neural-networks texture classifiers”. *Neurocomputing* 36, 85-102.

[6] Khoo Li-Phing, ZhaiLian-Yin, 2001 “A prototype genetic algorithm-enhanced rough set-based rule induction system”. *Computers in industry* 46, 95-106.

[7] Swiniarski Roman, 1992. “Intelligent Decision Support, Hand book of applications and advances of the Rough Sets Theory”. Kluwer, Netherlands.

[8] Gonzalez Rafael C. and Woods Richard E.,1992, “Digital Image Processing”. Addison-Wesley.

[9] Vinterbo S. and Øhrn A. 2000 “Minimal Approximate Hitting Sets and Rule Templates”., *International Journal of Approximate Reasoning*, 25(2), 124- 143. Elsevier.

[10] Hanan H. A. Adlan, AbdRahmanRamli, AdznanJantan, and BachokTaib. 2003, “Backpropagation for Recognition of Invariants and Spatial Detectors of Constituents of Perceptual Patterns”. *Brunei Darussalam Journal of Technology and Commerece*, Vol. 3, No. 1.

[11] Hanan H. A. Adlan, AbdRahmanRamli, Elsadig Ahmed MohdBabiker. 2005, “Improving Generalization in Backpropagation Networks Architectures”. CISAR, International Advanced Technology Congress, Malaysia.

[12] Singh K., Raghuwanshi M. 2009,”Approach to Enhance Performance of Face Recognition Systems Using Rough Sets”, in V. Sn_a_sel (Ed.): Digital Technology Journal 2009, Vol. 2, pp. 23{29, ISSN 1802-5811 (print), ISSN 1802-582X (online).

[13] Gupta S. and Patnaik K. 2008, “Enhancing Performance of Face Recognition System by Using Near Set Approach for Selecting Facial Features” , Journal of Theoretical and Applied Information Technology, © 2005 - 2008 JATIT. All rights reserved.

[14] Meng Yang, Lei Zhang, Simon Chi-Keung Shiu, and David Zhang, 2013. “Robust Kernel Representation with Statistical Local Features for Face recognition”, IEEE Transactions on Neural Networks and Learning Systems, vol. 24, No. 6 June

[15] SamanRazavi and Bryan A. Tolson, 2011. “ A New Formulation for Feedforward Neural Networks”, IEEE Transactions on Neural Networks vol. 22 No. 10, October 2011.

Hierarchical Compressed Sensing for Cluster Based Wireless Sensor Networks

Vishal Krishna Singh
Dept. of Information Technology
Indian Institute of Information Technology
Allahabad, India

Manish Kumar
Dept. of Information Technology
Indian Institute of Information Technology
Allahabad, India

Abstract—**Data transmission consumes significant amount of energy in large scale wireless sensor networks (WSNs). In such an environment, reducing the in-network communication and distributing the load evenly over the network can reduce the overall energy consumption and maximize the network lifetime significantly. In this work, the aforementioned problem of network lifetime and uneven energy consumption in large scale wireless sensor networks is addressed. This work proposes a hierarchical compressed sensing (HCS) scheme to reduce the in-network communication during the data gathering process. Co-related sensor readings are collected via a hierarchical clustering scheme. A compressed sensing (CS) based data processing scheme is devised to transmit the data from the source to the sink. The proposed HCS is able to identify the optimal position for the application of CS to achieve reduced and similar number of transmissions on all the nodes in the network. An activity map is generated to validate the reduced and uniformly distributed communication load of the WSN. Based on the number of transmissions per data gathering round, the bit-hop metric model is used to analyse the overall energy consumption. Simulation results validate the efficiency of the proposed method over the existing CS based approaches.**

Keywords—Compressed sensing; in-network communication; network lifetime; traffic load balancing; wireless sensor network

I. Introduction

Wireless sensor networks (WSNs) have revolutionised today's practice of numerous scientific and engineering endeavours, including ecosystems, environmental sciences, military applications, scientific research etc. WSNs are used for sensing physical variables of interest at unprecedented high spatial densities and long-time durations [1]. Applications like environmental monitoring, scientific research etc., explore the benefits of WSNs. Such applications require transferring a huge amount of sensed data from one point of the network to another. Considering the fact, that the energy consumed in transmission of 1 Kb of data over a distance of 100 meters is equal to the energy consumed in executing 300 million instructions with the rate of computation being 100 million instructions per second on a processor with general configurations [3], [4]. Almost 70% of the total energy is consumed in communication within the network [2]. Hence, the inherent constraints of WSNs such as limited bandwidth and limited battery life makes them prone to failure and compromise the network lifetime. Significant energy conservation in such networks can be achieved by: a) minimizing the cost of interaction between the nodes and b) achieving traffic load balancing during in-network communications [3]. Techniques such as data aggregation have been used to efficiently reduce the communication load of the network, however the issue of asymmetric load distribution in the network remain an important concern till date. Load balancing and optimized energy consumption are thus, much sought after parameters for multi hop data transmission in WSNs.

This work addresses the problem of uneven energy consumption and network lifetime maximization through a novel in-network data processing scheme which incorporates CS in a novel way over a clustered routing structure. The nodes are randomly deployed in a sensing area which is divided into homogenous sub-regions. Such a division is done to model the real world scenario of an area such as a thermal power plant. In such a deployment, the area can be divided into homogeneous regions in such a way that one homogeneous region is different from the other.

For example, the area with the thermal station (one homogeneous region) will exhibit high temperature readings as compared to the residential areas of the plant (another homogeneous region) and are known as a priori. The proposed HCS is divided into two phases, a) Clustering and communication phase b) CS and data processing phase. The proposed scheme incorporates CS in a way to efficiently distribute the communication load evenly over the network. To the best of our knowledge, the advantages of using CS and hybrid CS on tree based routing structure are many but its advantages over a clustered routing structure have not been explored yet. The contributions of this work can be summarised as:

- A CS based data processing scheme with minimum in-network transmissions.
- A scheme for enhancing the network lifetime by balanced network traffic load distribution

The remainder of the paper is organised as follows: in section II, a summary of the related work is discussed. In section III the problem statement for the proposed work is given. The section IV presents the proposed scheme along with the detailed analysis and framework for CS. A detailed analysis of the energy consumption, based on the bit hop metric model, is also discussed. Finally, in section V, a detailed explanation of the simulation results is presented. The section VI summarises the proposed work with concluding comments.

II. Related Work

Data aggregation in WSNs is considered to be the most easily deployable data reduction technique [2]. With varying network topology, such as cluster based [2], tree based [3], chain based [4], various data aggregation schemes have been proposed [4], [5], [6], [7], [8], [9]. Data aggregation mainly exploits the redundancy in the spatially and temporally co-related data sensed by the nodes [10]. Hence, significant energy conservation is achieved by reducing the amount of data being forwarded by any node. However, data aggregation approaches suffer from certain disadvantages. The most important concern with most of the data aggregation approaches, is the loss of information. Data aggregation approaches mainly focus on transferring only a summary of the sensed value to the sink. Hence, a lot of information about the measured value is sacrificed by the aggregation techniques [11]. Another drawback of data aggregation schemes is the asymmetric load distribution within the network. This results in parts of the network having relatively higher activity and thus becoming non-operational because of dead nodes, leaving the sink isolated. For example, in case of an event, WSNs have large amount of data flowing in the network. In such an environment, activity in the deployment area depends upon the occurrence of an event and position of the nodes [12]. The nodes near the sink have high energy consumption as compared to the nodes in other region because of heavy load of data transmission to the sink. Asymmetric distribution of the load results in high activity in parts of the network causing nodes with heavy communication load to die quickly.

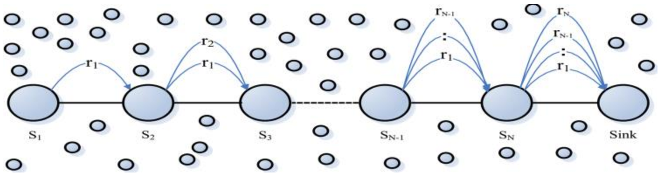


Fig. 1. Traditional data gathering in multi-hop environment

CS has evolved as a promising technique, which can efficiently overcome the drawbacks of the existing aggregation approaches. Data gathering in traditional multi hop environment can be understood from the Fig 1. which shows the highlighted path in a WSN where, ‘N’ sensor nodes form a multi-hop path for data collection. Let the reading generated by the node S1 is r1. Similarly, reading generated by the node S2 is r2 and so on. In a normal data acquisition process, the node S1 sends its reading to node S2. S2 in turn transmits both its data r2 and the data obtained from S1 to S3 .Finally in the end, the last node of the route, sends all the data received from previous nodes along with its own data to the sink. As seen in the Fig.1, the nodes closer to the sink consume more energy as compared to the nodes away from the sink. Due to this, the nodes closer to sink will be drained quickly compromising the lifetime of the network.

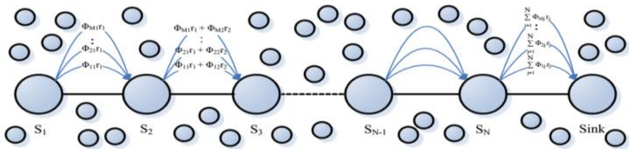

$\Phi_{M1}r_1$
:
$\Phi_{21}r_1$
$\Phi_{11}r_1$
$\Phi_{M1}r_1 + \Phi_{M2}r_2$
:
$\Phi_{21}r_1 + \Phi_{22}r_2$
$\Phi_{11}r_1 + \Phi_{12}r_2$
$\sum_{j=1}^{N}\Phi_{Mj}r_j$
:
$\sum_{j=1}^{N}\Phi_{2j}r_j$
$\sum_{j=1}^{N}\Phi_{1j}r_j$
S1
S2
S3
SN-1
SN
Sink

Fig. 2. Data gathering with Compressed Sensing

The Fig. 2 shows the compressed data gathering method in the highlighted path of a WSN where, ‘N’ sensor nodes form a multi-hop path for data collection. The sink upon receiving all the M samples from ‘N’ nodes, reconstructs the original data. In order to send the ith sample to the sink, S1 generates a random coefficient Φ_{i1} and multiplies it’s reading i.e. r1 with it. The product is then sent to the node S2. Similar multiplication is performed at the node S2 with the random coefficient Φ_{i2} and the sum $\Phi_{i1}r_1 + \Phi_{i2}r_2$ is sent to the node S3. This process is followed by every node in the route to send M samples of their data. Such a transmission results in the sink receiving $\sum_{j=1}^{n} \Phi_{ij} r_j$. The number of samples sent by each node is limited to M. Comparing the Fig.1 and Fig.2, three observations can be made: (a) all the nodes in compressed data gathering, perform the same number of computations (b) the number of transmissions of the first M nodes is more in compressed data gathering scheme as compared to normal transmissions, but the remaining nodes send less messages(c) compressed data gathering scheme distributes the load equally among the nodes of the routing path. Since M is a much smaller number as compared to n, this becomes clearly visible that the number of transmissions in compressed data gathering scheme is far less than normal data gathering [12]. Although, the number of transmissions for collecting M samples from N node is reduced to MN, the literature supports the fact that applying CS naively might not be as beneficial as applying it on a later stage. Considering the drawbacks of applying CS at initial stages of the data gathering process, hybrid approaches were proposed. The hybrid CS scheme proposed in [13] allows the leaf node to sense and send the data without using the CS method, but the nodes which are closer to the sink are the ones which are responsible for the application of CS. The approach in [13] proposes a threshold value K, beyond which CS scheme should be applied but before the threshold is reached, the data collection proceeds in a normal fashion. Owing to the fault tolerance and optimal load balancing properties of clustered routing, the tree based routing scheme becomes the secondary choice of routing strategy in this work. One other advantage of clustered routing is better traffic balancing, which makes the clustering approach preferable choice over the tree based routing scheme. A theoretical analysis presented in [14] proves the efficiency of the compressed data gathering. Aiming the network energy consumption, the authors in [15] proposed a greedy heuristic solution which uses CS with joint routing protocol. In [16], Zigbee protocols are considered in a wireless sensor environment with the assurance of reduced energy consumption. The authors proposed an adaptive scheme which uses CS to reduce the number of transmissions over the network. The scheme proposed in [17], makes use of data aggregation trees formed by subdividing the sensor network into sub networks. The authors in [18], [19] described a three phased CS data gathering scheme. The sensing region is classified into cells with a unique cell head. The cell head collects and forwards the data in compressed fashion along the columns to the Vth row. Finally the data is relayed to the sink along the Vth row cells. In a recent work in [20], the authors proposed a clustering based compressed data gathering scheme. The authors divide the nodes in a clustered fashion to transmit the compressed data through various levels.

Thus, variations of CS have proved to be advantageous in such an environment as WSN, but the existing approaches have certain disadvantages. An important concern is that most of the existing work on CS view compression from the signal processing perspective only [21], [22]. Applications on data compression, from the networking protocol perspective in WSNs are limited [23]. CS, if and when applied naively to a sensor network, imposes extra burden over the network. As shown in Fig. 2, suppose $N-1$ nodes are each sending one sample to the Nth node, the outgoing link of that node will carry 'N' samples if no aggregation is performed; or will carry 1 sample if lossy aggregation is performed. If we apply the CS principle directly, the CS aggregation will force every link to carry 'M' samples, leading to unnecessary higher traffic at the early stage transmissions [13]. To overcome these drawbacks, the idea of hybrid CS was proposed but hybrid CS has its own disadvantages. The selection of non-CS and CS points within the hybrid-CS scheme is critical in getting the benefit of CS [13], [14], [15]. Distributed CS [23] suffers as compared to a mixed protocol in large-scale WSNs, under real technological constrains. Unless the network size and compression are both taken into consideration in network design, distributed CS approaches tend to have average performance in terms of lifetime and energy conservation. Interestingly, existing works on compressed data gathering in WSNs, mainly exploit the tree based routing structure. Because of the drawbacks of the tree based routing structure, such as unstable network topology, most of the existing works face the problem of unreliability and poor quality of service [26].

Thus, the proposed HCS is developed on the principle of hybrid CS over clustered routing structure, with the aim of achieving reliable data transmission with minimum energy consumption (achieved by minimizing and balancing the network traffic evenly over the network).

III. Problem Definition

The main objective in WSNs is to reduce the in-network communication and improve the throughput of the network by increasing the network lifetime. However, network lifetime in large scale wireless sensor networks is also significantly affected by uneven energy expenditure by sensor nodes. Nodes with heavy communication load consume more energy and die quickly causing holes and isolation of some regions of the network. It is therefore desirable to process as much data locally as possible so as to reduce the number of bits transmitted. Techniques such as data aggregation are used to reduce the amount of data being forwarded by the nodes. However, data aggregation schemes have certain drawbacks such as, asymmetric load distribution and information loss. In case of an event, WSNs have large number of message transmissions in the network. In such an environment, activity in the deployment area depends upon the occurrence of an event and position of the nodes. The nodes near the sink have high energy consumption as compared to the nodes in other regions because of heavy load of data transmission to the sink. An asymmetric distribution of the load results in parts of the network having relatively higher activity and thus becoming non-operational because of dead nodes, leaving the sink isolated. Energy conservation may be achieved by transmitting only the summary of the sensed data that may result loss of information. Recent reported work state that, the idea of using CS for data transmission can be advantageous in the above scenario. Naïve application of CS to a sensor network imposes extra burden over the network, hence Hybrid CS would be the most suited solution. However, determining the CS and non-CS points is crucial in such approaches to explore the true potential of the scheme. Hence a CS scheme with the ability of uniform load distribution and efficient data transmission is desired. Reduced in-network communication and optimized energy usage in such a scheme will significantly improve stability period and reduced instable region. Hence, a scheme based on compressed sensing over clustered routing structure is proposed.

IV. Proposed Approach

A. Compressed sensing

The idea of compressive data gathering is relatively new in the field of wireless sensor networks. Some of the basic yet essential properties of the framework asserts that, a relatively small number of samples of a sparse data contains enough information to successfully recover the original data with almost no data loss [27]. Mathematically, If a sparse data 'x' can be denoted as $x=\left\{x_{1,}\ x_2,x_3\ \ldots\ldots\ x_N\right\}^{\mathrm{T}}$ such that $x\in\mathrm{R}^N$ and the orthogonal sparse basis or projection of x is given by $\Psi=\left[\Psi_1,\Psi_2,\Psi_3\ldots\ldots\Psi_N\right]$ where Ψ_i is the i^{th} column of Ψ, then x can be given by the following equation:

$$x=\ \Psi S=\sum_{i=1}^{N}\Psi_i.S_i \tag{1}$$

Such that, S is a vector of the coefficient matrix Ψ and "." represents the inner product. According to the theory of the CS, provided the target data x is K-sparse in the basis Ψ, then under specific conditions, M adaptive measurements of x are sufficient to fully recover the original data such that $M<<\mathrm{N}$. Where each weighted measurement, y, can be written as:

$$y=\Phi x \tag{2}$$

The recovery of the target data by M measurements is dependent on the following condition:

$$M\geq c.\mu^2\left(\Phi\Psi\right).\,K.logN \tag{3}$$

Where Φ is a $M\times N$ sensing matrix, 'c' is a positive constant and $\mu(\Phi\Psi)$ is the coherence between the sensing matrix Φ and the representation basis Ψ. The original data is recovered by solving a convex optimization problem, given by:

$$(\|s\|l1=\sum_i|s_i|)\qquad \underset{s\grave{o}\mathrm{R}^n}{min}\|s\|l1$$

$$\text{provided, } y=\ \Phi\Psi S \tag{4}$$

And considering, $x = \Psi \hat{s}$, with $\hat{s}$ being the optimal solution.

Thus, an important conclusion can be drawn from the assertion is that the basic foundation of CS relies on sensing matrices. For reliable and efficient compression, the data must be sparse in some intitutively known domain and the sensing matrix, Φ, must meet the restricted isometric property (RIP).

An essential property of the sensing matrix, Φ, is the Null space property (NSP) and is denoted by:

$$N(\Phi) = \{z : Az = 0\}. \text{ NSP}$$

To explain, a sparse data x, can be completely recovered by $\Phi \mathrm{x}$, if for every pair of distinct vectors such as $x, x^{'} \in \sum_k$, $\Phi \mathrm{x} = \Phi \mathrm{x}^{'}$. Considering the mentioned condition is not true i.e. if $\Phi \mathrm{x} = \Phi \mathrm{x}^{'}$, results $\Phi\left(x - x^{'}\right) = 0$ such that $x - x^{'} \in \sum_{2k}$. Therefore, it can be said that for 'A' to uniquely represent every $x \in \sum_k$, $N(\Phi)$ should not contain a vector that belongs to $\sum_{2k}$. This feature of the sensing matrix is known as spark of the matrix and is defined as:

Definition 1. The spark of a sensing matrix given by Φ, is the smallest number of linearly dependent columns of Φ.

Theorem 1. If the spark of a sensing matrix Φ, is greater than $2k$ then, for a given vector, y, where $y \in \mathrm{R}^m$, there exists a unique data $x \in \sum_k$, such that $y = \Phi x$.

Proof: To prove the theorem 1 by contradiction, let there be a vector 'y' such that $y \in \mathrm{R}^m$ and there exists a unique data $x \in \sum_k$, such that $y = \Phi x$. It is assumed that, spark $(\Phi) \le 2k$. Or it can be said that there are at most $2k$ linearly independent columns which implies that there is an $h \in N(\Phi)$ such that $h \in \sum_{2k}$. Now, since $h \in \sum_{2k}$, therefore $h = x - x'$, where $x, x' \in \sum_k$. As we know that $h \in N(\Phi)$, therefore $\Phi\left(x - x^{'}\right) = 0$ and $\Phi x = \Phi x^{'}$. But this is a contradiction of the above assumption that there exists a unique data $x \in \sum_k$, such that $y = \Phi x$. Hence, spark $(\Phi) > 2k$. Now, considering spark $(\Phi) > 2k$, and for a given 'y' there exist $x, x' \in \sum_k$, such that: $\Phi = \Phi x = \Phi x^{'}$. This implies that, $\Phi\left(x - x^{'}\right) = 0$, OR $\Phi \mathrm{h} = 0$, replacing $x - x^{'}$ with h.

Since spark $(\Phi) > 2k$, hence at most $2k$ columns of Φ are linearly independent and $\mathrm{h} = 0$. Therefore, the theorem 1 is proved as $x = x^{'}$. An important conclusion from the above theorem is that the number of measurements i.e. m, should follow the following condition:

$$m \ge 2k$$

Definition 2. For a measurement matrix Φ, to satisfy the null space property (NSP) of order k, there must exist a constant $C > 0$ such that,

$$\left\|h_{\wedge}\right\|_2 \le C \frac{\left\|h_{\wedge}\right\| 1}{\sqrt{k}} \qquad (5)$$

is true for all $h \in N(\Phi)$ and for all '$_{\wedge}$' such that $|\wedge| \le k$.

That is, a k-sparse vector in $N(\Phi)$ is $h = 0$, iff, the matrix Φ satisfies the NSP. The literature supports the fact that a NSP of order $2k$ is necessary and sufficient condition for a recovery algorithm (say $l1$ minimization).

The NSP guarantees do not cover for data, which is degraded because of noise. In [27], the authors proposed RIP on the sensing matrix 'Φ' for full recovery of the data even if it is corrupted.

Thus, the framework for data transfer using hierarchical CS in WSN can be summarised with the following advantages:

1) The computation load is shifted from encoder (CH) to the decoder end (sink).

2) Routing for data transmission is independent of the compression.

3) Same number of data packets for every node in the network.

Considering the above advantages, an efficient data transfer scheme for wireless sensor networks is proposed in *C and D.*

B. Sensor Network Model

The graph, $G = \langle \mathrm{V}, E \rangle$, is the sensor network where V consists of all (N) sensor nodes in the network and the sink node given by v_0. A link is assumed to be present between two nodes of V iff, the two nodes are within each other's communication range. A hierarchical clustering scheme, with at most two hop transmission, is used for the cluster formation. The nodes choose a random number between 0 and 1 and compare it to the threshold broadcasted by the sink. The nodes with values higher than the threshold are chosen as the cluster head. A major concern of hybrid CS is the point of application of CS and non CS strategies. In this work, the true potential of hybrid CS is explored by choosing two optimal points for the application of CS. Specifically, CS is applied at the FCHs and at the one hop neighbours of the FCHs, depending upon the transmissions received at both the points. Finally, Huffman encoded data is obtained at the SCH and is transmitted to the sink.

A network model with the following assumptions is considered:

1) A wireless sensor network is randomly deployed and the sensor nodes transmit the data on the occurrence of an event.

2) The network consists of only one sink.

3) The deployed area is randomly divided into sub-regions in such a way that readings from one region are different from the other.

4) Establishing routing information consumes relatively less energy as compared to the data gathering process, hence it is not considered for energy computation.

The two phases of the proposed scheme are discussed in the following sections:

C. Clustering and Communication

Selection of First level cluster head (FCH)

The FCH is chosen using the standard LEACH protocol from the deployed sensor nodes. Once the FCHs have been identified, the cluster formation and communication protocol is established as follows:

1) The FCH sends a join message to all the one hop neighbours.

2) All the one hop neighbours join the FCH if one of the following condition is true:-

✓ It has not received a join message from any other FCH.

✓ It has received a join message from more than one FCH. In this case the node joins the nearest FCH.

3) One hop neighbours, after joining the FCH, follow the two hop communication protocol and broadcast a join message to their one hop neighbours.

4) The two hop neighbours of the FCH follow the same communication protocol as explained in (ii).

5) The data is compressed using CS and is forwarded to respective SCHs.

Selection of Second level cluster head (SCH)

Assuming that the sink is aware of the position of all the nodes including the FCHs, the FCHs closest to the sink are identified as the SCHs. A multicast message from the sink to all the FCHs establishes the communication hierarchy. The communication protocol is established as follows:

6) The SCH sends a join message to all the one hop FCHs.

7) All the one hop FCHs join the SCH if one of the following condition is true:-

✓ It has not received a join message from any other SCH.

✓ It has received a join message from more than one SCH. In this case the node joins the nearest SCH.

8) One hop FCHs, after joining the SCH, follow the two hop communication protocol and broadcast a join message to their one hop neighbours.

9) Two hop neighbours of the SCH follow the same communication protocol as explained in (step 7).

10) The SCH applies Huffman encoding on the received data and forwards it to the sink.

D. Compressed Sensing and Data processing

The compression ratio (υ) is defined as the ratio between the amount of data available for transmission and the the amount of data actually transmitted. In a large scale WSN, the number of nodes in a cluster can be relatively high, leading to large amount of data at the cluster heads (CHs). Applying CS at the CH reduces the number of bits significantly, but the compression ratio might still be very low due to huge amount of data from the member nodes. In order to minimize the data to be transmitted and improve the compression ratio, a threshold (T), such that $\upsilon = T / \hat{M}$, is applied at the one hop neighbours of the FCH.

Sensor nodes in each cluster, on detecting an event, transmit their readings to their respective FCHs through one hop or two hop transmission only. If the incoming traffic, at the one hop neighbours, increases from the predefined threshold (T), the received data is compressed using the CS principle. Application of CS at this point of the network not only minimizes the network traffic but also imposes no extra load on the intermediate nodes.

Otherwise, the received data is transmitted to FCHs without compression. CS at one hop neighbours is applied in the same way as on the FCH's and is explained later in this section. One of the FCHs is designated as the SCH and receives the data forwarded by all the FCHs. It is assumed that each FCH already knows the value of the projection vectors in the measurement matrix Φ for all the nodes that belong to that cluster. In real environment a pseudorandom number generator is used to generate the value of the measurement coefficient Φ using the unique id of every node. Thus, with the node id's known, the measurement matrix is constructed locally at the sink and at the cluster heads. Sub-matrices for respective clusters are formed at each FCH, by decomposing the measurement matrix Φ. Let the sub-matrix for i^{th} cluster is given by Φ^{CHi}, the respective cluster head is given by CH_i, and the cluster's data vector is given by d^{CHi}. The cluster head, CH_i, computes the projection of all the data items within the cluster by multiplying the measurement matrix for the cluster with the data received from all its nodes, that is $\Phi^{CHi} d^{CHi}$.

Finally, M projections of the cluster data is forwarded by the CH_i to the SCH. The number of nodes in the cluster and the sparsity of the data determines the value of M. Each SCH, adds its own data (i.e. data of its own cluster) with the data obtained from all the members of the first level cluster and apply Huffman encoding on the received data.

The encoded data is then forwarded to the sink. The sink is responsible for recovering the original data from the received samples. At the sink, Huffman decoding algorithm is used to recover the compressed data sent to the SCH followed by L1 magic for recovering the original data.

Algorithm 1: Clustering and Communication

Input: *v(i).id (Node id), sink, sparsity (S),* $v(i).\min_d = \infty$

Output: *FCH(First level Cluster Head), SCH(Second level Cluster Head)*

```
Start
For all $v(i)$
check eligibility for cluster head
    if $v(i).E > 0$ and $v(i).G \le 0$
    set: temp _ rand ←rand ; generate random value
      set: Threshold ← $P/(1-P(r\%round(1/P)))$ ; using the probability 'P', set the threshold
        if temp_rand ≤ Threshold
          set v(i).type ← 'FCH'; determining FCH
End
    for all FCH(j) ; Communication setup
        Broadcast join message
      for all v(i)
        If $v(i).rec = 1$
          distance ← compute distance from $FCH(j)$ to one hop neighbors.
          if distance < $v(i).\min_d$
            set $v(i).\min_d$ ← distance
            join $v(i)$ with $\min_d$ for respective FCH ; establish one hop neighbours
      End
    End
    for all 1HN(k)
          send join message
          for all $v(i)$
              if $v(i).rec = 0$
                 distance ← compute distance from 1HN(k) $v(i)$
              $v(i).\min_d$ ← distance
              join $v(i)$ with $\min_d$ for respective 1HN; establish two hop neighbours
          End
    End
```

Algorithm: Compressed Sensing and Data Processing

Input: *Compression Threshold (T), FCH, SCH*

Output *: Huffman Encoded* $\sum_{i=1}^{N} \Phi^{FCHi} d^{FCHi}$

```
  for all 1HN
          receive data from its child node ; receive sensed data
          if $1HN_{trans} > T$
             take $CS_{measurement} \leftarrow d^{OHN_i}$ ; column vector of order $(N \times 1)$
               generate $\Phi^{OHNi}$ ; Order (M×N), using node.id
               set $Z_i \leftarrow \Phi^{OHNi} \times d^{OHNi}$ ; Order of Z is $(M \times 1)$
               save: $Z_i$
             $FCH_i \leftarrow Z_i$ ; forward Z to respective FCH
          else
             $FCH_i \leftarrow d^{OHN_i}$ ; data transfer from one hop neighbor to respective FCH
          End
          for all FCH
             $CS_{measurement} \leftarrow d^{FCH_i}$ ; column vector of order $(N \times 1)$
               generate $\Phi^{FCHi}$ ; Order (M×N), using node.id
               set $Y_i \leftarrow \Phi^{FCHi} \times d^{FCHi}$ ; Order of Y is $(M \times 1)$
               save: $Y_i$
             $SCH_i \leftarrow Y_i$ ; data transfer from FCH to respective SCH
          End
           for all SCH
               encode data using Huffman encoding
               sink ← huffman encoded($Y_i$); data transfer from SCH to sink
          End
```

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Network size	$100m \times 100m$ to $500m \times 500m$
Number of sensor node (n)	100 (minimum) and 900 (maximum)
Sink position	(200 m, 250 m) and (100 m, 100 m)
Initial energy	0.5 j
Transmitter/Receiver electronics(E_{elec})	50 nj/bit
Data aggregation (E_{DA})	5 nj/bit/report
Transmit amplifier ($\grave{0}_{fs}$)	10 pj/bit/m2
Sparse ratio of projection matrix (Φ)	$log_2(N)/N$
Message size (l)	1024 bits

V. RESULTS AND DISCUSSION

To evaluate the performance of the proposed framework, simulations were performed for different scenarios and the outcomes were compared with different existing CS based data gathering schemes [23], [24] and [25].

A. Simulation Setup

Lifetime of the network and the number of transmissions, in the proposed scheme, are tested with varying sink positions and are compared with the results presented in [23], [24] and [25]. Table I describes the parameters used for simulations.

B. Communication and Load Distribution analysis

The proposed HCS aims at optimizing the energy consumption by even distribution of the load over the network and maximizing the lifetime of the network by reducing the number of transmissions. The energy consumption at every node is mainly because of the following two kind of activities:

- **Transmissions received** – A major portion of energy is utilized in receiving the bits from other connected nodes in the network. Hence, the total number of transmissions received at every node is monitored and the activity of every node is mapped for each round of data transmission.

- **Transmissions sent** – The received packets are processed and are forwarded to the next hop. The energy dissipated in processing the data is minimal and hence the data forwarding is the next major energy consuming task.

Interestingly, in-network compression using CS allows minimum energy consumption for compression. The major energy consuming task being the recovery of the compressed data, is done at the sink which has sufficient energy resources. The proposed HCS exploits the advantages of CS in such a way that the load is evenly distributed over the network and the number of transmissions between the sensor nodes is minimum. Node activities are monitored and mapped for the following scenarios:

Sink located outside

The Fig. 3 shows a random deployment of 400 sensor nodes in $200m \times 200m$ area. The sink is located at a corner outside the deployment area. The activity of each node is monitored for each data gathering round and is mapped for the number of transmissions in every round. The activity of each node in one such round, specifically for the deployment shown in Fig. 3, is shown in the activity map 1.

As seen in the Fig. 3, the sink is located at the right corner (200, 250) while the nodes are randomly deployed in an area of $200m \times 200m$. While the FCHs (FCH) are many the SCH is only one for this round and is located at (81.78, 188.90). The number of transmissions i.e. sent and received, for every node is analysed and mapped to activity map shown in activity map 1(a) and activity map 1(b) respectively.

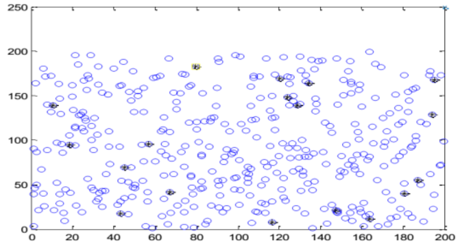
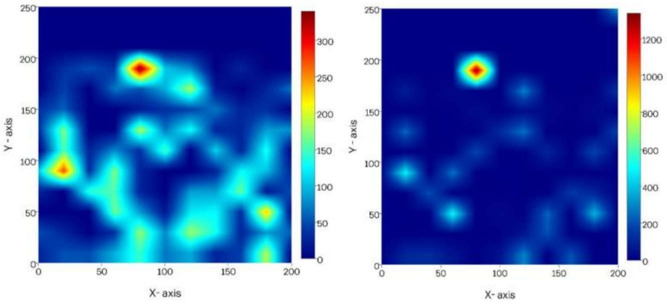


Fig. 3. Random deployment of nodes in $200m \times 200m$ area (sink outside)



(a) (b)

Activity map 1: Activity of nodes (a) sending transmissions (b) receiving transmissions

As seen in the activity map 1(a), data sending activity is marked all around the map leading to the mapping of moderate transmissions (between 80 and 135) by the sensor nodes at different locations in the map. The two dark patches around (81.78, 188.90) and (22.24, 90.01) signify relatively higher activity. The reason for this high activity around (81.78, 188.90) can be understood from the fact that the proposed HCS allows only the SCHs to communicate with the sink. Since, for this deployment, the SCH is located at (81.78, 188.90) hence it is responsible for sending all the information obtained from its member nodes. This distribution of ability to communicate with the sink allows all the other nodes to save a lot of energy which would have been dissipated by them otherwise. Huge activity is seen around (22.24, 90.01) because the FCH located at this position is the only one in the region and is responsible for sending the data from a large number of nodes. Moderate patches over the map show low number of transmissions (between 80 and 135) between the FCHs and respective SCHs. The sink at the corner shows no activity as it only receives data from the SCH. An important conclusion that can be drawn from this map is that the number of transmissions at every node is moderate and varies between 100 and 150. As seen in the activity maps every node shares the load equally and hence the number of sending activity along with the transmission load on every node is distributed evenly throughout the network.

As seen in the activity map 1(b), data receiving activity is marked all around the map leading to the mapping of moderate transmissions (between 400 and 500) by the sensor nodes at different locations in the map. It must be noticed that moderate receiving activity is seen at all the one hop neighbours and FCHs. The proposed HCS ensures that the number of received transmissions at every node is reduced to a low value (around 400 in this case) except at the SCH. The reduced amount of receiving activity at various FCH shows the advantage of applying CS at the one hop neighbours. The SCH, at (81.78, 188.90), shows the highest number of receiving activity for obvious reasons as it receives data from all the FCHs. As seen in the activity map 1(b), the SCH receives about 1200 transmissions but careful observation reveals that the sink at the corner (200, 250) receives about 300 transmissions only. Comparing the activity maps 1(a) and (b), it becomes clear that although the second level cluster receives a large number of transmissions (in this case about 1200) but with the proposed HCS, the outgoing traffic is reduced to nearly 300. With such a distribution of transmissions, the nodes save a lot of energy and hence the network lifetime is enhanced significantly. A significant deduction from this activity map is that the application of CS at one hop neighbours and at FCHs prevents the nodes from transmitting huge amount of data within the network. Thus, the activity maps proves that with reduced number of transmissions the proposed HCS can enhance the network lifetime significantly.

Sink located at the centre

The Fig. 4 shows a random deployment of 400 sensor nodes in $200m \times 200m$ area. The sink is located at the centre of the deployment area. The activity of each node is monitored for each data gathering round and is mapped for the number of transmissions in every round. The activity of each node in one such round, specifically for the deployment shown in Fig. 4, is shown in the activity map 2.

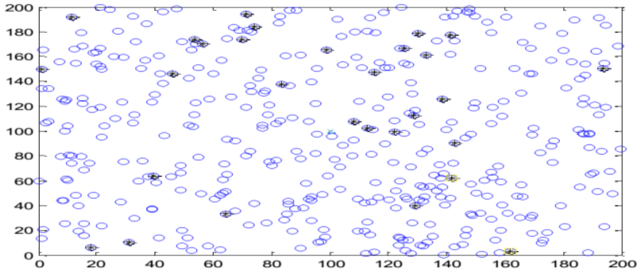
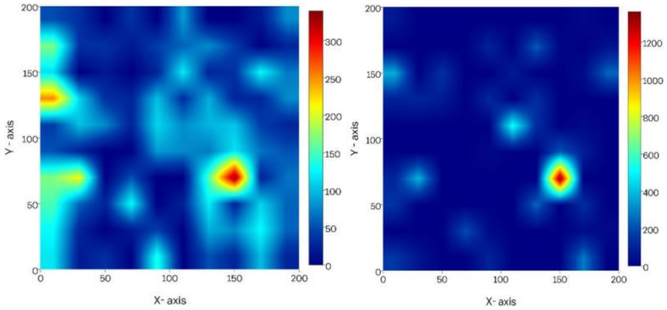


Fig. 4. Random deployment of nodes in $200m \times 200m$ area (sink outside)



(a) (b)

Activity map 2: Activity of nodes (a) sending transmissions (b) receiving transmissions

As seen in the activity map 2(a), data sending activity is marked all around the map leading to the mapping of moderate transmissions (between 55 and 110) by the sensor nodes at different locations in the map. The two dark patches around (144.56, 62.76) and (0, 151.01) signify relatively higher activity. The reason for this high activity around (144.56, 62.76) is the presence of SCH. Since the SCH sends majority of the data to the sink, the number of transmissions being sent by the SCH is relatively high (about 300 in this case). A small patch of relatively lower transmissions is seen around (161.08, 2.23) which is the location of the second SCH. The activity at this location is moderate because as seen in the deployment diagram (figure 6) the majority of the FCHs lie closer to (144.56, 62.76) and hence they send their data to it. Thereby leaving only a few FCH to send their data through the SCH at (161.08, 2.23), thus less data to send to the sink. Huge activity is seen around (0, 151.01) because the FCH located at this position is the only one in the region and is responsible for sending the data from a large number of nodes. A patch with moderate activity is seen around (50, 62.7) as there are relatively less number of FCHs in the region and hence increased two hop transmissions. The sink at the centre shows no activity as it only receives data from the SCHs. As compared to the activity map 1(a) the nodes in the activity map 2(a) have lower transmission values ranging between 55 and 110 and the load is more evenly distributed throughout the area. The advantage of having the sink at the centre of the deployment area is the significant reduction in the transmission distance for the SCH. With almost equal activity at every node the load is evenly distributed throughout the area. With sink at the centre the results tend to improve and the effect of this reduction is seen in the lifetime of the network. Another round of data gathering might result in different number of FCHs and SCHs with different positions.

As seen in the activity map 2(b), data receiving activity is marked all around the map leading to the mapping of moderate transmissions (between 350 and 500) by the sensor nodes at different locations in the map. Important deduction from this activity map is the better distribution of the transmission load as compared to the activity map 1(b) for sink outside the deployment area. The SCH, at (144.56, 62.76), shows the highest number of receiving activity as it receives data from all the majority of the FCHs. As discussed in activity map 2(a) the SCH at (161.08, 2.23) shows relatively low receiving activity as majority FCHs are near to the SCH at (144.56, 62.76). The receiving activity at the sink shows that the number of transmissions received at the sink is almost same as in activity map 1(b) but the difference being the distance between the SCHs and the sink. With sink at the centre, the energy consumption in transmitting the data is relatively low and hence is much better.

C. Transmission Analysis

The sink position remaining the same i.e. at the centre, the total number of packets circulating within the network for the proposed HCS is compared with the mixed algorithm proposed in [23]. The performance of the proposed algorithm is also compared with the other approaches i.e. Distributed compressed sensing (DCS) and Pack and forward (PF) strategies as used in [23]. The Fig. 5 shows the behaviour of the proposed HCS and the three existing approaches as the nodes are increased from 9 to 900.

The figure shows a gradual increase in the number of transmissions as the number of nodes is increased. However, Fig. 5 (a) clearly shows that with the proposed HCS, the number of packets within the network is far less as compared to the mixed algorithm proposed in [23] and DCS and PF schemes used for comparison in [23]. Even for a small network, the number of packets in the proposed HCS is much better than the mixed algorithm. Interestingly, a sudden burst, in the number of sent packets, is seen for a particular network size in the mixed algorithm, DCS and PF scheme. However, the proposed HCS remains unaffected with the network size, proving its efficiency over the existing approaches. As seen in Fig. 5 (b), the average number of transmitted packets per node ranges between 0.1175968 and 0.58376 which is much better than the mixed algorithm [23] where the range of per node packet ranges approximately between 0.333856 to 1.1625. The proposed HCS has better performance than the existing approaches, presenting very less number of sent packets which is always better than the mixed algorithm [23], DCS and PF as used in [23].

D. Energy and Network Lifetime analysis

The lifetime of the network is analyzed for both the discussed scenarios. The stability period, first dead and final dead are considered as the parameters for the energy and lifetime analysis.

Sink located outside

The Fig. 6 shows the lifetime of the network with the proposed HCS, with sink outside the deployment area.

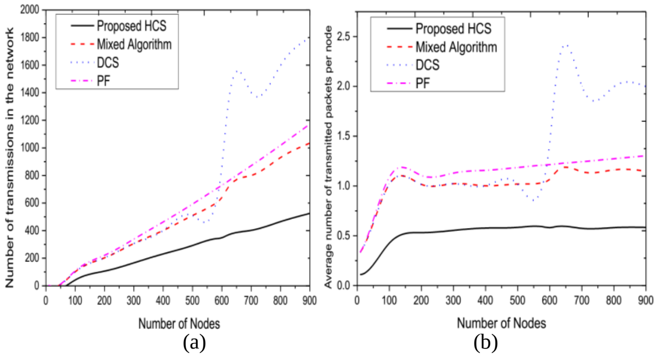


Fig. 5. Comparison among proposed HCS, mixed algorithm, DCS and PF (sink at the centre). (a) Total number of packets circulating within the network. (b) Average number of transmissions per node

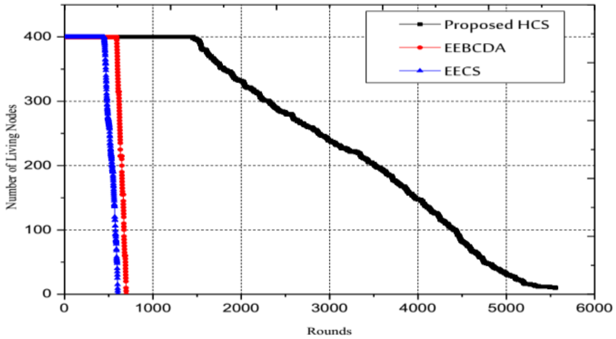


Fig. 6. The number of living nodes over rounds (sink outside)

As evident from the Fig. 6, the proposed HCS is able to improve the lifetime of the network to almost 200 % as compared to EEBCDA as proposed in [24]. The first node with the proposed HCS dies in the 1362nd round whereas with EEBCDA [24] the first node dies in the 591st round. The proposed HCS outperforms the EECS [25] and EEBCDA [24], both in terms of stability period and lifetime of the network. The simulations are run until only 10 nodes are alive and, with the proposed scheme, the 390th node dies in the 5380th round after which the network is considered to be dead. An important aspect of EEBCDA [24] and EECS [25] is the even distribution of the load over the network. The relatively small unstable region i.e. the duration between the first and the last dead, in EEBCDA [24] signifies efficient traffic load balancing within the network. However, in the proposed HCS, with sink at the corner outside the deployment area, the SCHs spend huge amount of energy in transmitting their data to the sink. Over the lifetime of the network, every node becomes a SCH and bears this heavy energy consumption. Hence, although the load is distributed efficiently for almost all the nodes in the network, there might be one or more (depending upon the number of SCHs) nodes dissipating huge amount of energy in data transmission to the sink. The presence of such nodes, in every round of data collection, does not allow the proposed HCS to achieve its true potential in terms of stability period and lifetime. The effect is seen in the duration between the first and the last dead of the proposed HCS. After the first node dies in the 1362nd round, the last considered alive node i.e. 390th node dies in the 5380th. Thus, in the current scenario, though the lifetime of the network is improved greatly but the advantage of distributing the load throughout the network is lost. Changing the position of the sink can not only facilitate the load distribution in the network but can also improve the stability period of the network.

Sink located at the centre

The Fig. 7 shows the lifetime of the proposed HCS with sink at the centre.

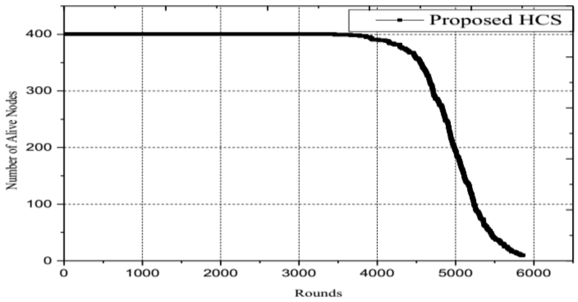


Fig. 7. The number of living nodes over rounds (sink at centre)

As evident from the figure, with sink at the centre, not only a perfect distribution of the load is obtained but the stability period is doubled as well. In the current scenario, the first node dies in the 3873rd round whereas when the sink is placed outside the deployment area, the first node dies in the 1362nd round. The proposed HCS, in the current scenario, is able to improve the stability period of the network to almost 200 % as compared to the scheme proposed in scenario 1. The simulations are run until 10 nodes are alive and the 390th node dies in the 5878th round and after this the network is considered to be dead. Comparing the lifetime of EEBCDA proposed in [24] with the proposed HCS (both scenarios), it becomes clear that the proposed HCS (with sink at the centre) improves the lifetime of the network significantly and is able to achieve about 300 % efficiency over EEBCDA [24] and is even better than EECS [25]. With sink at the centre, the proposed HCS is able to efficiently distribute the load throughout the network and hence better traffic load balancing in the network. The effect of this even distribution is seen in the reduced unstable region. Thus, in the current scenario, not only the advantage of distributing the load throughout the network is achieved but the stability period is also improved significantly. Changing the position of the sink to the centre allows the proposed HCS to achieve better results both in terms of network lifetime and reliability as well.

VI. Conclusion

This paper addressed the problem of network lifetime and uneven energy consumption in large scale WSNs. A Hierarchical Compressed Sensing (HCS) scheme is proposed to achieve efficient traffic load balancing and reduced in-network communication. A major concern for the existing CS based methods is the point of application of CS and this has been successfully resolved in the proposed scheme. HCS is able to reduce and uniformly distribute the number of transmissions in the network. The effect is seen in the lifetime of the network which grows significantly as compared to the existing approaches. Due to efficient load distribution among the nodes of the WSN the availability of all the nodes for data gathering is increased significantly i.e. the stability period is improved. The same has been validated through results. The results prove that the proposed HCS outperforms the existing CS based schemes.

VII. Future Work

We continue to extend our work to analyse the performance of the proposed HCS in a real world deployment. A mathematical model to determine the compression ratio and the communication overhead, remains to be developed in the future.

References

[1] Abbasi AA, Younis M. A survey on clustering algorithms for wireless sensor networks. Computer communications. 2007 Oct 15;30(14):2826-41.

[2] Singh S, Chand S, Kumar B. Energy Efficient Clustering Protocol Using Fuzzy Logic for Heterogeneous WSNs. Wireless Personal Communications. 2016 Jan 1;86(2):451-75.

[3] Xue Y, Cui Y, Nahrstedt K. Maximizing lifetime for data aggregation in wireless sensor networks. Mobile Networks and Applications. 2005 Dec 1;10(6):853-64.

[4] Tan HÖ, Körpeoğlu I. Power efficient data gathering and aggregation in wireless sensor networks. ACM Sigmod Record. 2003 Dec 1;32(4):66-71.

[5] Sinha A, Lobiyal DK. A multi-level strategy for energy efficient data aggregation in wireless sensor networks. Wireless personal communications. 2013 Sep 1;72(2):1513-31.

[6] Banerjee T, Chowdhury KR, Agrawal DP. Using polynomial regression for data representation in wireless sensor networks. International Journal of Communication Systems. 2007 Jul 1;20(7):829-56.

[7] Othman SB, Bahattab AA, Trad A, Youssef H. Confidentiality and Integrity for Data Aggregation in WSN Using Homomorphic Encryption. Wireless Personal Communications. 2015 Jan 1;80(2):867-89.

[8] Liu T, Li Q, Liang P. An energy-balancing clustering approach for gradient-based routing in wireless sensor networks. Computer Communications. 2012 Oct 1;35(17):2150-61.

[9] Sutagundar AV, Manvi SS. Wheel based Event Triggered data aggregation and routing in Wireless Sensor Networks: Agent based approach. Wireless Personal Communications. 2013 Jul 1;71(1):491-517.

[10] Khedo K, Doomun R, Aucharuz S. Reada: Redundancy elimination for accurate data aggregation in wireless sensor networks. Wireless Sensor Network. 2010 Apr 1;2(4):300.

[11] Guo W, Xiong N, Vasilakos AV, Chen G, Cheng H. Multi-source temporal data aggregation in wireless sensor networks. Wireless personal communications. 2011 Feb 1;56(3):359-70.

[12] Enachescu M, Goel A, Govindan R, Motwani R. Scale free aggregation in sensor networks. InAlgorithmic Aspects of Wireless Sensor Networks 2004 Jul 16 (pp. 71-84). Springer Berlin Heidelberg.

[13] Luo J, Xiang L, Rosenberg C. Does compressed sensing improve the throughput of wireless sensor networks?. InCommunications (ICC), 2010 IEEE International Conference on 2010 May 23 (pp. 1-6). IEEE.

[14] Razzaque MA, Dobson S. Energy-efficient sensing in wireless sensor networks using compressed sensing. Sensors. 2014 Feb 12;14(2):2822-59.

[15] Xiang L, Luo J, Vasilakos A. Compressed data aggregation for energy efficient wireless sensor networks. InSensor, mesh and ad hoc communications and networks (SECON), 2011 8th annual IEEE communications society conference on 2011 Jun 27 (pp. 46-54). IEEE.

[16] Caione C, Brunelli D, Benini L. Compressive sensing optimization over ZigBee networks. InSIES 2010 Jul 7 (pp. 36-44).

[17] Mehrjoo S, Shanbehzadeh J, Pedram MM. A novel intelligent energy-efficient delay-aware routing in wsn, based on compressive sensing. InTelecommunications (IST), 2010 5th International Symposium on 2010 Dec 4 (pp. 415-420). IEEE.

[18] Zheng H, Xiao S, Wang X, Tian X. On the capacity and delay of data gathering with compressive sensing in wireless sensor networks. InGlobal Telecommunications Conference (GLOBECOM 2011), 2011 IEEE 2011 Dec 5 (pp. 1-5). IEEE.

[19] Zheng H, Xiao S, Wang X, Tian X, Guizani M. Capacity and delay analysis for data gathering with compressive sensing in wireless sensor networks. Wireless Communications, IEEE Transactions on. 2013 Feb;12(2):917-27.

[20] Xu X, Ansari R, Khokhar A. Power-efficient hierarchical data aggregation using compressive sensing in WSNs. InCommunications (ICC), 2013 IEEE International Conference on 2013 Jun 9 (pp. 1769-1773). IEEE.

[21] Chung WY, Villaverde JF. Implementation of Compressive Sensing Algorithm for Wireless Sensor Network Energy Conservation. InInternational Electronic Conference on Sensors and Applications 2014. Multidisciplinary Digital Publishing Institute.

[22] Qaisar S, Bilal RM, Iqbal W, Naureen M, Lee S. Compressive sensing: From theory to applications, a survey. Communications and Networks, Journal of. 2013 Oct;15(5):443-56.

[23] Caione C, Brunelli D, Benini L. Distributed compressive sampling for lifetime optimization in dense wireless sensor networks. Industrial Informatics, IEEE Transactions on. 2012 Feb;8(1):30-40.

[24] Yuea J, Zhang W, Xiao W, Tang D, Tang J. Energy efficient and balanced cluster-based data aggregation algorithm for wireless sensor networks. Procedia Engineering. 2012 Dec 31;29:2009-15.

[25] Ye M, Li C, Chen G, Wu J. EECS: an energy efficient clustering scheme in wireless sensor networks. InPerformance, Computing, and Communications Conference, 2005. IPCCC 2005. 24th IEEE International 2005 Apr 7 (pp. 535-540). IEEE.

[26] Chand S, Singh S, Kumar B. Heterogeneous HEED protocol for wireless sensor networks. Wireless personal communications. 2014 Aug 1;77(3):2117-39.

[27] Davenport MA, Duarte MF, Eldar YC, Kutyniok G. Introduction to compressed sensing. Preprint. 2011;93(1):2.

[28] Han Z, Li H, Yin W. Compressive sensing for wireless networks. Cambridge University Press; 2013 Jun 6.

[29] Razzaque MA, Bleakley C, Dobson S. Compression in wireless sensor networks: A survey and comparative evaluation. ACM Transactions on Sensor Networks (TOSN). 2013 Nov 1;10(1):5.

[30] Donoho DL. Compressed sensing. Information Theory, IEEE Transactions on. 2006 Apr;52(4):1289-306.

A Model for Classification Secondary School Student Enrollment Approval Based on E-Learning Management System and E-Games

Hany Mohamed El-katary
College of Computing and Information Technology, Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt

Essam M. Ramzy Hamed
College of Computing and Information Technology, Arab Academy for Science, Technology & Maritime Transport, Cairo, Egypt

Safaa Sayed Mahmoud
Ain Shams University
Cairo, Egypt

***Abstract*—Student is the key of the educational process, where students' creativity and interactions are strongly encouraged. There are many tools embedded in Learning Management Systems (LMS) that considered as a goal evaluation of learners. A problem that currently appeared is that assessment process is not always fair or accurate in classifying students according to accumulated knowledge. Therefore, there is a need to apply a new model for better decision making for students' enrollment and assessments. The proposed model may run along with an assessment tool within a LMS. The proposed model performs analysis and obtains knowledge regarding the classification capability of the assessment process. It offers knowledge for course managers regarding the course materials, quizzes, activities and e-games. The proposed model is an accurate assessment tool and thus better classification among learners. The proposed model was developed for learning management systems, which are commonly used in e-learning in Egyptian language schools. The proposed model demonstrated good accuracy compared to real sample data (250 students).**

Keyword—evaluation; learning management system; e-games; classification

I. Introduction

Learning process goes through many generations starting from traditional learning till electronic learning. Traditional learning would be the oldest and the first process of learning then distance learning (DL) and finally electronic learning appears which would be the earliest educational phase process. Traditional learning used for centuries it is based mainly on face to face learning where a lecturer and a group of students meet with each other at certain place.

DL has various ways of descriptions with the more popular formats such as audio, video, broadcast radio and television. DL is characterized by the separation of geographic distance and time difference. E-learning is the use of internet and digital technologies to create experiences that educate the follow of human beings.

Learning Management Systems (LMSs) , which is one of the E-learning tools , provide a wide set of functionalities to support students' learning such as file storage, forums, calendar, news, mail , submission management system, groups surveys , organization, assessments, FAQs (Frequently Asked Questions) or scheduling and educational games. All these types of education have different ways of learning but evaluation process become one of the big challenges for classifying students in fair enrollment approval based on new educational technologies as LMS.

Sotiris, Athanasios and Savvas, mentioned some advantages of LMS in [1] which are:

- If a pupil loses a tutorial because of illness or participation in school activities, he/she has the ability to have access to the presentations, the examples and all the teaching material.
- The pupils have better assimilation of the course concepts in comparison to the ones of previous years, since they can do exercises and tests from their home and evaluate their knowledge.
- The pupils recognize that computers do not exist only for playing games but also as a mean to gain knowledge. Since they are familiar with LMS, they will probably correspond very easily later, in the requirements of their academic studies.

Although there are many successful LMS systems, Sabine ,Kinshuk and Tzu-Chien Liu [2] , concluded in their paper that matching students with learning material and activities Which may fit their preferred ways of learning and study can make learning easier for them. This matching hypothesis is supported by educational theories. The characteristics of each type of student is initiated upon his behavior for example sensing learners like solving problems based on standard procedures.

Classifying student upon his behavior is the main conclusion of this paper.

Yücel urlu, Dai Hasegawa, Hiroshi Sakuta [3] tried to discover a relation between student and LMS by considering the characteristics of the students in order to understand their needs. Student access rates were correlated with their needs, interests, and personal motivations.

Dave Moursund mentioned the importance of educational games for both student and teacher. Using crossword puzzles can help in maintaining and improving vocabulary, spelling skills and knowledge of many miscellaneous tidbits of information [4].

This paper mentioned through results that e-learning and educational games learning was clearly better used in education process rather than traditional learning , the new designed model was more accurate in classifying students based on their interests using WEKA as a data mining tool.

In this paper, related work are explained in section II. In section III, the discussion of the proposed model and the implementation of the proposed model. In section IV results of ten classification algorithms including their performance measures. Comparative analysis, and conclusions are explained in sections V and VI respectively.

II. Related Work

Waraporn Jirapanthong [5] designed a new model which can support students in Thailand to choose their courses based on number of factors as input sector, weight of vector and total number of neuron and applying number of classification algorithm to complete this process.

Qing Yang1, Junli Sun1, Jinqiao Wang1[6] implementing system which created an ontology file for classifying students upon their interest by calculating the similarity of different users by the using of the relationship between the concepts in domain ontology.

Jili Chen, Feng Wang, Kebin Huang, Huixia Wang [7] proposed a new method of classifying student behavior by measuring different student activities by using Fuzzy clustering method to mine E-learning behavior patterns using browsing behavior with Web pages and other learning resources. The learner's behavior can be perceived by clicking on a link, staying at a page.

Andrea, Marco [8] designed a system SOCIALX which is a web application. This system classifies students into 6 classes, first: Involvement which student can be measured by number of contributions that submitted by any student or grades given or acceptability. Second: usefulness students which measures how students contribute others. Third: competence: which measure complains from students and teachers. Fourth: judgment which can measure the ability of student judgment others. Fifth: self-judgment: which measure how student be fair with himself related to teacher evaluation. Sixth: active critical system: which measure the creativity of that student.

Yücel urlu, Dai Hasegawa, Hiroshi Sakuta [9] divided students’ topics and count number of accessing for each topic for each student to measure the highest for each student. Also they concluded the highest accessed material which student access it, they concluded that e-learning systems can be used to improve student-learning patterns and help us in improving traditional courses as well as e-Learning systems.

III. The Proposed Model

The proposed approach, with its main features is essentially based on two components: the student phase and the evaluation phase. Figure1 shows a block diagram for the components of the proposed model used.

A. Model Architecture

This diagram consists of two basic levels or phases, the first one is student phase and the second one is the evaluation phase. In student phase gathering information is done rather automatically based on the online behavior and activities of students as registration ,notification ,course documents , interface tutorial , announcements , useful links , student papers , exercises , quizzes and semantic search which the model can extract information from all these activities.

Huge amount of data are collected continuously from the student interactions with materials, exams and educational games as illustrated there are different resources embedded in the LMS systems which in this model can extract information from students for the next evaluation phase. In evaluation phase student interests gathered and evaluated by comparing each class of study and take the decision for the dedicated student class of interest.

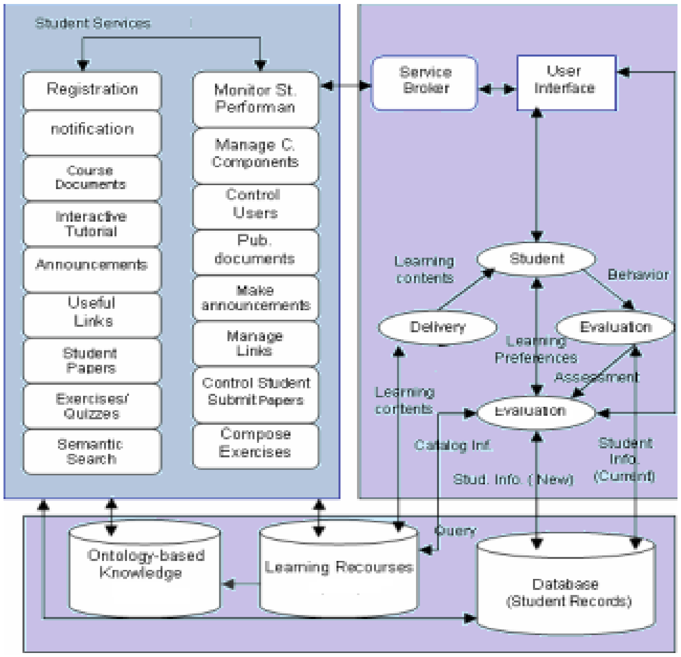


Fig. 1. The proposed model Architecture

B. Methodology of the proposed model and implementation

To implement and evaluate the proposed approach, a conceived system composed by a set of components, where each component is performing a number of student activities. The main features of the proposed recommender system are shown in the next paragraph.

This system provides an analysis of the attributes. Which can trace the distribution of students within each section according to their courses, their activities and educational games desire when achieving school section criteria, which include Grades Qualified Materials.

At this study divided student evaluation into two stages first stage for classifying students, either his interest is scientific or literary, second stage to classify scientific students either, science scientific interest students or mathematical scientific interest students.

As Sebastian Arnold, Jun Fujima, Andreas Karsten and Harald Simeit [10] designed a new model based on game theory which can be adapted for each learner upon his own preferences which divided into four classes each class can meet with every learner behavior.

Evaluation process starts with filling basic information for each student as his name and sex and birthdate and academic School year as illustrate in fig 2 .then second screen appears for gathering courses, activities and e-games grades. Each material classified previously as scientific material or literary material. Gathering material starting from KG1 till

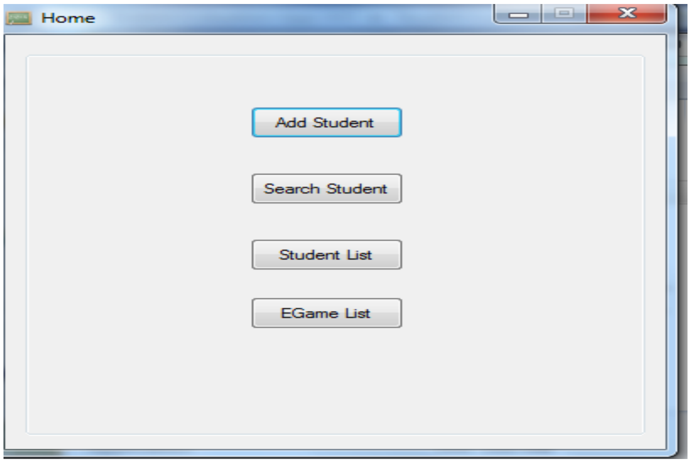


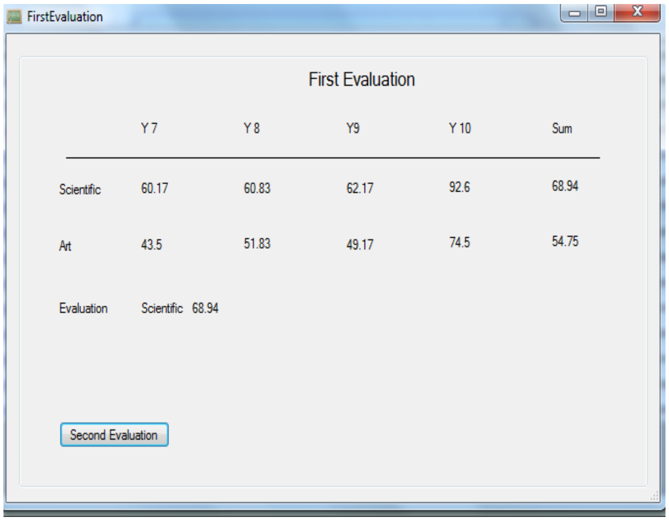
Fig. 2. Main Menu for student evaluation model

Year 11 and every academic year have its own evaluation result. Last year evaluation (Y11) depends on the accumulative previous years evaluations. Ontology based knowledge set the rules and relations between material, activities and e-games for the same section so model can decide either this material or activity or e-game is scientific or literary. Actually system didn’t allowed to extract last year evaluation which is Y11 until evaluated the last 4 years (Y7, Y8, Y9 AND Y10) which called first evaluation process as illustrated in fig.3 which shows students summation for each year at each section and the decision taken based on the comparison between each section to choose the highest which represent student interest. Then go through second evaluation as showed in fig. 4 which is on Y11 with its final result for this student. Material used for this classification for each section was as follow:

Literary section = (Geography, History, and Arabic).Science section which is divided into two sub sections which are:

- Scientific science section = (Physics, chemistry and biology).
- Mathematical science section = (Geometry and Algebra).

These materials including student activities, grades and educational games which model compare the sum for each section and take the highest.

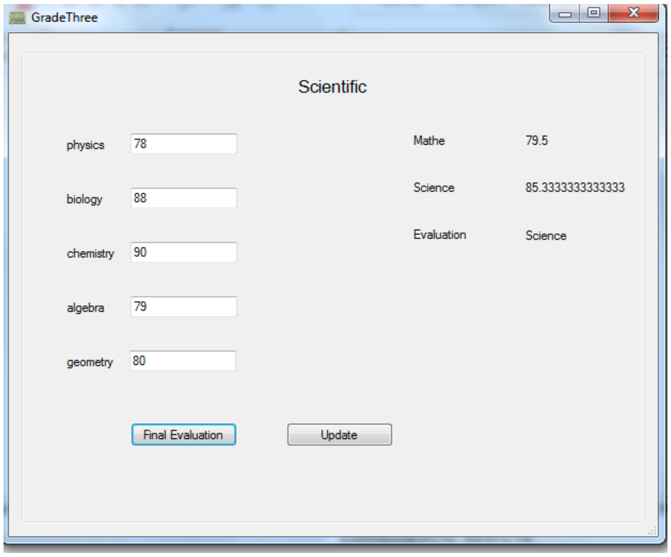


First Evaluation

	Y 7	Y 8	Y9	Y 10	Sum
Scientific	60.17	60.83	62.17	92.6	68.94
Art	43.5	51.83	49.17	74.5	54.75

Evaluation Scientific 68.94

Fig. 3. First student screen with major student behavior evaluation



Scientific

physics 78

biology 88

chemistry 90

algebra 79

geometry 80

Mathe 79.5

Science 85.3333333333333

Evaluation Science

Fig. 4. Second student evaluation screen with minor student behavior evaluation (Science or Math)

C. Results Methodology

Figure 5 illustrate the steps after collecting data from implemented model and feeding it to WEKA as a miming tool. After gathering student data either online or offline , this data get cleaned and preprocess to convert it to .ARFF file as WEKA can deal with this kind of files, then applying 10 different kinds of classification algorithms for mining students data. Finally calculate accuracy to choose the best.

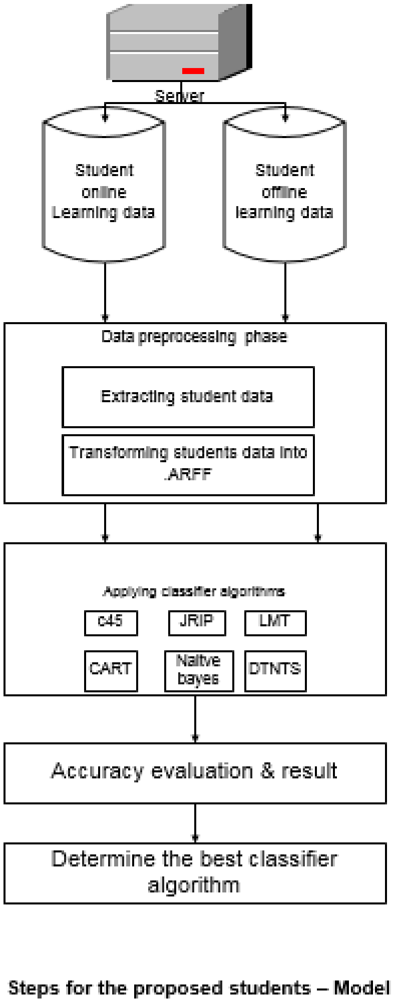

Server
Student online Learning data
Student offline learning data
Data preprocessing phase
Extracting student data
Transforming students data into .ARFF
Applying classifier algorithms
c45
JRIP
LMT
CART
Naitve bayes
DTNTS
Accuracy evaluation & result
Determine the best classifier algorithm
Steps for the proposed students – Model

Fig. 5. Data Collection from students Mining Procedure

IV. Experimental Results

The previous section described the patterns which are incorporated for each dimension as well as whether a high or low occurrence indicates a specific learning section preference. Based on this information, data about students' behavior can be used to calculate hints for specific learning section preferences. For example, if a learner often visited LMS courses or activities or games, this gives us a hint that the learner attend otherwise it gives us hints that this student is not a regular student for LMS.

There are many patterns which clarify the importance of using LMS than traditional learning such as attendance time of access.

Cavus, Uzunboylu and Ibrahim [11] underlined that a learning management system (LMS) provides the platform for web-based learning environment by enabling the management, delivery, tracking of learning, testing, communication, registration process and scheduling.

Fig.6 shows the importance of using LMS than using traditional learning and verified the advantages for using LMS as mentioned previously.

Figure 6 illustrate a comparison between different types of learning based on the access time duration (hours / month) for the three courses chosen which are Math, Science and English. Main major notification was that students prefer using LMS and E-games than traditional learning also minor notification was the number of students increased in later months in using LMS system and E-games than traditional learning. So, the traditional learning comes at third priority after LMS and E-games learning.

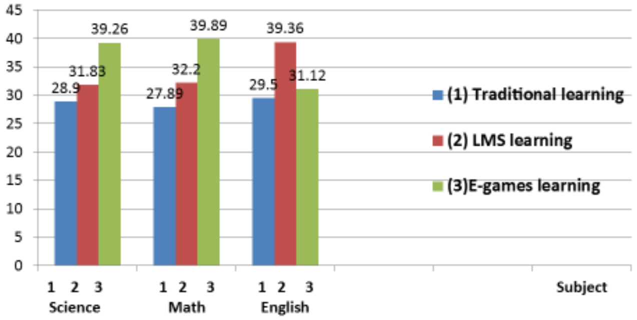


Fig. 6. Science ,Math and English courses comparision using number of access time duration (hours/month) for learning types

Classification is one of the data mining techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class with the aim of achieving least classification error. It is used to extract models that correctly define important data classes within the given dataset. It is a two-step process. In first step the model is created by applying classification algorithm on training data set.

Then in second step, the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So, classification is the process to assign class label for this dataset whose class label is unknown. Versatile list of techniques are available for classification like decision tree induction, Bayesian classification, and Bayesian network.

Figure 7 showes results for TP, FP , Precesion and recall after using WEKA for classifying students for scientific class using 10 different types of algorithms which mostly used in educational field [12].

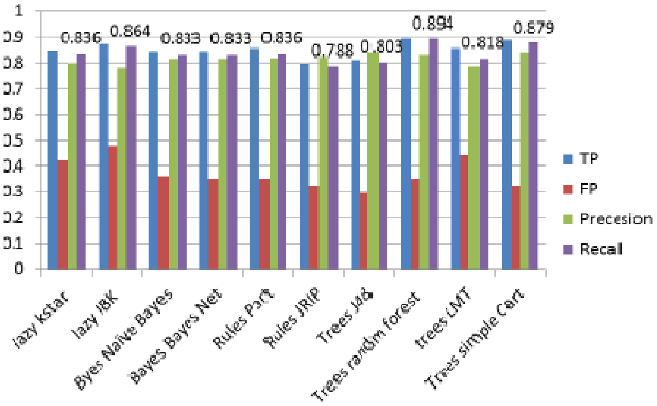


Fig. 7. Different parametes for 10 algorithms for E-game for science class

Results show that TREE RANDOM FOREST gives the best performance, then CART, LAZY IBK, LAZY K-STAR, RULES PART, Byes Naïve Bayes, Bayes Bayes Net, TREE LMT respectively. The TREE J48 algorithm comes in the Ninth place and Rules JRIP come at the end of the order.

Fig. 8 showes results for TP, FP , Precesion and recall after using WEKA for classifying students for literary class using 10 different types of algorithms which mostly used in educational field.

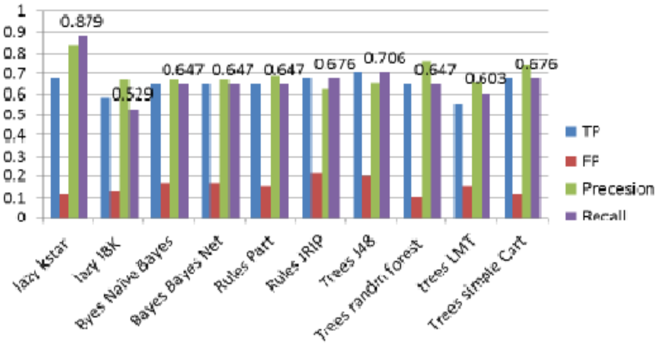
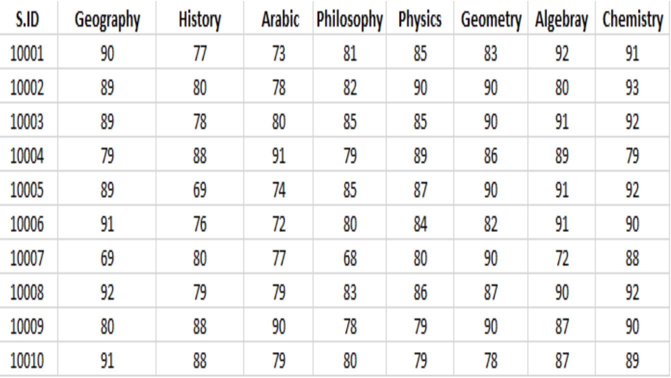


Fig. 8. Different parametes for 10 algorithms for E-game for Literary class

Results show that lazy k-star gives the best performance, then TREE J48, RULES JRIP, TREES SIMPLE CART, Byes Naïve Bayes, Bayes Bayes Net and tree random forest, rules part the same performance, respectively. The Tree LMT algorithm come in the ninth place and lazy IBK come at the end of the order.

V. Comparative Analysis

When evaluating students (N=250) which was collected from national language school in Egypt from different level of education as the sample shown in Fig.9 through the model and comparing the result deduced from the system for Y11 (second secondary school) with the data collected from result of Y12 (third secondary school) at their school. Conclusion in Table I shows the classification of each section and the percentage of each. Fig.10 displays final comparison between three types of learning included in this study which are LMS , E-games and Traditional learning, which shows that LMS has 85% accuracy (Rules Part Algorithm), E-games has 82% accuracy (Tree Random Forest Algorithm) and finally traditional learning has 80.4% (49/250=19.6% , 100 - 19.6 = 80.4%).



S.ID	Geography	History	Arabic	Philosophy	Physics	Geometry	Algebray	Chemistry
10001	90	77	73	81	85	83	92	91
10002	89	80	78	82	90	90	80	93
10003	89	78	80	85	85	90	91	92
10004	79	88	91	79	89	86	89	79
10005	89	69	74	85	87	90	91	92
10006	91	76	72	80	84	82	91	90
10007	69	80	77	68	80	90	72	88
10008	92	79	79	83	86	87	90	92
10009	80	88	90	78	79	90	87	90
10010	91	88	79	80	79	78	87	89

Fig. 9. Data set sample according to their grades

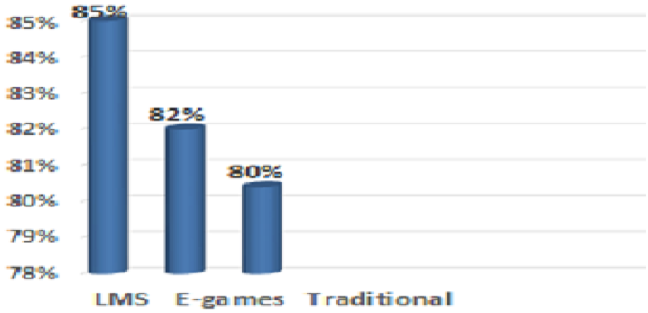


Fig. 10. Accuracy results for three types of learning in evaluation enrollment study

The results for each branch is calculated by the next equation.

Branch Success rate = number of succeeded students in section / total number of students at this section.

Math success rate = (71 / 76) * 100 = 93.4 %

Science success rate = (69/73) * 100 = 94.5 %

Literary success rate = (90 / 101) * 100 = 89.1 %

And the total average success for all the recommender system = ((Math success rate + science success rate + literary rate) / 3) * 100.

The success average percentage of the recommender system = ((93.4 + 94.5 + 89.1) / 3)*100 = 92.3 %

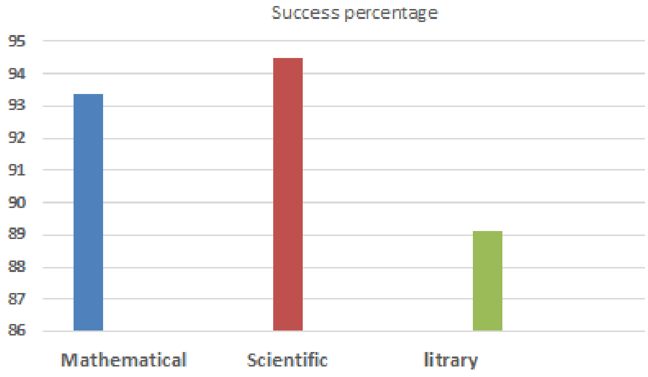


Fig. 11. System success for each branch

On the other hand real data gathered for students for Y12 was calculated as next equation:

Branch Success rate = number of succeeded students in section / total number of students at this section.

So, Math success rate = (76 / 85) * 100 = 89.4 % ,

and Science success rate = (65/71) * 100 = 91.5 % ,

and Literary success rate = (83 / 94) * 100 = 88.2 %

Then, the total average success for all the recommender system = ((Math success rate + science success rate + literary rate) / 3) * 100.

So, the success average percentage of the real data = ((89.4 + 91.5 + 88.2) / 3)*100 = 89.7 %

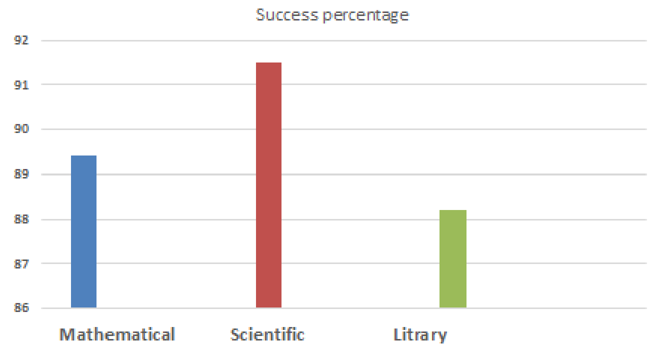


Fig. 12. Real success data for each branch

VI. Conclusions and Future Work

This work introduced an automatic student modeling approach for identifying learning skills based on LMS. The proposed model used behavior of students during they are learning in order to gather hints about their learning skills. By applying a simple rule-based mechanism, learning skills are calculated based on the gathered indications. By comparing different types of materials as online courses, online activities, and online educational games, this work can deduce the most suitable section which students can specialize on it. Improving educational games would be one of the main points for improving educational process. Compared with statistical analysis methods, this model is more effective, the process is more intelligent, and the result is more accurate. It shows that by using suggested model, teachers can understand the students better in interest, material and other information. Educational games not only improve educational process but also improve evaluation process through calculating different parameters which will be used in the future as access duration, material type and access level. The evaluation of the approach demonstrated good results and showed that the approach is suitable for identifying learning skills with respect to the new model.

Future research should include multiple schools and examine differences based on region, available resources. Future research could also be done to include undergraduate students and compare the perceptions of undergraduate, graduate students, institute students and faculty.

References

[1] Sotiris Manitsaris, Athanasios Perdos, Savvas Pavlidis, "An open – source Learning Management System (ASDL) using ICT for High Schools" , Proceedings of the Sixth International Conference on Advanced Learning Technologies , 2006 IEEE, Macedonia Universty, Thessaloniki.

[2] Sabine Graf, Kinshuk, Tzu-Chien Liu , Identifying Learning Styles in Learning Management Systems by Using Indications from Students' Behavior, Eighth IEEE International Conference on Advanced Learning Technologies, July 2008, PP 482 – 486, Santander, Cantabria.

[3] Yücel urlu, Dai Hasegawa, Hiroshi Sakuta, Student Interactions with E-learning Systems: User and Topic Analysis, 3-5April 2014., Military Museum and Cultural Center, Harbiye, Istanbul, Turkey.

[4] Dave Moursund, "Introduction to Using Games in Education: A Guide for Teachers and Parents" Teacher Education, College of Education University of Oregon 97403, 2006.

[5] Waraporn Jirapanthong , Classification Model for Selecting Undergraduate Programs , 2009 Eighth International Symposium on Natural Language Processing , IEEE , Bangkok, Thailand, pp. 89-95.

[6] Qing Yang1, Junli Sun1, Jinqiao Wang1, Semantic Web-Based Personalized Recommendation System of Courses Knowledge Research, 2010 International Conference on Intelligent Computing and Cognitive Informatics.

[7] Jili Chen, Feng Wang, Kebin Huang, Huixia Wang , E-learning Behavior Analysis based on Fuzzy Clustering , 2009 Third International Conference on Genetic and Evolutionary Computing.

[8] Andrea Sterbini , Marco Temperini, "Social Exchange and Collaboration in a Reputation-Based Educational System" , Information Technology Based Higher Education and Training , 2010 9th International conference, IEEE , PP 201-207.

[9] Yücel urlu, Dai Hasegawa, Hiroshi Sakuta , Student Interactionz with E-learning Systems: User and Topic Analysis, 2014 IEEE Global Engineering Education Conference (EDUCON).

[10] Sebastian Arnold∗, Jun Fujima∗, Andreas Karsten† and Harald Simeit†, Adaptive Behavior with User Modeling and Storyboarding in Serious Games, 2013 International Conference on Signal-Image Technology & Internet-Based Systems.

[11] Cavus, N., Uzunboylu, H., & Ibrahim, D. (2007). Assessing the success of students using a learning management system and together with a collaborative tool in Web-based teaching of programming languages. Journal of Educational Computing Research, Vol. 36(3) 301-321, 2007.

[12] Abdul Hamid M. Ragab, Amin Y. Noaman, Abdullah S. AL- Ghamd, Ayman. Madbouly," a comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining" , IDEE 14 Proceedings of the 2014 Workshop on Interaction Design in Educational Environments, ACM International Conference Processing Series.

A Random-Walk Based Privacy-Preserving Access Control for Online Social Networks

You-sheng Zhou
College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing 400065, CHINA
School of Electronic Engineering,
Dublin City University,
Dublin 9, IRELAND

En-wei Peng
College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, CHINA

Cheng-qing Guo
National Computer Network Emergency Response
Technical Team/Coordination Center of China,
Beijing 100029, CHINA

***Abstract*—Online social networks are popularized with people to connect friends, share resources etc. Meanwhile, the online social networks always suffer the problem of privacy exposure. The existing methods to prevent exposure are to enforce access control provided by the social network providers or social network users. However, those enforcements are impractical since one of essential goal of social network application is to share updates freely and instantly. To better the security and availability in social network applications, a novel random walking based access control of social network is proposed in this paper. Unlike using explicit attribute based match in the existing schemes, the results from random walking are employed to securely compute L1 distance between two social network users in the presented scheme, which not only avoids the leakage of private attributes, but also enables each social network user to define access control policy independently. The experimental results show that the proposed scheme can facilitate the access control for online social network.**

Keywords—online social networks; access control; random walk; privacy-preserving

I. Introduction

In recent years, the popularity of online social networks, such as Facebook and Twitter has grown tremendously. Users of social networks can easily establish relationships with people worldwide, and social network has become an indispensable communication platform in daily life. However, users usually obsessed with security risk when they shared pictures or news using on social network [1]. To address this problem, the traditional solution is employ attribute based access control. For example, when an online social network user A views his friend B's page, he notices a message posted by C, then he attempts to view C's page by clicking the links of B and the request of A will be sent to C. However, A is not a friend of C as shown in Fig. 1,the access control mechanism will be enforced before C making a decision on A's request. In the traditional attribute based access control scheme, A's attribute information will be requested to match the access control policy. As shown in Fig. 1, each online social network user should have multiple attributes a_i , which represent his social attributes, such as gender, age, email, contact, school etc. However, some attributes are sensible for the user, and the online social network user, such as A, is unwilling to reveal the attributes to others. Once the access control process is executed, C knows the attribute information of A. Undoubtedly; this attribute based access control method cannot prevent the user's privacy.

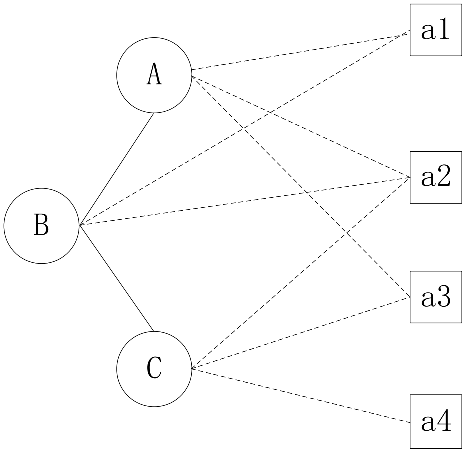


Fig. 1. Social-attribute Network

To address these privacy leaking problems, a novel random walker based access control for social network has been proposed in this paper. Unlike the classic access control schemes use attribute directly, our scheme utilizes the results of random walking as the inputs for distance protocol [2], which is used to evaluate the close relationship between two users. Furthermore, Paillier homomorphic encryption is integrated to our scheme to prevent the derivation of relationship using the results of random walking.

The rest of this paper is organized as follows: The related works are described in Section II. Section III focuses on preliminaries of truncated random walking and Paillier homomorphic encryption. In Section IV, the concrete construction of the proposed access control is introduced. Implementation of our scheme and analysis are presented in Section V. Finally, the conclusion is made.

II. Related Work

Previous researches on access control for OSNs mainly focus on social graph structure, such as [3-5]. The D-FOAF system [3] is primarily of a friend ontology-based distributed identity management system for social networks, where access rights and trust delegation management are provided as additional services. In D-FOAF, relationships are associated with a trust level, which denotes the level of friendship existing between the users participating in a given relationship. Although the work discusses only generic relationships, corresponding to ones modeled by the FOAF: knows RDF property in the FOAF vocabulary [6], another D-FOAF-related paper [7] considers also the case of multiple relationship types. As far as access rights are concerned, they denoted authorized users in terms of the minimum trust level and maximum length of the paths connecting the requester to the resource owner. In work [4], authors adopt a multi-level security approach, where trust is the only parameter used to determine the security level of both users and resources. In the work [8], a semi-decentralized discretionary access control model and a related enforcement mechanism for controlled sharing of information in OSNs is presented. The model allows the specification of access rules for online resources, where authorized users are denoted in terms of the relationship type, depth, and trust level existing between nodes in the network. Barbara [9] has proposed an extensible, fine-grained OSN access control model based on semantic web technologies, and the main idea is to encode social network-related information by means of ontology. Those works all base on classical access control, they have ignored the process of classical access control may also leak users' social attributes. Fu [10] has proposed an attribute privacy preservation scheme based on node anatomy. It allocates original node's attribute links and social links to new nodes to improve original node's anonymity, thus protects user from sensitive attribute disclosure. Meanwhile, it measures social structure influence on attribute distribution, and splits attributes according to attributes' correlations.

Random walking algorithm is used for privacy preserving widely. Pili et al. [11] designed a protocol, which transforms community detection to a series of Private Set Intersection instances using Random walking algorithm. Gabor et al. [12] introduced a light-weight protocol to quickly and securely compute the sum of the inputs of a subset of participants assuming a semi-honest adversary. In this protocol, random walkers are performed over the network. Prateek et al. [13] developed a system that mediates privacy-preserving access to social relationships. It takes users' social relationship graph as an input, then it performs Random walking algorithm to obfuscate the social graph topology.

Recently, approximation of ℓ_1 distance has been used for privacy preserving in social networks. EWPM [14] is a protocol which provides a realistic matching approach considering both the number of common interests and the corresponding weights on them. P-match [15] has been proposed to privately match the similarity with potential friends in vicinity. P-match also considers both the number of common interests and the corresponding priorities on each of them individually. Ben has proposed Weighted Average Similarity (WAS) algorithm [16], which considers both the number of common interests and the corresponding weights on them, to protect users' privacy without reliance on any Trusted Third Party.

III. Preliminaries

In this section, some preliminaries about random walking and Paillier homomorphic encryption are briefly reviewed.

A. Truncated Random Walking Model

Since there is no direct attributes in our approach, another sort of attribute-like property to define the closeness between two users should be employed, that is results from random walking. So, some preliminaries about random walking model are briefly reviewed here.

Given social network graph $G = < V, E >$, where V is a vertex set representing the social network users, and E is edge set representing the social relationships between users. The adjacent matrix is denoted as B

$$b_{ij} = \begin{cases} 1 & if < vi, vj > \in E \\ 0 & Otherwise \end{cases}$$

Then every node sends out W random walkers, and the random walker, who comes from user v_i, is denoted as w_i. And every random walker has a time-to-live (TTL) t, initially set to T, denoted as $w_{TTL} = T$, which represents the hops number of every random walker can walk on the social graph. Once a node receiving a random walker, he records the ID of w and deducts its TTL t, and sends it to a random neighbor if $t > 0$. We generate the random connection matrix as follows:

$$\Pr\left(b_{ij} = 1\right) = \begin{cases} \alpha & if\ b_{ij} = 1 \\ \beta & Otherwise \end{cases},$$

where $\beta = 1 - \alpha$. The corresponding random connection matrix is

$$B_R = \begin{pmatrix} \alpha & \beta & \alpha \\ \beta & \alpha & \beta \\ \alpha & \beta & \alpha \end{pmatrix}.$$

Then random walkers go on random walking until $t = 0$.

B. Paillier Homomorphic Encryption

The Paillier homomorphic encryption secure computation consists of the following stages.

- Key generation: The Key Generation Center (KGC) chooses two large prime numbers p and q, and computes $N = pq$ and $\lambda = \operatorname{lcm}\left(p-1, q-1\right)$. It then selects a random $g \in \mathbb{Z}^{*}_{N^2}$ such that

$\gcd\left(L\left(g^{\lambda} \bmod N^2\right), N\right)=1$, where $L(x)=\frac{x-1}{N}$. The KGC's Paillier public key is (N,g) and the private key is (p,q).

- Encryption: Let $m \in \mathbb{Z}_N$ be a plaintext and the ciphertext is given by $E_{\chi}(m)=g^m \cdot \chi^N \bmod N^2$, where $\chi \in \mathbb{Z}_N$ is a random number, and $E(\cdot)$ denotes the Paillier encryption operation.
- Decryption: Given a ciphertext $c \in \mathbb{Z}_{N^2}$. Then, the corresponding plaintext is given by

$$D\left(E_{\chi}(m)\right)=\frac{L\left(c^{\lambda} \ \bmod N^2\right)}{L\left(g^{\lambda} \ \bmod N^2\right)}=m \ \bmod N ,$$

where $D(\cdot)$ denotes the Paillier decryption operation.

Note that the entity who executes decryption does not learn the value of χ used during encryption. The Paillier cryptosystem is probabilistic and semantically secure, because χ is chosen randomly for every encryption. The Paillier cryptosystem has two useful properties.

- Homomorphic. For any $m_1, m_2, \chi_1, \chi_2 \in \mathbb{Z}_N$, we have

$$E_{\chi_1}(m_1)E_{\chi_2}(m_2)= E_{\chi_1\chi_2}(m_1+m_2) \bmod N^2 ,$$

$$E_{\chi_1}^{m_2}(m_1)=E_{\chi_1}(m_1 m_2) \bmod N_2 .$$

- Self-blinding.

$E_{\chi_1}(m_1)\chi_2^N \bmod N^2 = E_{\chi_1\chi_2}(m_1)$, which implies that any ciphertext can be modified arbitrarily without knowing the plaintext.

IV. Construction of Privacy-Preserving Access Control

A. Pre-processing of Random Walking Results

Through the execution of random walking algorithm stated in Section III, every online social network user (denoted as node v_i) should collect a set of random walkers. According to the identities of random walkers, every node can count the amount of walker u_j issued by node v_j . Next, this node forms a random walker vector $\mathbf{u}$ with u_j , whose length is equal to $|V|$.

For example, a social network has 100 users, and user Alice has obtained a random walker set $\{3rw_1, 5rw_4, 1rw_5\}$, where $3rw_1$ represents 3 random walkers come from node v_1 . Then the corresponding random walker vector of Alice can be formed as $\mathbf{u}=(3,0,0,5,1,0,\cdots,0)$, whose length is100, and the *i-th* element of vector $\mathbf{u}$ represents the amount of random walkers from v_i .

B. Computation of Closeness

With proper parameters W and T , the random walker issued by v_i will more likely reach other nodes which is more close to v_i . So that by inspecting the approximation of random walker vector $\mathbf{u}$ and $\mathbf{v}$, namely $\|\mathbf{u}-\mathbf{v}\|_1$, we can figure out how close node v_i is with another node v_j using ℓ_1 distance protocol [17].

Assume Alice is the resource owner and Bob is requestor. According to the random walk model described in section II, both of them have formed their own random walker vectors. The walker vectors of Alice and Bob are denoted as $\mathbf{u}$, $\mathbf{v}$ respectively. On one hand, Alice has to compute the approximation of vector $\mathbf{u}$ and $\mathbf{v}$ to determine how they are close before she permits Bob's request; On the other hand, to prevent the privacy of Bob, $\mathbf{v}$ should not be presented to Alice directly. Fortunately, Paillier homomorphic encryption [18] can be used to deal with this dilemma.

Since Alice and Bob have the corresponding random walker vectors $\mathbf{u}=(u_1,u_2,\cdots,u_n)$ and $\mathbf{v}=(v_1,v_2,\cdots,v_n)$, and we have

$$\|\mathbf{u}-\mathbf{v}\|_2^2=\sum_{i=1}^{n}(u_i-v_i)^2=\sum_{i=1}^{n}\left(u_i^2+v_i^2-2u_iv_i\right) .$$

One can see that Alice knows $\sum_{i=1}^{n}u_i^2$, Bob knows $\sum_{i=1}^{n}v_i^2$, but $\sum_{i=1}^{n}(-2u_iv_i)$ contains the cross terms and is unknown to both of Alice and Bob. For secure computation, Alice generates a public/ private key pair and shares only public key with Bob. Alice and Bob will follow the steps of the protocol for secure computation of the squared ℓ_2 distance as below.

For every $i \in (1,2,\cdots,n)$, Alice encrypts u_i into $E_{\chi_i}(u_i)$ according to the encryption process of Paillier cryptosystem. Here, $\chi_i \in \mathbb{Z}_N^*$ is chosen randomly. Then Alice transmits the encrypted vector $E_{\chi}(\mathbf{u})$ to Bob.

For every $i \in (1,2,\cdots,n)$, Bob computes

$$E_{\chi_i}^{-2v_i}(u_i) \bmod N^2 \equiv E_{\chi_i}(-2u_iv_i) .$$

Bob computes

$$E_{\chi_C}\left(\sum_{i=1}^{n}(-2u_iv_i)\right) \equiv \prod_{i=1}^{n}E_{\chi_i}(-2u_iv_i) \bmod N^2 ,$$

where $\chi_C=\prod_{i=1}^{n}\chi_i \bmod N \in \mathbb{Z}_N^*$. Note that Bob operates solely in the encrypted domain in this step, so the values of $\sum_{i=1}^{n}(-2u_iv_i)$ and χ_C are unknown to him.

Bob chooses $\chi_B \in \square_N^*$ randomly, and $\chi_D = \chi_B \chi_C \bmod N \in \square_N^*$. Then, he computes

$$E_{\chi_D}\left(\sum_{i=1}^{n} v_i^2 + \sum_{i=1}^{n}(-2u_i v_i)\right) \equiv E_{\chi_B}\left(\sum_{i=1}^{n} v_i^2\right) E_{\chi_C}\left(\sum_{i=1}^{n}(-2u_i v_i)\right) \bmod N^2$$

Bob transmits this result to Alice. One can see that the value of χ_D is implicit in the encryption result but is unknown to Bob, since he does not know the value of χ_C.

Alice chooses $\chi_A \in \square_N^*$ randomly and $\chi = \chi_A \chi_D$, then she computes

$$E_{\chi}\left(\|\mathbf{u}-\mathbf{v}\|_2^2\right) = E_{\chi}\left(\sum_{i=1}^{n}\left(u_i^2 + v_i^2 - 2u_i v_i\right)\right)$$
$$\equiv E_{\chi_A}\left(\sum_{i=1}^{n} u_i^2\right) E_{\chi_D}\left(\sum_{i=1}^{n} v_i^2 + \sum_{i=1}^{n}(-2u_i v_i)\right) \bmod N^2$$

Note that, the value of χ is also implicit in the encryption result but unknown to Alice because she does not know the value of χ_B.

Alice decrypts $\sum_{i=1}^{n}\left(u_i^2 + v_i^2 - 2u_i v_i\right) = \|\mathbf{u}-\mathbf{v}\|_2^2$ using the private key according to the decryption process of Paillier cryptosystem.

We can see that, this protocol does not reveal $\mathbf{v}$ to Alice or $\mathbf{u}$ to Bob.

C. Decision on Request

After the execution of computation of closeness, Alice would obtain the value of $\|\mathbf{u}-\mathbf{v}\|_2^2 \approx \|\mathbf{u}-\mathbf{v}\|_1$. Then, she can make the decision by checking whether $\|\mathbf{u}-\mathbf{v}\|_1 \le \tau_A$, where τ_A is a permissible threshold set by Alice herself. If yes, she will allow Bob to access her data. Otherwise, she declines the request.

V. Implementation and Evaluatuon

A. Data Sets and Preparation

We have implemented a preliminary prototype of the proposed scheme, which provides access control with privacy preserving. We use the Facebook friendship graph from the New Orleans regional network [20] to simulate the social graph in our scheme. This dataset describes the links between users from the Facebook New Orleans network, consisting of 63,732 nodes and 1.545 million edges. To show our experiment results clearly, only 100 access requests are shown.

B. Results

Three parameters (W, T, τ_i) need to be set initially before the experiments, where W is the random walker number of every node have issued, T represents time-to-live (TTL) of every random walker, and τ_i is the permissible threshold value set by node v_i.

TABLE I. Approximation from ℓ_1 Distances

1	6	26	12	51	12	76	2
2	8	27	3	52	14	77	16
3	1	28	12	53	12	78	14
4	7	29	20	54	18	79	20
5	10	30	14	55	14	80	8
6	12	31	3	56	17	81	12
7	9	32	12	57	8	82	9
8	16	33	8	58	17	83	14
9	1	34	16	59	2	84	3
10	18	35	3	60	10	85	17
11	2	36	14	61	12	86	18
12	15	37	16	62	4	87	14
13	6	38	7	63	17	88	10
14	4	39	17	64	5	89	3
15	9	40	10	65	20	90	16
16	14	41	3	66	16	91	17
17	1	42	17	67	4	92	10
18	10	43	1	68	1	93	17
19	14	44	17	69	14	94	14
20	9	45	5	70	15	95	18
21	12	46	20	71	17	96	8
22	4	47	10	72	8	97	12
23	8	48	8	73	1	98	16
24	14	49	12	74	16	99	1
25	1	50	18	75	12	100	12

Note that, in order to investigate how much the variation of parameters would affect the access control, we also set the value of τ uniformly. When setting $W=10$ and $T=10$, we have obtained100 approximation of ℓ_1 distances, who are shown in Table I.

We employ a variety of parameters combination to observe the influence on results. Firstly, we set $(W, T, \tau_i) = (10, 10, 5)$, the outcome is shown as in Fig.2, which depicts the number of passed the closed-relationship verification of access requests, such as the first dot in Fig.2 represents two pairs of nodes could pass the access control when there are ten access requests.

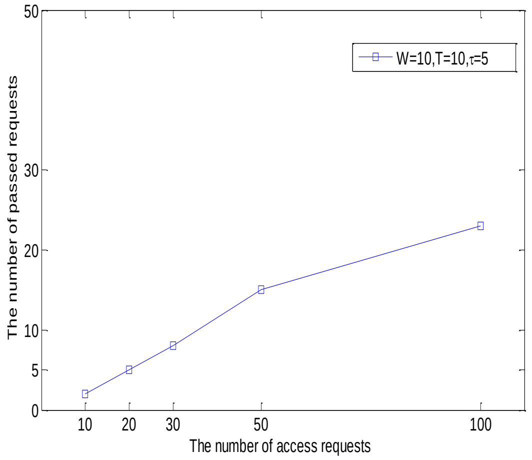


Fig. 2. The number of passed the closed-relationship verification of access requests

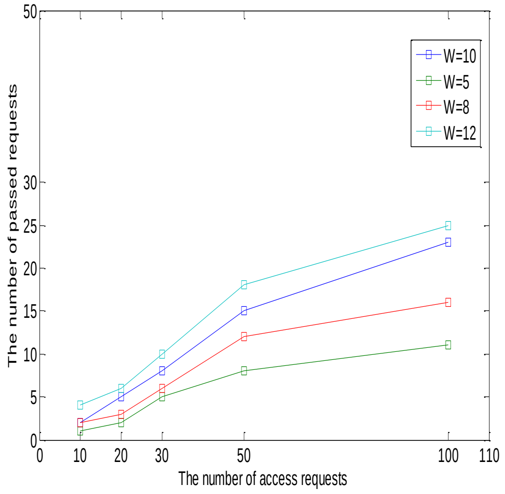


Fig. 3. The number of passed the closed-relationship verification influenced by W

To figure out how much parameter W would affect the access control, we vary the value of W while keep T and τ stable. We set $(W,T,\tau)=(5,10,5)$, $(W,T,\tau)=(8,10,5)$ and $(W,T,\tau)=(12,10,5)$. As shown in Fig.3, the amount of passed requests is consistent with the variation of parameter W . This phenomenon is in accordance with our theoretical study. If W decrease, the amount of collected random walkers by every node would decrease as well. Since the number of common random walker is smaller, the amount of passed requests would be smaller than before when it proceeds to the computation of closeness.

Next, we investigate the influence from T .we set $(W,T,\tau)=(10,5,5)$, $(W,T,\tau)=(10,15,5)$ and $(W,T,\tau)=(10,20,5)$. The outcome shows as in Fig.4.

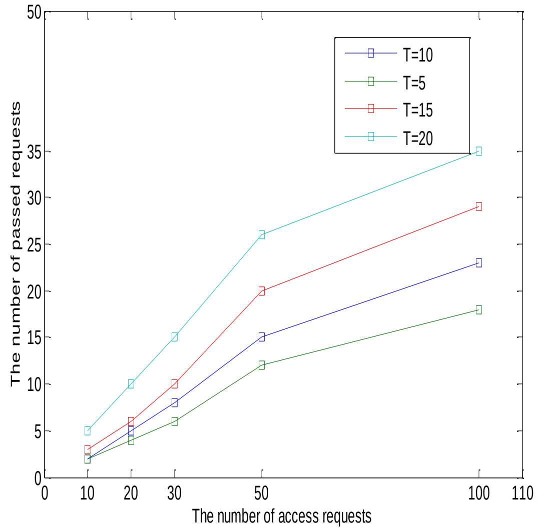


Fig. 4. The number of passed the closed-relationship verification influenced by T

One can see that the variation of the amount of passed requests is also consistent with the variation of parameter T . However, we have observed that the influence from T is much smaller than W 's when we decrease the same value of W and T . This is caused by chose relationship. So we have increased variation of T .

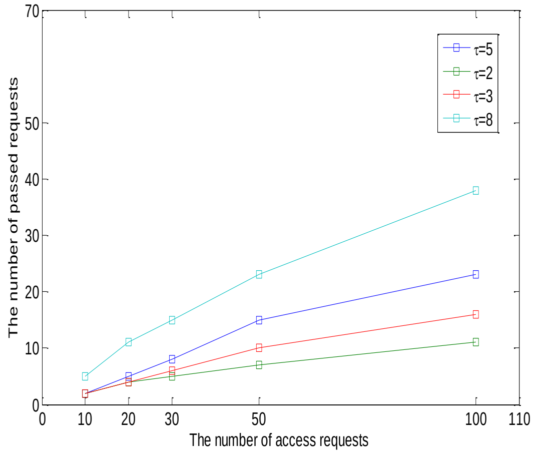


Fig. 5. The number of passed the closed-relationship verification influenced by τ

At last, we vary the value of parameter τ . We set $(W,T,\tau)=(10,10,2)$, $(W,T,\tau)=(10,10,3)$ and $(W,T,\tau)=(10,10,8)$. Fig.5 depicts the variation of the amount of passed requests is in the opposite trend of τ 's variation. This result is also consistent with our theoretical analysis. Although the amount of random walkers remains unchanged, when set the permissible threshold τ to a smaller value, this means only those has much closer relationship with the owner can be allowed to access, so that the number of passed requests is smaller than before.

Adamic et al. [21] has proposed a classical scheme (Adamic-Adar for short) to measure similarity between two users, and their scheme is based on common neighbors and the degrees of those common neighbors. The formulation expression of Adamic/Adar is:

$$Similarity(u,v)=\sum\nolimits_{i\in\Gamma(u)\cap\Gamma(v)}\frac{1}{\lg|\Gamma(i)|},$$

where $\Gamma(i)$ denotes the set of neighbors of node i in social graph. We have employed Adamic-Adar to evaluate the accuracy of our scheme. To find out which parameters setting is more practical for our proposed scheme in reality. We have counted the distribution of the similarity value in Adamic-Adar and our scheme. We have sampled the first ten percent, the middle ten percent and the last ten percent of the 100 similarity values to compare our scheme clearly.

After executing our scheme and computing the similarity of node pairs according to Adamic-Adar, we get the best parameter setting shown as in Fig.6. One can see that the accuracy of our scheme is almost equal to the outcome of Adamic-Adar scheme when setting $W=10, T=10$.

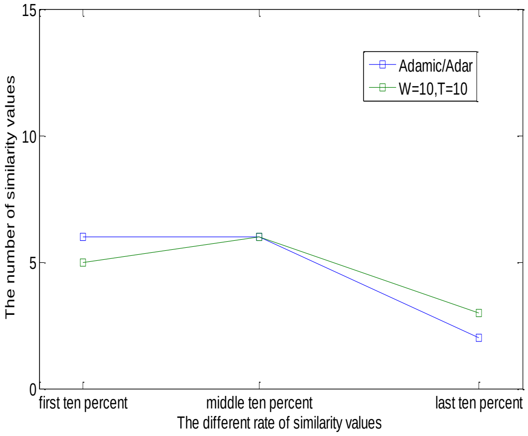


Fig. 6. The best parameter setting

VI. Conclusion

In this paper, a random walk based access control scheme is investigated in this paper. . The proposed novel approach employs random walking to form the profile for online social users. In terms of the formed profile, users can carry out access control according to the secure computation of closeness. Furthermore, the user can set the permissible threshold independently according to his access policy. In this way, the leakage of privacy existing in traditional attribute based access control has been removed. Experimental results show that the proposed scheme is reasonable and practical. In our future work, more efficient approach for computation of closeness will be investigated.

Acknowledgement

This work was jointly supported by the National Social Science Foundation of China (no. 14CTQ026), the National Natural Science Foundation of China (no. 61472464), the Chongqing Research Program of Application Foundation and Advanced Technology (no. cstc-2014jcyjA40028, no. cstc-2013jcyjA40017), the Natural Science Foundation of Shandong Province, China (no. ZR2015FL024).

Reference

[1] BCarminati, E Ferrari, M Viviani. Security and trust in online social networks[J]. Synthesis Lectures on Information Security, Privacy, & Trust, vol.4, no.3, pp.1-120, 2013.

[2] R. Shantanu, S.Wei and V. Anthony,"Privacy-preserving approximation of L1 distance for multimedia applications," Multimedia and Expo (ICME), 2010 IEEE International Conference, Singapore,pp. 492-497, 2010.

[3] S.R.Kruk, S.Grzonkowski, A.Gzella, T.Woroniecki andH.C.Choi, "DFOAF: distributed identity management with access rights delegation,"In The Semantic Web–ASWC 2006, Springer Berlin Heidelberg, pp.140-154,2006.

[4] B.Ali, W.Villegas and M.Maheswaran,"A trust based approach for protecting user data in social networks,". In Proceedings of the 2007 conference of the center for advanced studies on Collaborative research,IBM Corp, pp.288-293, 2007.

[5] B.Carminati, E.Ferrari and A.Perego,"Security and privacy in social networks," Encyclopedia of information Science and Technology, pp 3369-3376, 2008.

[6] D. Brickley andL. Miller, "FOAF vocabulary specification 0.91," Available at, http://xmlns.com/foaf/0.1, 2007.

[7] H.C.Choi, S.R. Kruk, S.Grzonkowski, K. Stankiewicz, B.Davids andJ.G Breslin,"Trust models for community-aware identity management," IRW2006/WWW2006 Workshop, 2006.

[8] B. Carminati, E. Ferrari and A. Perego,"Enforcing access control in webbased social networks,"ACM Transactions on Information and System Security (TISSEC), 13(1), 6, 2009.

[9] C. Barbara, F. Elena, H. Raymond, K. Murat andT. Bhavani, "Semantic web-based social network access control," Computers & Security In ELSEVIER, pp. 108-115, 2011.

[10] Y.Y.Fu, M. Zhang, D.G.Fengand K.Q. Chen,"Attribute privacy preservation in social networks based on node anatomy,"RuanJianXueBao/Journal of Software, pp. 768−780, 2014.

[11] H.Pili, S.M.C Sherman, C.L. Wing,"Secure friend discovery via privacy-preserving and decentralized community detection," In ICML 2014 Workshop on Learning, Security and Privacy, 2014.

[12] D.Gabor and J. Mark, "Fully distributed privacy preserving mini-batch gradient descent learning," International Federation for Information Processing, pp. 30-44, 2015.

[13] L. Changchang and M. Prateek,"LinkMirage: Enabling privacy-preserving analytics on social relationships," In NDSS, 2016.

[14] Z.Xiaoyan, L.Jie, J. Shunrong, C. Zengbaoand L. Hui, "Efficient weight-based Private Matching for proximity-based mobile social network," In IEEE ICC, 2014.

[15] N.Ben, Z. Xiaoyan, Z. Tanran, C. Haotian and P. Hui,"P-match: Priority-aware frirnd discovery for proximity-based mobile social networks," In 2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systenms,Hangzhou, pp. 4114-4119, 2013.

[16] L.Ben, Z. Xiaoyan, L. Jie, L. Zan and L.Hui,"Weight-aware private matching scheme for proximity-based mobile social network," In Globecom 2013- Symposium on Selected Areas in Communications, pp. 3170-3175, 2013.

[17] W.Du, M. Atallah and F. Kerschbaum,"Protocols for secure remote database access with approximate matching,"the 7th ACM Conference on Computer and Communications Security, Athens, pp. 523–540, 2000.

[18] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," in Advances in Cryptology, EUROCRYPT 99. 1999, vol. 1592Springer-Verlag, Lecture Notes in Computer Science, , pp. 233–238, 1999.

[19] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz Mapping Into Hilbert Space,"Contemporary Mathematics, vol. 26, pp. 189–206, 1984.

[20] B.Viswanath, A. Mislove, M. Cha and K.P. Gummadi,"On the evolution of user interaction in Fcaebook,". In Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks, Barcelona, pp. 37-42, 2009.

[21] L.A Adamic andE. Adar,"Friends and neighbors on the Web,"Social Networks, 25(3), pp. 211–230, 2003.

Pricing Schemes in Cloud Computing: An Overview

Artan Mazrekaj
Department of Computer Engineering
Faculty of Electrical and Computer Engineering
University of Prishtina
Prishtina, Republic of Kosovo

Isak Shabani
Department of Computer Engineering
Faculty of Electrical and Computer Engineering
University of Prishtina
Prishtina, Republic of Kosovo

Besmir Sejdiu
Department of Computer Engineering
Faculty of Electrical and Computer Engineering
University of Prishtina
Prishtina, Republic of Kosovo

***Abstract*—Cloud Computing is one of the technologies with rapid development in recent years where there is increasing interest in industry and academia. This technology enables many services and resources for end users. With the rise of cloud services number of companies that offer various services in cloud infrastructure is increased, thus creating a competition on prices in the global market. Cloud Computing providers offer more services to their clients ranging from infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS), storage as a service (STaaS), security as a service (SECaaS), test environment as a service (TEaaS). The purpose of providers is to maximize revenue by their price schemes, while the main goal of customers is to have quality of services (QoS) for a reasonable price. The purpose of this paper is to compare and discuss several models and pricing schemes from different Cloud Computing providers.**

Keywords—Cloud Computing; Pricing Models; Pricing Schemes

I. Introduction

Cloud Computing is a new paradigm which has changed the traditional business schemes/plans and incorporating new economic and financial models of IT services market.

This technology allows end users to process, store and manage their data efficiently with fast and reasonably price.

Cloud computing customers do not need to install different software and they could access their data wherever they are via the Internet.

There are different definitions for Cloud Computing, Foster et al. [1] defines Cloud Computing as ”a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet”.

Cloud Computing Providers offer numerous online services based on SLA (Service Level Agreement) between the provider and the customer.

However an important role between providers and customers relationship has pricing model for which they must agree.

Each provider has his scheme for calculating the price (has an accounting system) for the cloud services offered for clients. The provider’s goal is to have a greater benefit, while each client’s goal is to have the maximum service for low price.

Therefore, satisfying both parties requires an optimal pricing methodology. The price charged is one of the most important metrics that a service provider can control to encourage the usage of its services [2].

Price is an important factor for the company which provides cloud services because it affects the clients directly and organization profit.

The price also has a major impact in economic aspect, where key concepts such as fairness and competitive pricing in a multi-provider marketplace affect the actual pricing [18].

Pricing for competition and fairness affects choices in the design of user applications and system infrastructures.

In fact pricing fairness balances user cost and cloud service provider profit.

Pricing model in Cloud Computing is more flexible than traditional models. Every cloud provider has its own pricing scheme. Main focus of Cloud Computing is to fulfill and guarantee quality of service (QoS) for customers.

The price in Cloud Computing and value chain is based on business models and framework. The value chain from the traditional IT services is changing as a result of cloud computing.

This issue is illustrated from authors Jaekel and Luhn [30], which is shown in the figure below.

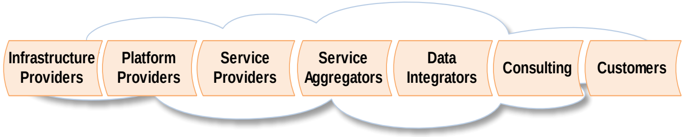


Fig. 1. Cloud Computing Value Chain

The key issue is how price / costs are measured, accounted, and distributed between different service layers and organizational units responsible of them.

There are many business models based on different service models that determine the price of services in the cloud.

For example below we see an illustration of a model for Cloud Computing cost accounting [14] which addresses cost accounting issues in production of cloud services.

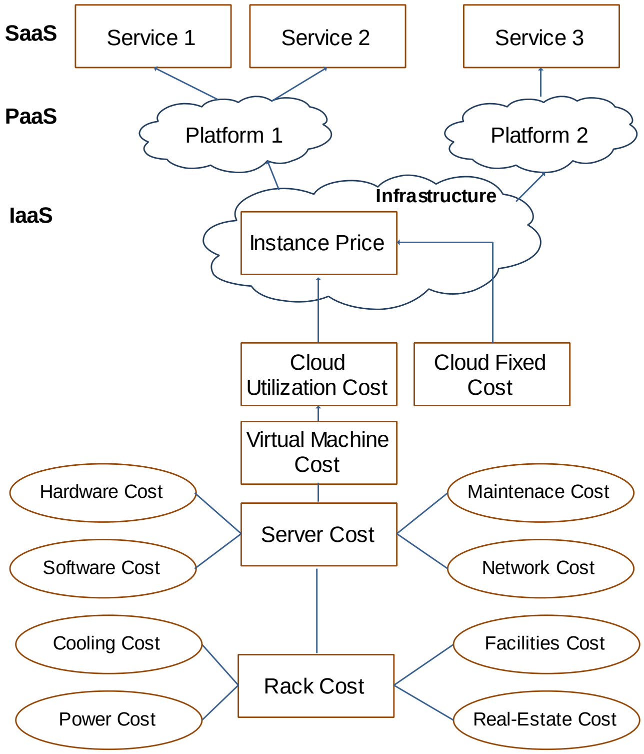


Fig. 2. Cloud Computing cost accounting model

In this paper we focus on review and comparing prices of some models, some pricing schemes which are provided by cloud service providers, based on services provided, their quality, fairness price and significance in the market.

II. Related Work

In this section we discuss related work with regard to the pricing schemes in Cloud Computing.

The issue of price in the cloud in terms of relationship between providers and customers is treated by many authors, which have analyzed different schemes and models in theoretical aspects and simulated through different software.

Sharma et al., [3] proposed a novel financial economic model capable of providing a high level of QoS to customers.

They developed a financial option theory treating cloud resources as assets. The price determined by their model represents an optimal price where the provider charges the client in order to cover the initial cost.

Also, indirectly they have used Moore's law to determine the price of resources in the cloud and Black-Scholes-Merton (BSM) model that treats cloud resources as assets.

Through their experiments and simulations, they analyzed the effect of initial investment, effect of contract period, effect of rate of depreciation, effect of quality of service, effect of age of the resources on the resource price. The authors focused on the initial price however did not take under consideration the maintenance costs.

Patel and Shah [4] studied for prices caused by datacenters, which focused on three issues: space, power and cooling on cost model. They analyzed the price for each of the three cases and the sum of these prices to show a price comparison operating in datacenter. The authors of this study do not go any further in finding the cost of Cloud resources meant to be sold as a service.

Pal and Hui [5] have studied an economic model for fixing prices of resources. They used game theory and have presented some economic models. In first model, QoS guarantees provided by cloud provider are pre-specified and fixed competitors compete for prices. In another model, cloud providers compete for QoS level as well as prices for a particular application.

Wang et al. [6] proposed an algorithmic solution to optimize data center net profit with deadline-dependent scheduling by jointly maximizing revenues and minimizing electricity costs. They developed two distributed algorithms for the net profit optimization: Net Profit Optimization for Divisible jobs (NPOD), and Net Profit Optimization for Indivisible Jobs (NPOI). The authors proved via simulations their algorithm's capabilities to increase revenues and reduce electricity costs by comparing it to the Largest Job First (LJF) algorithm. The authors considered only static job arrivals and departures. They also assumed that the servers at all data centers were homogenous, which is not realistic [2].

Yeoa et al. [7] analyzed difference between fixed and variable prices. Fixed prices were easier to understand and more straightforward for users. However, fixed prices could not be fair to all users because not all users had the same needs. They proposed charging variable prices with advanced reservation, in which case users know the exact expenses that are computed at the time of reservation even though they were based on variable prices.

Macias and Guitart [8] proposed a genetic model for pricing in cloud computing markets. Choosing a good pricing model via their genetic algorithms involved three main steps: define a chromosome, evaluate it, and finally select the best pairs of chromosomes for reproduction and discarding those with the worst results. The results of the simulation illustrated that genetic pricing acquired the highest revenues in most of the scenarios. Service providers employing genetic pricing achieved revenues up to 100% greater than the other dynamic pricing strategies and up to 1000% greater than the fixed pricing strategy.

Li at al. [19] proposed a pricing algorithm for cloud computing resources. Authors proposed the cloud bank agent model as a resource agency from the global perspective, which provides analysis and guidance for all members.

The model analyses the historical utilization ratio of the resource, and iteration current prices constantly, get the availability of resources next time, the final price to users are expected to calculate. The proposed pricing model could not adapt to the rapid changes that occur in the market.

III. Pricing Schemes in the Cloud

Here we present an overview of pricing schemes from the perspective of the accounting process and the relevance from the business model. There are various pricing schemes depending on the cloud service provider. The challenge of service providers is to provide good services for reasonable price to users. The pricing should be based on customer's perceived value instead of production costs of services.

Some of the definitions and short description of pricing schemes and which vary depending on the services are [11]:

- *Time based*, pricing based on how long a service is used;
- *Volume based*, pricing based on the volume of a metric;
- *Flat rate*, a fixed tariff for a specified amount of time.
- *Priority pricing*, services are labeled and priced according to their priority;
- *Edge pricing*, calculation is done based on the distance between the service and the user;
- *Responsive pricing*, charging is activated only on service congestion;
- *Session-oriented*, based on the use given to the session;
- *Usage-based*, based on the general use of the service for a period of time, e.g. a month;
- *Content-based*, based on the accessed content;
- *Location-based*, based on the access point of the user;
- *Service type*, based on the usage of the service;
- *Free of charge*, no charge is applied for the services;
- *Periodical fess*, payment of time to time quantities for the use of a service;
- *Pre-paid*, the payment of the service is done in advance.
- *Post-paid*, the payment of the service is done after the use;
- *Online*, the accounting performed while the user makes use of a service;
- *Offline*, the accounting process is done after a service is used;

A. Fixed Pricing

Each service provider defines price for resources that could be prohibitive and thus lead to a reduced customer base and decrease in revenue and profits. Fixed pricing includes pricing mechanism as pay-per-use pricing, subscription and list price / menu price [14].

Pay-per-use pricing, users only have to pay for what they use. Customer pays in function of the time or quantity he consumes on a specific service. Pay-per-use makes users aware of the cost of doing business and consuming a resource. In the following table are presented some of the pricing schemes for some providers for pay-per-use pricing mechanism [16].

TABLE I. Pricing Schemes for Pay-Per-Use

Service	**Pricing scheme**
Amazon Web Services / Elastic Compute Cloud (EC2)	• Charges on hourly for usage the RAM, CPU • Charge per transferred GB basis
Amazon Web Service / Simple Storage Service (S3)	- Charge per GB of storage - Charge per data transferred in GB
Microsoft Azure	- Charge on hourly basis for processing power and on per GB for storage
AppNexus / AppNexus Cloud *Source*: appnexus.com	- Charge on hourly basis

Subscription pricing, users pay on a recurring basis to access software as an online service or to profit from a service.

The customer subscribes to use a preselected combination of service units for a fixed price and a longer time frame, usually monthly or yearly [15].

In the following table are presented some of the pricing schemes for some providers for subscription pricing mechanism.

TABLE II. Pricing Schemes for Subscription

Service	**Pricing scheme**
Dropbox / Dropbox cloud storage	Pricing assigned for stored in GB [17]
Google / App Engine	- Charges on monthly basis - Charge per user basis - Pricing on hourly basis
Amazon Web Services / Elastic Compute Cloud (EC2)	It allows for the reservation of units
Microsoft Azure	- Charges on monthly basis, depending from number of transactions - Charges on monthly basis subscription for database *Source:* Microsoft Azure website
Salesforce / Salesforce.com	- price assigned on monthly basis per user

Hybrid pricing model (pay-per-use plus subscription), in this model between per-per-use and subscription, dedicated servers must be provided in advance for a period of time [16].

In the following table are presented some of the pricing schemes for some providers for hybrid (pay-per-use plus subscription) pricing mechanism.

TABLE III. Hybrid Pricing Schemes

Service	**Pricing scheme**
Google / App Engine	- Price assigned on monthly basis - If limit exceeds then charge on per GB and processing power on hourly basis
Joynet / Smart Machines *Source*: joynet.com	- Price assigned on monthly basis for the package - If the usage exceeds the limit the charges on per GB

List Price / Menu Price, is a fixed price that is often found in a list or catalog.

B. *Dynamic Pricing*

The price is calculated based on pricing mechanism whenever there is a request. In some cases, the price of the resources is determined according to demand and supply [9]. As compared to fixed prices, the dynamic pricing that reflects the real-time supply demand relationship represents a more promising charge strategy that can better exploit user payment potentials and thus larger profit gains at the cloud provider [13].

C. *Market-Dependent Pricing*

Customer pays depending on the real-time market conditions and constraints. This schemes includes:

Bargaining, the price is determined on the basis of the relationship of the parties involved.

Yield Management, the best pricing policy for optimizing profits is calculated based on real-time modeling and forecasting of demand behavior [14].

Auction, is a negotiation mechanism which allows both parties to communicate and to agree on the offer. The price is set as buyers bid in increasing increments of price.

Dynamic Market, in that case buyers and sellers determine their price reference, but are not able to influence this price as individual sellers.

IV. Pricing Models in the Cloud

The pricing in Cloud Computing has its root in system design and optimization. Resource's consumption based pricing is particularly sensitive to how a system is designed, configured, optimized, monitored, and measured. Cloud services vendors use a variety of pricing mechanisms, including usage-based fixed pricing, usage-based dynamic pricing, subscription-based pricing, reserved services contracts with a combination of usage-based fixed pricing and up-front fees, auction-based pricing, etc. [12].

Also pricing is more important in economic terms as fairness and competitive pricing in a multi-provider marketplace affect the actual pricing [10].

Pricing presents exchange process when customer/end user pays for services which have been offered by the service provider. Some of the most common factors affecting pricing in the cloud resources are presented in table IV.

Also there are other factors which affect the price in the cloud resources. These factors could be fixed or variable. Some of these factors that influence the price of cloud resources are presented in figure3.

Monitoring Service, few Cloud Providers have the confidence to provide customers with monitoring tools for service availability [28]. Monitoring services could be managed from the providers or a third party.

TABLE IV. Most Common Factors that Affect in Pricing

Initial Cost / Investment	Represent the amount of money that Cloud service provider will spend per year, to buy a resource.
Lease Period / Contract Time	It is the time in which the client will lease resources from the cloud service provider. Cloud service providers usually offer lower unit prices for longer subscription periods.
Quality of Service	This factor represent the quality assurance from cloud service provider for the customer. The key aspects of quality of Service (QoS) are: integrity of service provider, availability, security, privacy, scalability. For the better quality of service the price will be higher.
Rate of depreciation [3]	It is the rate at which the hardware of service provider is expected to lose its financial value.
Age of Resources	It represents the age of a particular resource the service provider is leasing to the client [3].
Cost of Maintenance	Represent the amount of money per year that the cloud service provider spends to maintain and secure the cloud.

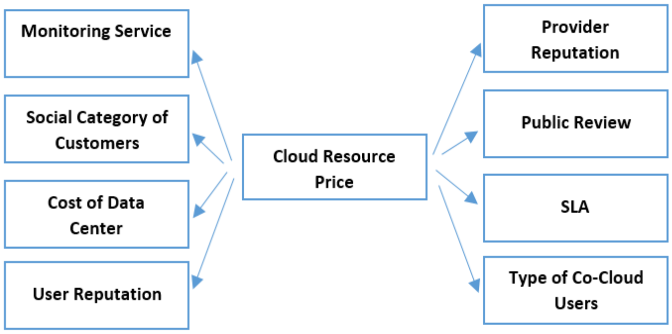


Fig. 3. Some factors that influence the price of cloud resources

Social Category of Customers, all clients should be offered a fair price, however, it should be viewed social aspect of clients or social classifications. Classification should be done depending on client's location.

Cost of Data Center, the price should be calculated for data centers, as cost of real estate, backup power, maintenance, cooling resources, network connectivity, security features etc.

User Reputation, the reputation of the users has a special importance in cloud services considering various attacks, sniffing programs, Trojans etc.

Provider Reputation, Cloud provider's reputation is also necessary to create a trust from the community when it is known that may have sensitive data. The reputation is the component of trust and it also measures reliability. Using Cloud infrastructure for critical business computation necessitate that the reputation of the Cloud provider is well established [28].

Public Review, public reviews on issues such as downtime, phishing, and data loss and password weakness can be valuable in pricing of cloud services [28].

SLA (Service Level Agreement) is a negotiated agreement for services between Cloud providers and cloud costumers. Most often SLAs are dictated by the Cloud Providers [29].

Co-Cloud Users, the nature of multi-tenancy in a Cloud could enable competitive companies to use the same Cloud platform. Information about co-tenants in the Cloud can be used to influence service price.

The service price could be affected if the information about co-tenants in the Cloud is used.

The table below compares some pricing model.

TABLE V. COMPARISON OF SOME PRICING MODELS

Pricing Model	**Description**	**Features and Fairness**	**Implementation**
Pay-as-you-go [2]	Price is set by service provider and remains constant. This model is static.	Unfair to the client. He might pay more than necessary	Implemented [20 , 21]
Subscription	Price assigned based on subscription. This model is static.	According to this model client sometimes can charge more or less.	Implemented [20 , 21]
Dynamic Resource Pricing [22]	It is a dynamic pricing model used for federated cloud and supports various resource types.	In this model, resource payments are assigned based on demand and supply.	Theoretical study with simulation
Pay-for-resources	This model (static) is cost based. Offers maximum utilization for resources	Is fair for client and cloud service providers.	Implemented [20, 21]. It is hard to implement.
Hybrid pricing	Price changed according to the job queue wait times [2]	This model is fair to clients.	The model is implemented
Dynamic Auction [23]	This model based on truthfulness and dynamic adjustment.	Tolerate fluctuation of users' distributions.	Theoretical study with simulation
Double Auction Bayesian Game-Based [24]	This model enables people to buy resources from various providers	Free resources can be exchanged with more flexibility.	Theoretical study.
Double sided Combinational Auctions to Resource Allocation [9, 25]	This model is for service allocation that enables users and providers to deal through double-sided combinational auction.	Users and service providers should be satisfied by the resource allocation mechanism.	Theoretical study with simulation
Pricing algorithm for cloud computing resources [19]	This model based on real time pricing. This model is dynamic.	Fair for provider because it reduces costs and maximizes revenues	Theoretical study with simulation
Genetic model for pricing in cloud computing markets [8]	This model based on real time pricing. This model is dynamic.	The algorithm increasing revenues for service providers.	Theoretical study with simulation
Value-based pricing	Price assigned on client's perceive basis. This model is dynamic.	Fair to producers where prices are set on the value perceived by the client	The model is implemented
Cost-based pricing [26]	In this model the priority is to increase the profit.	It is not difficult to calculate the price. Client role is not primary.	The model is implemented
Competition-based pricing [27]	Price assigned according to competitors' prices. This model is dynamic.	The model is fair to clients if the price is assigned on the basis of competition.	The model is implemented
Customer-based pricing [11]	Price assigned according to what client want and the need to pay.	This model is fair for clients if there are taken into account.	The model is implemented
A novel financial economic model [3]	This model is dynamic, which based on usage.	Is fair for service provider and client. Provides a high level of Quality of Service for clients.	Theoretical study with simulation

The following we present some pricing structure examples for some services.

An example of IaaS is Amazon S3, which is an online storage web service offered by Amazon Web Services. Amazon Web Services uses *Amazon Spot Instances* to allow customers to bid for their unused capacity. Amazon runs the customer's instances as long as the bid price is higher than the spot price, which is set by Amazon based on their data center utilization [15].

The pricing structure (pay-per-use pricing) of several Amazon S3 services is shown in the table below.

TABLE VI. AMAZON S3 PRICING

	Standard Storage	**Reduced redundancy Storage**
First 1 TB / month	$0.0390 per GB	$0.0312 per GB
Next 49 TB / month	$0.0383 per GB	$0.0306 per GB
Next 450 TB / month	$0.0377 per GB	$0.0301 per GB
Next 500 TB / month	$0.0370 per GB	$0.0296 per GB
Next 4000 TB / month	$0.0364 per GB	$0.0291 per GB
Over 5000 TB / month	$0.0357 per GB	$0.0285 per GB

Source: Amazon website

An another example of PaaS is Google App Engine which is a platform for developing and hosting web applications in data centers who are managed by Google.

The pricing structure (pay-per-use pricing) for Google AppEngine is shown in the table below.

Another example of SaaS is Sales Cloud which provides some features as sales representatives with a complete customer profile and account history. It enables users for decision makers, to manage marketing campaign and other information for the company's sales process.

TABLE VII. GOOGLE APPENGINE PRICING

Resource	Unit	Unit cost
Frontend Instances	Instance hours	$0.05/$0.10/$0.20/$0.30
Outgoing Network Traffic	Gigabytes	$0.12
Datastore Storage	Gigabytes per month	$0.18
Dedicated Memcache	Gigabytes per hour	$0.06
Blobstore, Logs, and Task Queue Stored Data	Gigabytes per month	$0.026
Logs API	Gigabytes	$0.12

Source: Google Cloud Platform / App Engine Pricing

The pricing structure for Sales Cloud is shown in the table below.

TABLE VIII. SALES CLOUD PRICING

Product	Description	Price (per user per month)
Contact Manager	Contact management for up to 5 users	$5
Group	Basic sales & marketing for up to 5 users	$25
Professional	Complete CRM for any size team	$65
Enterprise	Customize CRM for entire business	$125
Unlimited	Unlimited CRM power and support	$250

Source: salesforce.com

An another example of PaaS and SaaS is Microsoft Windows Azure as a Cloud Computing platform and infrastructure for building, deploying and managing applications and services in datacenters. The pricing structure for Windows Azure block blobs is shown in the table below.

TABLE IX. WINDOWS AZURE PRICING

Storage Capacity	Locally Redundant Storage	Zone Redundant Storage	Geographically Redundant Storage
First 1 TB / Month	$0.024 per GB	$0.03 per GB	$0.048 per GB
Next 49 TB (1 to 50 TB) / Month	$0.0236 per GB	$0.0295 per GB	$0.0472 per GB
Next 450 TB (50 to 500 TB) / Month	$0.0232 per GB	$0.029 per GB	$0.0464 per GB
Next 500 TB (500 to 1,000 TB) / Month	$0.0228 per GB	$0.0285 per GB	$0.0456 per GB
Next 4,000 TB (1,000 to 5,000 TB) / Month	$0.0224 per GB	$0.028 per GB	$0.0448 per GB

Source: azure.microsoft.com

V. CONCLUSIONS

In this paper we have reviewed and discussed some basic concepts for the pricing schemes and models in Cloud Computing.

Also we made some comparisons between recent pricing schemes and models which are implemented by providers. Each of the pricing schemes have advantages and their disadvantages, which often can be unfavorable to customers.

Future work must address the changes in risk sharing model between services provider and customer.

In the future a major consideration should be towards the development of an efficient and adequate pricing mechanism that will meet even more customer's requirements.

REFERENCES

[1] I. Foster, I. Yong, Z. Raicu and S. Lu, Cloud Computing and Grid Computing 360-Degree Compared, Grid Computing Environments Workshop, 2008.

[2] M. Al-Roomi, Sh. Al-Ebrahim, S. Buqrais and I. Ahmad, Cloud Computing Pricing Models: A Survey, International Journal of Grid and Distributed Computing Vol.6, No.5, pp.93-106, 2013.

[3] B. Sharma, R. K. Thulasiram, P. Thulasiraman, S. K. Garg and R. Buyya, Pricing Cloud Compute Commodities: A Novel Financial Economic Model, Proc. of IEEE/ACM Int. Symp. on Cluster, Cloud and Grid Computing, 2012.

[4] C. D. Patel and A. J. Shah, Cost model for planning, development and operation of a data center, hp technical report- hpl-2005-107(r.1), 2005.

[5] Pal, R. and Hui, P., Economic models for cloud service markets: Pricing and Capacity planning. Theoretical Computer Science 496, 113-124, July. 2013.

[6] W. Wang, P. Zhang, T. Lan and V. Aggarwal, Datacenter Net Profit Optimization with Individual Job Deadlines, Proc. Conference on Inform. Sciences and Systems 2012.

[7] C. S. Yeoa, S. Venugopalb, X. Chua and R. Buyyaa, Autonomic Metered Pricing for a Utility Computing Service, Future Generation Computer Syst., vol. 26, no. 8, 2010.

[8] M. Macias and J. Guitart, A Genetic Model for Pricing in Cloud Computing Markets, Proc. 26th Symp. of Applied Computing, 2011.

[9] P. Samimi and A. Patel, Review of Pricing Models for Grid & Cloud Computing, IEEE Symposium & Informatics, 2011.

[10] H. Wang, Q. Jing, R. Chen, B. He, Zh. Qian and L. Zhou, Distributed Systems Meet Economics: Pricing in the Cloud, Inproceedings, HotCloud '10, June 2010.

[11] I. R-Agundez, Y. K. Penya and P. G. Bringas, A Flexible Accounting Model for Cloud Computing, Annual SRII Global Conference, 2011.

[12] J. Huang, Pricing Strategy for Cloud Computing Services, PACIS Proceedings, paper 279, 2013.

[13] J. Zhao, H. Li, C. Wu, Z. Li, Z. Zhang, F. C.M. Lau, Dynamic Pricing and Profit Maximization for the Cloud with Geo-distributed Data Centers, INFOCOM, Proceedings IEEE, 2014.

[14] J. Jäätmaa, Financial aspects of cloud computing business models, Aalto University, master's thesis, 2010.

[15] S. Chun and B.S. Choi, Service models and pricing schemes for cloud computing, Springer Science + Business Media New York 2013.

[16] S. Kansal, G. Singh, H. Kumar and S. Kaushal, Pricing Models in Cloud Computing, Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies, ACM, 2014.

[17] Cloud Storage Providers: Comparison of Features And Prices, http://www.tomshardware.com/reviews/cloud-storage-provider-comparison,3905-3.html.

[18] H. Wang, Q. Jing, R. Chen, B. He, Zh. Qian and L. Zhou, Distributed Systems Meet Economics: Pricing in the Cloud, HotCloud '10, June 2010.

[19] H. Li, J. Liu and G. Tang, A Pricing Algorithm for Cloud Computing Resources, Proc. Int. Conference on Network Computing and Inform. Security, 2011.

[20] Amazon Web Services, http://aws.amazon.com/, last accessed 10.12.2015.

[21] Google App Engine, https://cloud.google.com/appengine/, last accessed 11.12.2015.

[22] M. Mihailescu and Y.M. Teo, Dynamic Resource Pricing on Federated Clouds, 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 2010.

[23] W.-Y. Lin, G.-Y. Lin, and H.-Y. Wei, Dynamic Auction Mechanism for Cloud Resource Allocation, Cluster, Cloud and Grid Computing, 10th IEEE/ACM, 2010.

[24] S. Shang, J. Jiang, Y. Wu, Z. Huang, G. Yang, and W. Zheng, DABGPM: A Double Auction Bayesian Game-Based Pricing Model in Cloud Market, Network and Parallel Computing, 2010.

[25] I. Fujiwara, K. Aida and I. Ono, Applying Double-sided Combinational Auctions to Resource Allocation in Cloud Computing, 10th Annual International Symposium on Applications and the Internet, 2010.

[26] S. Lehmann and P. Buxmann, Pricing Strategies of Software Vendors, Business and Information Systems Engineering, 2009.

[27] J. Rohitratana and J. Altmann, Agent-Based Simulations of the Software Market under Different Pricing Schemes for Software-as-a-Service and Perpetual Software, Economics of Grids, Clouds, Systems, and Services, ser. Lecture Notes in Computer Science, Springer, 2010.

[28] S. A.Bello, C. L¨uthje and C. Reich, Cloud Resource Price System, The Sixth International Conference on Emerging Network Intelligence, Emerging 2014.

[29] P. Hofmann and D. Woods, Cloud computing: The limits of public clouds for business applications, IEEE Internet Computing, vol. 14, no. 6, 2010.

[30] M. Jaekel and A. Luhn, Cloud Computing – Business Models, Value Creation Dynamics and Advantages for Customers. White Paper. Siemens IT Solutions and Services, 2009

Analysis of a Braking System on the Basis of Structured Analysis Methods

Ben Salem J.
Research Unit Signals and Mechatronic Systems, SMS, UR13ES49
National Engineering School of Carthage, ENICarthage
University of Carthage

Lakhoua M.N.
Research Unit Signals and Mechatronic Systems, SMS, UR13ES49
National Engineering School of Carthage, ENICarthage
University of Carthage

El Amraoui L.
Research Unit Signals and Mechatronic Systems, SMS, UR13ES49
National Engineering School of Carthage, ENICarthage
University of Carthage

***Abstract*—In this paper, we present the general context of the research in the domain of analysis and modeling of mechatronic systems. In fact, we present à bibliographic review on some works of research about the systemic analysis of mechatronic systems. To better understand its characteristics, we start with an introduction about mechatronic systems and various fields related to these systems, after we present a few analysis and design methods applied to mechatronic systems. Finally, we apply the two methods SADT (Structured Analysis Design Technique) and SA-RT (Structured Analysis Real Time) to the Anti-lock Braking System (ABS).**

***Keywords*—mechatronic system; ABS braking system; analysis and modeling; SADT method; SA-RT method**

I. Introduction

The systemic approach is used to define a system of analyzing it as a whole and in its sub-systems of studying and measure their goals or teleology to analyze the elements of the system that supports the achievement its objectives, analyze the links, interactions, mechanisms, and factors of balance and imbalance that regulate the system's objectives with its environment or with other systems with which it interacts [1].

The systemic analysis is defined as a discipline that studied and analyzes the concept of the systems with the aim of solving complex problems, to treat jointly effects and causes. So it's a discipline that brings together theoretical, practical and methodological approaches, for studying complex systems [2].

There are several methods of analysis and design systemic allowing the modeling of a system and its technical specification.

The purpose of this paper is to present the interests of the systemic approach, based on the two methods SADT (Structured Analysis Design Technique) and SA-RT (Structured Analysis Real Time). A case study of an ABS braking system is presented and discussed.

II. Analysis and Design of Mechatronic Systems

The Mechatronics term was introduced by an engineer of the Japanese company Yaskawa Electric Corporation in 1969 for characterizing a system integrating a mechanical part and an electronic part. Recently, the definition of mechatronics has evolved with the increasing complexity of systems [3]. Indeed, several definitions exist in the literature:

The definition of the newspaper IEEE Transactions on mechatronics or that chosen in 2000 by the Technical Committee on Mechatronics Systems define the mechatronic as the integration of mechanical and electromechanical (mechanical components, machinery…) with electronic systems (microelectronics, power electronics sensors, actuators) and information technology (Systems theory, modeling, automation, software, artificial intelligence) [3].

Mechatronic systems have emerged to meet the new requirements in terms of performance, comfort, safety and energy (Fig. 1). The mechatronics brings undeniable benefits such as: reduced costs, customer satisfaction, by the proposed innovative solutions, the positive response to social requirements increasingly important: pollution, consumption, safety of passengers...

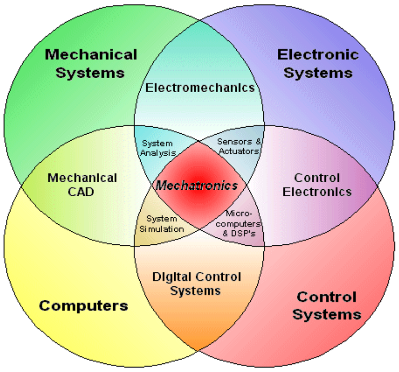


Fig. 1. Structure of a mechatronic system [4]

Nowadays, mechatronics is very present in the daily lives, as well as in industry. It touches many applications in various fields; it affects more and more the world of transport and particularly the automotive sector [5].

In this paragraph, a literature research work interested by the application of SADT and SA-RT methodologies are presented to analyze a mechatronic system

Researchers, Demri A. & al. (2008) [6], proposed using SADT, FMECA (Failure Modes, Effects and Criticality

Analysis), and the Petri nets methods to study the reliability of a mechatronic systems. By using functional analysis of defining the material limits, the various functions and operations performed by the system and the various configurations.

Researchers, Plateaux R. & al. (2009) [7], proposed to integrate all of the part of the cycle V design in order to attain a continuity of the modeling through the different levels of approach and design (functional and structural requirements). For this, they have proposed a hybrid method based on several tools, and Methods such as SADT, SysML (Systems Modeling Language) and Modelica, in the environment Dymola.

The researcher Vincent B. (2008) [8] applied the method Safe-SADT on a railway braking system on the one hand, and on the other hand he used this method to study the reliability of an automated system and to show the applicability of this method in industrial domains through a railway systems.

In order to study a production quality control system the researcher Khalaf Alahmad (2008) used the SADT method for constructing the Petri net model of a thermal central, and it evaluated the dynamic properties from the Petri network model. He studied the dynamic aspect of the system using the SADT temporal tool in the end; he produced a passage algorithm from SADT to Petri Network [9].

Works presented by Bruno Denis under the title "Design assistance and evaluation of the driving architecture of complex production systems" in (1994) [10] involve the design of the driving architecture of automated production systems". It’s about a prospective work whose objective is to construct a formal methodological framework of design and architectures evaluation for using SA-RT, SADT and temporized Petri nets methods. To validate and clarify the proposed method he studied architecture design driving of a launch station with a free transfer. After presenting some research works on the application of some methodologies, a state of the art on the two methods SADT and SA-RT are presented.

III. PRSENTATION OF THE METHODS SADT AND SA-RT

The SADT represents an image of the system. It is a method of analysis to understanding why a system exists, or must be designed, what functions it must fulfill and finally, how they are realized, and whatever of the complexity. The method is based on a graphical model, proceeds by down approach in the sense that are going from general to more detailed, by focusing on system activity (Fig. 2) [11], [12].

The SADT method seems adapted to the modeling of mechatronic systems for at least one reason: this method applies perfectly to the multi-technological systems, that is to say, it adapts to mechanical, electronic and software systems [11], [13]. But it does not take into account the dynamic aspect of system.

The SART method is a method of specifying computer systems falling within of real time domains. It takes into account the dynamic aspect of the analyzed system. It is based on the SA method (Structured Analysis) which has been used extensively.

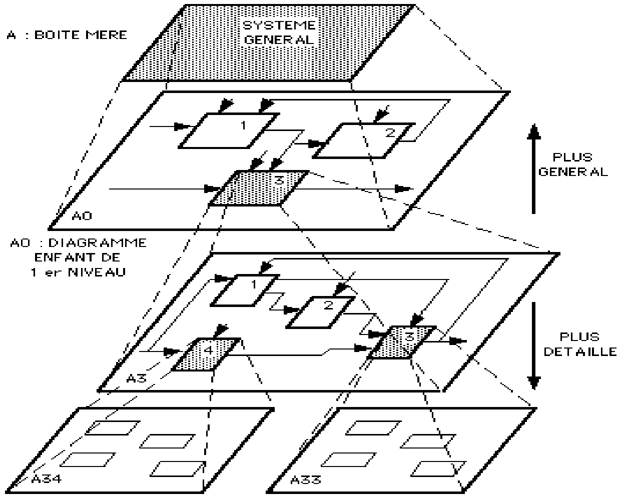


Fig. 2. Structure of an SADT model

SA-RT is well suited for applications with high dynamic behavior. This aspect is missing from the SADT method for modeling a mechatronic system [14].

In the SA-RT model, the definition of a system is given on the one hand of the description of system functions and control of the information flow and on the other hand of the description of the hardware architecture of the internal system (Fig. 3) [15], [16].

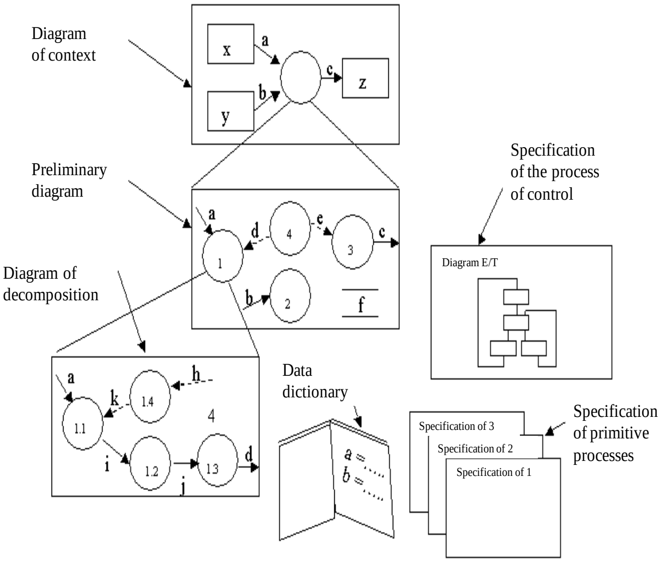


Fig. 3. Organization of the SA-RT model

IV. RESULTS OF THE ANALYSIS AND THE MODELING

The presence of the ABS braking system in the vehicle is important, its functioning is based on an electronic calculator which constantly analyzes the speed of the vehicle and its variation, as well as of the four wheels (integrated sensors). When it detects a blockage of one or more wheels, the system responds by instructing the brake system and reducing its action on the designated wheel. Figure 4 shows the components of a braking system.

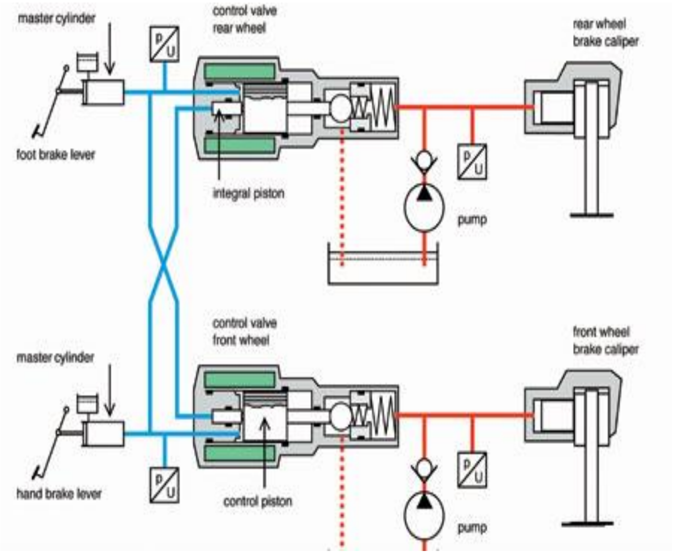


Fig. 4. Structure of an ABS braking system

A. Review on braking system control

Some studies on braking system control that has been presented in various researches:

Researchers Sidek S.N. & al. (2000) [17], have considered the use of an intelligent controller to achieve the objective to modify the current conventional braking system so as to make it work automatically. To ensure high speed of system response, a DSP controller TMS320C24x with an embedded fuzzy algorithm is used in the implementation of this new device. Results of simulation studies using MATLAB have demonstrated the feasibility of this new system under investigation

Researchers Fletcher I. & al. (2003) [18], have investigated one aspect of system design, the braking system. The design exercise is based upon a simulation of cars braking system enables several alternative control strategies to be assessed. The findings illustrate the problems involved and the opportunities available for the application of an 'intelligent' control strategy.

Researchers Li Junwei & al. (2009) [19], have explained why ABS is an important part to improve the automobile's active safety. In general, ABS is designed to achieve maximum negative acceleration by preventing the wheels from locking. Researches show that the friction between road and tire is a nonlinear function of wheel slip. In this paper, to deal with the strong nonlinearity in the design of ABS controller, a variable structure controller has been designed and index reaching law and integral switching surface with saturation function methods are used to reduce chattering. In the simulations, several situations such as braking in dry road, wet road and snow road are considered.

Researchers Chun-Liang Lin & al. (2011) [20], have presented how the slip ratio control problem in ABS is highly nonlinear and complicated. A sliding mode controller is developed to generate appropriate torque for the driving motor of two-wheel electric vehicles that ensures optimality of the slip ratio for efficient vehicle brake. The design is based on a novel short-circuit braking mechanism to emulate the mechanical ABS for the traditional gas-powered vehicles.

Researchers Lu Bo & al. (2010) [21], have presented a simplified vehicle vertical two-wheeled model, an improved fuzzy PID controller was devoted to the ABS. This controller was obtained based on inosculating with fuzzy PID controller and the algorithm of objective function Automatic optimization. A simulation result was gotten through simulation of MATLAB Simulink environment. This result shows that the control system had better stability, adaptive, control precision and shortened the braking distance, braking time, reduced the slip angle through using this improved fuzzy PID controller. Comparing with the conventional PID controller, the fuzzy controller and the fuzzy PID controller, the better control effect and the control system stability could be acquired by this improved fuzzy PID controller.

Researchers Cabasino, M.P. & al. (2011) [22], have considered the brake system of a vehicle whose wheels are equipped with ABS. They assume that the sensors that are responsible of the activation of the ABS are subject to faults. They first show how such a system can be modeled using labeled Petri nets and the notion of concurrent composition. Then, they show how fault diagnosis and diagnosability analysis can be performed on such a system using appropriate techniques based on Petri nets.

Researchers Qi Zhang & al. (2004) [23], have presented the precondition of realizing logic thresholds based ABS which is the accuracy calculation of wheel angular acceleration. By analyzing the properties of the output signal of the electromagnetism induction sensor, which is used to measure wheel speed in vehicle antilock braking system, this paper proposes a kind of circuit that can satisfy the requirements of the system. Utilizing the wheel speed signals and applying the Kalman filter technology, the wheel angular acceleration is calculated. The trial results verify the feasibility and validity of this method.

Researchers Yonghua Xiong & al. (2004) [23], have proposed an improved ABS ECU test system for the pneumatic ABS. Composed of finished ECU, data acquisition card, V/F convertor and PC, the system accomplishes the establishing of vehicle, brake and braking air pressure model for realistic simulation, and the dynamic loading method is used to load the different model on PC. Using the system, the test and evaluation of the signal processing ability, communication capacity and the effectiveness of the controlling algorithm of ECU can be done. Finally, a type of ECU produced by one company is taken as an example to verify the feasibility of the testing system, and the results show that the system is available for practical application for its low-cost and short-cycle.

Researchers He Jidu & al. (2004) [23], have presented a fuzzy immune adaptive PID control algorithm for ABS system after analyzed the lack of traditional slip rate ABS control algorithm, combining with the biological immune principle and the adaptive ability of fuzzy logic ratiocination. And the fuzzy immune PID controller adjust the proportional coefficient, the fuzzy controller adjust the integral and differential coefficient. Compared with the PID control algorithm, fuzzy control algorithm, fuzzy PID control algorithm and fuzzy immune adaptive PID control algorithm, the simulation results indicate that the method has characteristics of small overshoot, fast response, shorter braking distance and strong anti-interference ability and robustness. There is a higher application value in ABS control system.

B. SADT results

First, SADT analysis of the braking system ABS is presented. This analysis provides a synthetic description of the ABS system operating modes. The realization of this analysis is intended to identify the technical functions of the system. It allows us to make a hierarchical decomposition of the ABS system elements (Fig. 5).

The first diagram shown above is the level A-0 of the SADT method. The first diagram shows the main functions of the ABS system, which is to slow down the vehicle without blocking the wheels. The second diagram which is the level A0 shown in the following figure 6 is a decomposition of the previous level A-0, it contains 6 boxes.

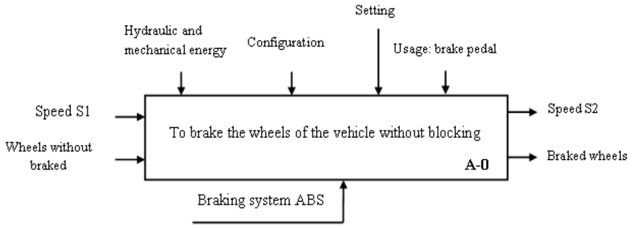


Fig. 5. Overall function of a braking system ABS

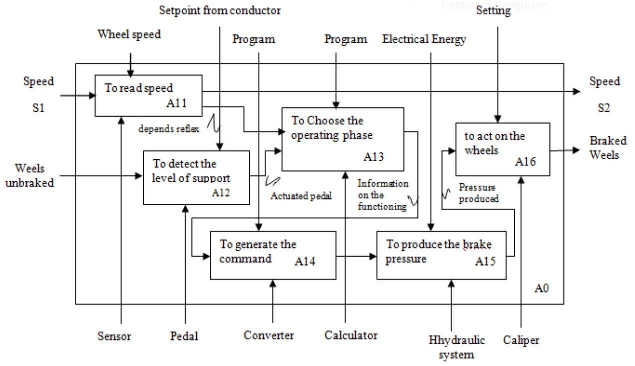


Fig. 6. Node A0 of the braking system

C. SA-RT results

Second, an SA-RT analysis of an automotive braking system composed on the one hand of a classic whole of a brake pedal (braking demand) and a brake (braking actuator) and on the other hand of an ABS system [24] is presented. A sensor sliding of wheel is associated to this ABS system. To simplify, the working of the ABS is based on a stop of braking when a sliding is detected on the wheels and it even though the driver's demand is always efficient. The driver has the possibility to activate this ABS system with the help of a specific button (button to two steady states: switch). A seer permits to indicate it (the activation of the ABS system). But then, it is not possible to deactivate the ABS system during the braking that is during the support on the brake pedal.

The whole of data or events exchanged with the outside of the functional process that represents the application, constitute specifications of entrances and exits of the application. The description of these Inputs/Outputs will be made in the dictionary of data.

The context diagram (Fig. 7) is constituted of the functional process "to control the braking system" and of five terminators:

- Brake pedal providing the data braking demand;

- Activation Button of the ABS providing the data ABS activation;

- Sliding sensor" providing the data Wheel sliding;

- Braking system consuming the data Braking command;

- ABS light consuming the data Display ABS.

This context diagram perfectly defines the interfacing between the inventor and the customer that is data to either provide or to generate [25].

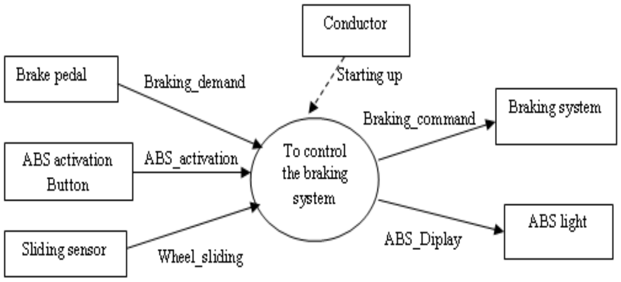


Fig. 7. Context Diagram of the braking system ABS

The preliminary diagram is constituted of five functional processes (Fig. 8). It can immediately underline at the level the obligatory consistency between the context diagram and the preliminary diagram at the level of the data flows in entrance and in exit. The passage of data between the functional processes is done of direct way. It is important to note that the data "Sliding_state" and the data "Button_ABS_state" are Boolean type.

A control process in the preliminary diagram is implemented in order to coordinate the different functional process execution (Fig. 9). This control process will therefore interact with a functional process either to launch or to activate its execution and, in return, the functional process will provide if necessary an event indicating the result of its treatment in order to give some useful information to change the control states.

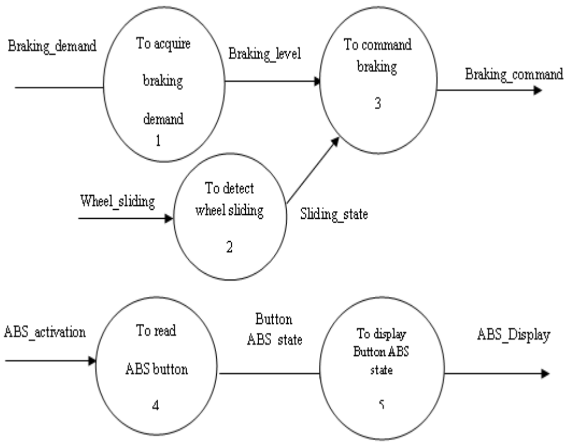


Fig. 8. Data Flow Diagram of the baking system ABS

In order to specify the control process of the application, the representation of the state-transition diagram is presented (Fig. 10).

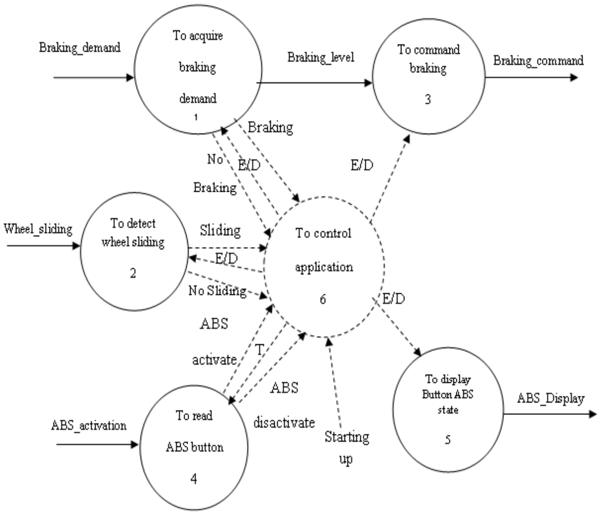


Fig. 9. Control Flow Diagram of the braking system ABS

The analysis of the system by the SADT method is descending, hierarchical, modular and structured, it brings out the maximum details as when in proportion as one progresses in the decomposition. The SADT representation system to describe the process in several levels to show the relationships, to explain the details gradually, controlled and summarized while avoiding the inaccuracies inherent in natural language. The SADT model allows clear view system functionality and interactions between these elements [37], [38].

After this analysis, we guarantee a functional description of the system regardless of the various possible solutions in the realization. Therefore it resulted in two levels of abstracti on, the conceptual level and an organizational level [9].

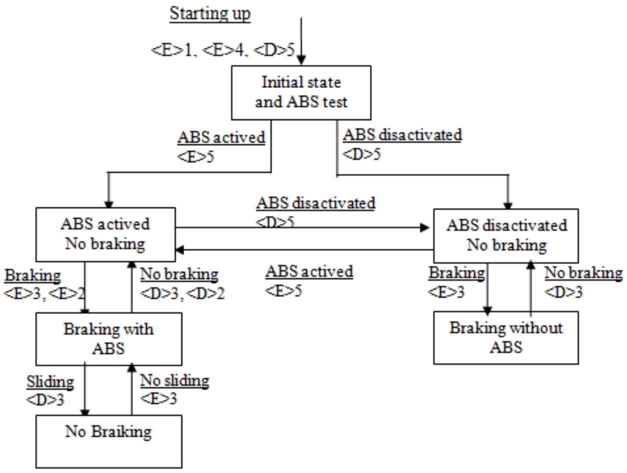


Fig. 10. State - Transition diagram of the ABS braking system

Watching the SADT model of the system it is noted that the depth of the decomposition depends on the informational availability between the levels. Furthermore, it is limited by the lower level of readability of the diagram. It contains a limited amount of information on a specific topic. Moreover it is difficult to analyze clearly the activities of a system with dynamic behavior [39], [40].

For the analysis of the system by the SA-RT method, it represents communication between designers and users of an application [30], [31], [32]. The SART models make it possible to express the wishes and needs of the user by removing the ambiguities of everyday language. SART is a move towards a method of dynamic specification. SART can thus be characterized by three aspects [33], [34], [35], [36]:

- A functional aspect, which is achieved through the use of data flow diagrams, help to show how functional processes transform their inputs (billows of incoming data) in outputs (billows of outgoing data). They also allow distinguishing between the types of discrete or continuous data, and they represent that the data storage can be used by multiple processes.
- An event aspect for the dynamic of the model is described by conditions or event occurrences. Two elements are introduced for this purpose: the control process, which pilots the functional processes by generation of events activation or deactivation [27] [28]; the control billows, representing the occurrence of events generated by different processes (functional or control). Here again, we can distinguish between consumables events (discrete) and permanent (continuous). The functioning of a control process is described via a state-transition diagram which models the life cycle processes. This diagram thus allows materialize the incidence of an event (or a combination of events) on the state of the system by indicating the actions to be taken when he arrived.

V. Conclusion

In this paper, the need for system modeling of an ABS braking system for understanding its operation and exploitation with consistency and in a correct manner was presented. In fact, the methods used for the description of this system are characterized by their graph formalism which is necessary for the description of the system operating modes studied and identified sub systems and knowledge of its internal and external functions. Indeed, two structured analysis methods SADT and SA-RT are presented, on the one hand and an application of these methods on a practical case of an ABS braking system was presented, on the other hand. This application shows the interest of the graphical formalism of these two methods to the analysis and design of a mechatronic system in particular to describe the environmental and behavioral modeling.

Staring from this case study of the application of the structured analysis of an ABS braking system discussed in this paper, work is in progress to develop a general methodology of analysis and modeling for different mechatronic systems.

References

[1] G. Turchany, Act together on Education for Sustainable Development, Bordeaux 27-290 October 2008.

[2] M.N. Lakhoua, Systemic analysis of an industrial system: case study of a grain silo, Arabian Journal for Science and Engineering, vol.38, 2013, pp. 1243-1254.

[3] El Feki M., Analysis and tolerance synthesis for the design and dimensioning of mechatronic systems, Central School of Lyon, 2011.

[4] Web site: http://www.engr.ncsu.edu/mechatronics/what-mech.php [consulted January 2016]

[5] M. Zerelli, Mechatronic systems with variable parameters: behavior analysis and tolerancing approach, LISMMA SUPMECA Toulon, École Centrale Paris, 2014.

[6] A.Demri, F. Charki, A. Guerin and H. Christofol, Functional and dysfunctional analysis of a mechatronic system, Annual Reliability and Maintainability Symposium RAMS, 2008, pp. 114-119.

[7] R. Plateaux, J.Y. Choley, O. Penas and A. Riviere, Towards an integrated mechatronic design process, IEEE International Conference on Mechatronics, ICM, 2009, pp. 1-6.

[8] B. Vincent, Application of the safe-sadt method on a railway braking system INRETS ESTAS, Villeneuve d'Ascq, France.

[9] K. Alahmad, Control system of production quality control methodology of modeling and optmization of production systems. Univ. Paul Verlaine in Metz, 2008.

[10] A. Bruno, Assistance in the design and evaluation of the architecture of conduct of complex production systems University Research Laboratory Automated Production (LURPA - EA 1385) Cachan, 1994.

[11] P. Jaulent, SADT a langage for communication, IGL Technology, Eyrolles, Paris, 1989.

[12] M. Galinier, SADT a language to communicate, Eyrolles, Paris 1989.

[13] Augusto, Modeling of complex systems. Higher National School of Mines of Saint-Etienne, 2013.

[14] Demri, Contribution to the evaluation of the reliability of a mechatronic system functional and dysfunctional modeling, Univ. Angers, 2009.

[15] D.J. Hatley, I.A. Pirbhai, Specification, Strategies for Real-Time Systems (SA-RT), Masson, Paris, France, 1991.

[16] M. N. Lakhoua, Using structured analysis for the control of real-time systems, Journal of Engineering and Technology Research Vol. 4(5), pp. 82-88, October 2012.

[17] S.N. Sidek and. M.J.E Salami, Design of intelligent braking system, Vol.2, 2000, pp. 580- 585.

[18] Fletcher I., Automatic braking system control, IEEE International Symposium on Intelligent Control, 2003, pp. 411- 414.

[19] L. Junwei and J. Wang, Design of antilock braking system based on variable structure control, IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009.

[20] L. Chun-Liang and Y. Meng-Yao , Design of anti-lock braking system for electric vehicles via short-circuit braking, Second International Conference on Mechanic Automation and Control Engineering (MACE), 2011.

[21] Lu, Bo; Yu Wang; Jing-jing Wu; Jin-ping Li, ABS system design based on improved fuzzy PID control, Sixth International Conference on Natural Computation (ICNC), 2010.

[22] M.P. Cabasino, A. Giua, C. Seatzu, A. Solinas and K.. Zedda, Fault diagnosis of an ABS system using Petri nets, IEEE Conference on Automation Science and Engineering (CASE), 2011.

[23] Qi Zhang; Guofu Liu; Yueke Wang; Tingting Zhou, Study of calculation method of wheel angular acceleration in ABS system, International Conference on Information Acquisition, 2004.

[24] Yonghua Xiong; Xiaoyan Li; Yong He; Min Wu; Yonghong Long, Design and application of testing system for pneumatic ABS ECU, 11th World Congress on Intelligent Control and Automation (WCICA), 2014.

[25] He Jidu; Zheng Yongjun; Tan Yu; Wu Gang, Research on Vehicle Anti-braking System Control Algorithm Based on Fuzzy Immune Adaptive PID Control, Third International Conference on Digital Manufacturing and Automation (ICDMA), 2012.

[26] F. Cottet, Real time systems of control command, Dunod., Paris. 2005.

[27] W. S. Liu, Real-time Systems, Prentice Hall, 2000.

[28] M.N. Lakhoua, Review of Structured Analysis and System Specification methods, Journal of Electrical Engineering, Vol.10, N°2, 2010.

[29] D. Laurent, Contribution to the management of the regularity of execution of real-time tasks from one application to strict constraints, in an online scheduling context. Laboratory of Computer Science and Industrial Scienti_que (LISI), University of Poitiers, 2002.

[30] L.E. Colmenares-Guillen, O. Niño Prieto and A. Aguila Jurado, An Approach of Real-Time System for River Monitoring and Flood-Warning System in Puebla, Mexico, World Applied Programming, Vol.3, Issue (8), August 2013. 328-340.

[31] M.I. Capel-Tuñón, Modelling and Simulation of a Real-Time Hybrid System, Proceedings of EOMAS'08.

[32] C. de la Riva and J. Tuya, Automatic generation of assumptions for modular verification of software specifications, Journal of Systems and Software, Vol.79, Issue 9, Sept 2006, pp. 1324-1340

[33] L. Jóźwiak and S-A Ong, Quality-driven model-based architecture synthesis for real-time embedded SoCs, Journal of Systems Architecture, Vol.54, Issues 3–4, March–April 2008, pp. 349–368.

[34] J.M. Fernandes, J. Lilius and Dragos Truscan, Integration of DFDs into a UML-based model-driven engineering approach, Softw Syst Model 2006, 5 pp. 403–428.

[35] M. Al-Mohamed and D. Esteve, Tools and models for systems design and synthesis of MEMS based on asynchronous circuits, IEEE International Conference on Industrial Technology vol.1, 2000, pp. 64 - 69.

[36] Specifying systems and applications with SA/SD/RT method, Training course: Modeling/SA, Insoft November 2013, web: http://www.insoft.fi

[37] D.A. Marca, SADT/IDEF0 for Augmenting UML, Agile and Usability Engineering Methods, Software and Data Technologies, 2012, pp. 38-55.

[38] N. Ouasli, R. Ben Mehrez and L. El Amraoui, Parameter estimation of one wheel vehicle using nonlinear observer, International Conference on Electrical Sciences and Technologies in Maghreb, CISTEM 2014.

[39] M.N. Lakhoua, Contributions à l'analyse systémique, à la supervision et à la sûreté de fonctionnement des systèmes de contrôle-commande, Rapport de synthèse, HDR, ENICarthage, Tunisie, 2015.

[40] M.N. Lakhoua, Application of Functional Analysis on a SCADA system of a Thermal Power Plant, Advances in Electrical and Computer Engineering journal, 9(2), 2009.

Contributions to the Analysis and the Supervision of a Thermal Power Plant

Lakhoua M.N
The National Engineering School of Carthage,
University of Carthage, Research Unit:
Signals & Mechatronic Systems, Tunisia

Ben Hamouda M.
The National Engineering School of Carthage,
University of Carthage, Research Unit:
Signals & Mechatronic Systems, Tunisia

Glaa R.
The National Engineering School of Carthage,
University of Carthage, Research Unit:
Signals & Mechatronic Systems, Tunisia

El Amraoui L.
The National Engineering School of Carthage,
University of Carthage, Research Unit:
Signals & Mechatronic Systems, Tunisia

***Abstract*—Supervision systems play an important role in industry mainly due to the increasing demand for product quality and high efficiency, and to the growing integration of automatic control systems in technical processes. In fact, the supervision system has a great number of components and interconnections, and it is difficult to describe and understand its behavior. Furthermore, the supervision system in industrial plants, implemented in supervisory control and data acquisition (SCADA) software, must undertake, at least, the following three main tasks: monitoring, control and fault tolerance. So it can be classified as a complex system. The objective of this paper is to show interests of the use of functional analysis techniques such as SADT (Structured Analysis and Design Technique) and SA-RT (Structured Analysis Real Time) for the design of supervisory systems. This is why we present a general model of analysis and supervision of production systems. This model was based on the one hand on the functional analysis (FA) and on the other hand on the SCADA system.**

Keywords—SCADA systems; SADT method; SA-RT method; thermal power plant

I. INTRODUCTION

In our days in different industrial areas it become very important to analysis, control and supervise the production systems. In fact, the most used are supervisory control and data acquisition (SCADA) systems; these are used to control systems in different areas, to collect information and to centralized data acquisition. One of the most important concerns in the SCADA design is to assure whether the system could go with the reliability and performance requirements specification.

The production systems of increasing complexity present difficulties at the level of their functional analyses. Several types of functional and structural analysis methods exist and are used in many industrial domains since about thirty years. In fact, the functional analysis (FA) permits the synthetic description of fashions of working of a system and the knowledge of functions to guarantee [1]. It establishes of systematic and exhaustive way the functional relations inside and outside of the system. Again, the FA consists in searching for and to characterize functions offered by a system to satisfy its user's needs.

The objective of FA methods is to describe and to understand the system. It is gaits that put the accent on what makes the system, to master the complexity and to detect defaulting. So the FA serves to model the system while clearing constraints and while specifying the different fashions of working [1].

Many well-known graphical modeling techniques, for example, the Structured Analysis Real Time (SA-RT) method developed by Hatley and Pirbhai [2], provide mechanisms that allow one to: (1) abstract over details best left to later stages of development, (2) model an application along different views, and (3) modularize problems and solutions. It has also been affirmed that the simple and graphical manner of the modeling constructs facilitates the construction of concise and understandable models [3].

This paper can be loosely divided into five parts: first, we summarize the complexity of supervision systems particularly to design SCADA systems and second briefly describe the characteristics of a SCADA system and the problems related to its design. Subsequently, we present the issues involved in the analysis and supervision of the production systems. In section 3, after presenting concepts of SADT and SA-RT, a model of analysis and supervision of production systems is presented. Then, the benefits of using SADT and SA-RT in the design steps are developed. Next, the results of the proposal methodology was used on a case study of a SCADA system of a thermal power plant (TPP) particularly in programming and visualizing new chemical analysis parameters of the water - steam cycle. In the last section, we present a discussion about the advantages and inconveniences of the methodology used in order to describe and understand the behavior of supervision systems.

II. DESIGN OF SCADA SYSTEMS

With the advances of electronic and software technologies, the SCADA systems are generally used in industrial plant automation. It provides an efficient tool to monitor and control equipment in industrial processes. In fact, SCADA systems can be found in critical infrastructures such as power plants [4] [5] and power grid systems, water, oil and gas distribution systems, building monitoring, production systems and other products.

In a typical SCADA system, data acquisition and control are performed by remote terminal units (RTUs) and field devices that include functions for communications and signaling. In fact, the control performed by SCADA is much more superficial and generally applied to correct a fault and the SCADA system monitors and provide control actions based on this monitoring [6].

The whole automation process is done using programmable logic controller (PLC) which has number of unique advantages like speed, reliability, less maintenance cost and reprogram ability [7]. Without reasonably designed system software architectures and hardware structures, it is impossible to handle these tasks efficiently, safely and reliably, with the possibility of online reconfiguration and flexibly embedding applications.

Designing, monitoring and controlling such systems is becoming more and more challenging as a consequence of the steady growth of their size, complexity, level of uncertainty, unpredictable behavior, and interactions.

Moreover, SCADA systems have been applied in flexible manufacturing cells for educational purposes in different automation engineering fields [8]. Design and implementation of a low cost compact modular production system controlled by a SCADA system was developed for educational purposes [9]. In this way, students have access to instruments in a lab via the Internet, all by improving their skills related to SCADA systems, used in industry [10].

We present in this part some applications of SCADA systems design that have been presented in various researches.

Researchers, Ponsa P. & al. [11], have described how Human-Machine-Interfaces have received appropriate attention in order to improve the design of SCADA applications in industrial domain. It is because of a major concern about aspects related to maintenance, safety, achieve operator awareness, etc has been gained. Even there are in the market software solutions that allow for the design of efficient and complex interaction systems, it is not widespread the use of a rational design of the overall interface system. The researchers have provided an example of such development also by showing how to include the automation level operational modes into the ergonomic interfacing system.

Researchers, Gezer D. & al. [12], have presented a methodology and a case study through which system architecture and dynamic models of related system components are identified in order to design and simulate the SCADA system of a new hydro turbine test laboratory. System architecture model is prepared in System Modeling Language, a system modeling language based on Unified Modeling Language, while the dynamic model of the laboratory is formed in Matlab/Simulink. Some simulations are performed in order to verify the preliminary system design studies and system requirements.

Researchers, Dong W. & al. [13], have presented a remote monitoring system that has been designed by using SCADA. Specifically, the monitor signals of boiler dry spot's worker are been composed first, and a real time database is realized on the basis of the collected information. In summary, these approaches could extend the system function and optimize the system structure. More importantly, it obtains the real-time data of web monitor system, which can draw the real-time curve of the dynamic systems and the curve of history, etc. Furthermore, it could achieve real-time brushless description.

Researchers, Morosan A. & al. [14], have proposed an architecture for the SCADA system, used for making well-organized production in manufacturing system, in order to can combine all the elements of a flexible manufacturing line. The main characteristic of this SCADA system is to supervise and to control the manufacturing process from a flexible manufacturing line.

Researchers, Chun-Lien S. & al. [15], have proposed an approach to take the system performance into account in the reliability analysis of the SCADA system. This approach is based on a model for evaluating data transmission time, which allows us to find the operation time needed to complete SCADA functions. With information on performance, the reliability evaluation technique using fault tree analysis is described and applied for analysis of SCADA system component connectivity to assess the availability of SCADA controls. A sensitivity analysis is also described and used to illustrate the effects of input data uncertainty on the system reliability.

Researchers, Guobing H. & al. [16], have studied an embedded SCADA system in order to manage and control industry objects directly at the production line terminal. A solution is recommended for an embedded SCADA system based on the development of embedded computer and network communication. The design and implementation methods are introduced for technical structure, hardware and software of this system. The first 10 units have been brought into use at power plants and power monitor centers since 2010. The result of site acceptance indicates that measuring precision and reliability follow correlative standards.

In the following part, we present the two structured analysis methods.

III. Presentation of Structured Analysis Methods

This section presents two structured analysis methods used in this research. These methods are: Structured Analysis Design Technique (SADT) and Structured Analysis Real Time (SA-RT). SADT, which was designed by Ross in the 1970s [17], was originally intended for software engineering but quickly other areas of application were found, such as training, operations, manufacturing, finance, etc. Although SADT (Figure 1) does not require any specific supporting tools, a number of computer programs implementing SADT methodology have been developed. Among of them is Design: IDEF, which implements IDEF0 method [18]. Furthermore, SADT/IDEF0 is a confirmed way to model any kind of domain [19].

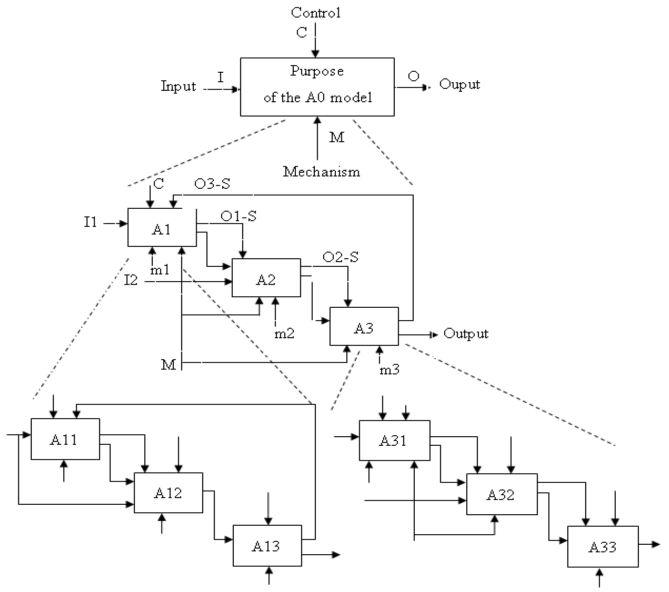


Fig. 1. Top-down, modular and hierarchical decomposition of SADT

Researcher, DeMarco, one of the pioneers of structured analysis (SA), described the need for such a method more than 30 years ago. He suggested that the products of analysis must be maintainable; problems of size must be dealt with using the effective method of partitioning; graphics must be used whenever possible [20]. Then, he proceeded further to establish requirements for the SA method as follows: the method should help us partition our requirements and document that partitioning before specification; it should give us means of keeping track of and evaluating interfaces; it should facilitate the development of new tools to describe logic and policy better than narrative text.

The notation used for creating SA diagrams consists of:

Data Flow Diagrams (DFDs): these diagrams model the processing of information in terms of data flows; data processing nodes are defined to represent the data processing functions from the requirements of the system.

Control Flow Diagrams (CFDs): these diagrams model the processing of information in terms of control flows; control nodes are defined to represent the control functions from the requirements of the system under analysis.

Process specifications: these specifications are used to describe the details of the data processing nodes defined in a DFD. These specifications consist of scripts of pseudo-code or just plain text which explains how the output flows of a particular processing node is generated from its input flows.

Control specifications: these specifications are used to describe the details of controls nodes in a CFD. These specifications define the behavioral model of the controllers and specify how the output control flows are obtained from the input control flows. They also specify when the data processing nodes are activated or deactivated.

Data Dictionary: it defines all the information flows and the data and control stores in the system. It contains text that defines each information item and its value range.

Entity-Relationship Diagrams (ERDs): these diagrams provide an information model for the data items and control signals and the relationships among these data items.

Figure 2 shows three dimensional views of the SA model. The top layer, which is the highest level in the SA-RT model, consists of what is commonly termed as the context diagram (CD). It describes the boundary of the software under analysis as well as the external interfaces and the external entities.

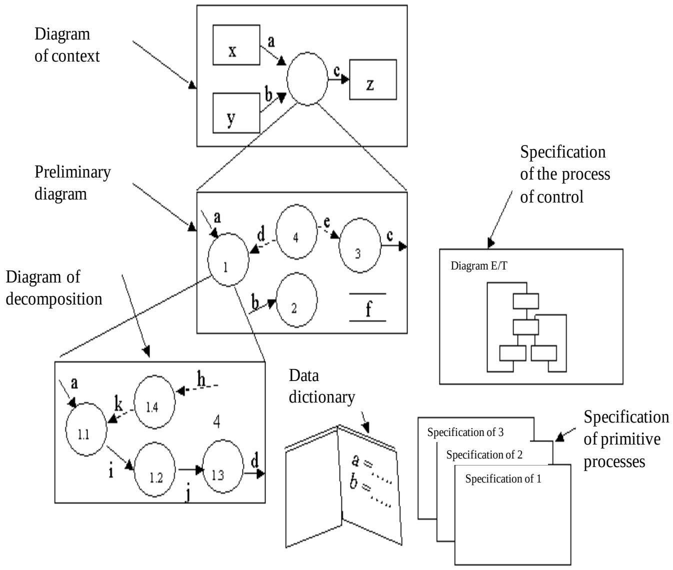


Fig. 2. Components of SA-RT and their relationship with each other

The following level of the hierarchy is a DFD which represents the major functions outlined in the functional requirements. These functions represent a top level decomposition of the software under developed.

Consequently the whole DFD0 is viewed as the child of the process representing the system in the top level. Each process in DFD (numbered 1, 2, 3, etc.) will have either its own child DFD or a P-spec sheet describing it in more detail. Also C-spec sheets are used to specify the controllers shown in each DFD.

We present in this part some applications of the SADT and SA-RT methods that have been presented in various researches:

Researchers, Benard V. & al. [21], have described the Safe-SADT method that enables the explicit formalization of functional interactions, the recognition of the characteristic values affecting the dependability of complex systems, the quantification of the reliability, availability, maintainability, and safety parameters of the system's operational architecture, and its validation in terms of the dependability objectives, as well as constraints set down in the functional requirement specifications.

Researchers, Lauras, M. & al. [22], have presented an approach based on GRAI, SADT/IDEFO that enables the integration of the best practices defined by these methods. In addition the indicators next the results and determinants, three types of indicators are introduced to analyse the performance: the facility viewpoint measures, the appropriateness of the resources available and the determinants of the activity.

Researcher, Marca D.A. [23], has explained how SADT/IDEF0 domain modeling can bring precise and complete context, to today's commonplace disciplines of the UML (Unified Modeling Language), Agile System Development, and Usability Engineering methods. In fact, the power and the rigor of SADT/IDEF0 come from: (1) a synthesis of graphics, natural language, hierarchical decomposition, and relative context coding, (2) distinguishing controls from transformations, (3) function activation rules, and (4) heuristics for managing model complexity.

Researchers, Jimenez F. & al. [24], have developed models and tools for system design and synthesis of MEMS-micro based on SDL (specification description language), SA-RT and PNs. In fact, a main problem concerns the design of these varied circuits because it associates disciplines such as electronics, mechanics, chemistry, etc.

IV. General Model of Analysis and Supervision

We present in this part, the development of a general model for the FA and the supervision of a production system in a SCADA environment.

The objective of the work is therefore the establishment of a general methodological gait of functional modeling and supervision of the production systems permitting to reach the aimed objectives [35].

The proposed model articulates around three essential phases: functional analysis of production systems; analysis of control-command applications; supervision of the control-command applications in a SCADA environment.

Phase 1: Functional analysis of the production systems

The first phase consists in proceeding to a functional modeling of the production systems while using the SADT method. The structure of this method permits to master the complexity of the process thanks to its downward and modular analysis.

The different stages of such a FA by the SADT method of the production systems are:

- To prepare the SADT model: the creation of a SADT model starts with the definition of two concepts that is the goal that stationary objectives of the model and the point of view that establish for what auditorium model it is created. These two determined concepts permit to guide the decomposition of every box;
- To create the diagram A-0 representing the general activity to analyze in an actigram;
- To create the diagram of activity A0: one writes down inside boxes drawn on a new form (3 to 6 boxes) the name of activities and one draws arrows of interfacing between the different boxes;
- From the diagram A0, to select the least complex box and that will give more information to create the diagram by decomposition have (1 < i <6) with corresponding i to the number of the box in A0;
- To redo the same principle of decomposition for the other diagrams of activity;
- To represent all diagrams of the model on the SADT forms.

This first phase of the model enables us the decomposition of the production system under a hierarchized manner to bring back it to the elementary situations and to elaborate a static model describing the activities of the process.

Phase 2: Analysis of the control-command applications

Once functions and activities of every function have been identified, the following stage consists in analyzing control-command applications while using the SA-RT method.

The different stages of such a FA by the SA-RT method of control-command applications are:

- To establish the CD while representing the different terminations as well as the incoming and retiring data streams;
- To establish the preliminary diagram (DFD0) that represents the necessary functional process list to the application with the stream of data correspondents.
- To establish the diagram of decomposition partner streams to one of the functional processes identified;
- To establish the DFC representing the controls aspect the SA-RT method;
- To establish the State/Transition diagram representing the behavior aspect or real time of the application;
- Functional process specification (dictionary of data) while leaning on a procedural specification.

This second phase of the model enables us the decomposition of the studied application under a hierarchized manner to bring back it to the elementary situations and to drive to the development of a dynamic model describing the various processes of the application as well as flows of data and control.

Phase 3: Supervision of the control-command application in a SCADA environnement

After a various process analysis and flows of data and control of the application, different static and dynamic model are elaborated. The last stage of the model of the analysis and the supervision of the production system consists in the establishment of tools and methods of supervision, control and date acquisition using a SCADA system as well as the development of Man - Machine interfacings. In fact, the supervision is the whole of tools and methods that permit to drive some industrial facilities so much in normal working that in presence of failings. It is the tool of reference of the conduct operator but can interact also directly with the control-command application.

The supervision has for objective to supervise and to control the working of an installation so that it remains in the normal working whatever are the outside disruptions. It permits to detect in real time mistakes from alarms and to identify their reasons.

In the following part, we present the results of FA and supervision on an example of a SCADA system of a thermal power plant (TPP) as well as the development of different phases of the general model proposed.

V. Results of Functional Analysis and Supervision

In this section, we present on the one hand an example of a SCADA system in a TPP in Tunisia and on the other hand a control-command application of the water steam cycle of the TPP. By means of a significant example, the objective of this work is to validate the general model of analysis and supervision presented in the previous section [31], [32].

A. Presentation of the SCADA system and a case study of a water steam cycle

In Tunisia, the TPP of Rades (near to Tunis) is composed of thermal power stations that are one of the most important stations of the electric energy production in Tunisia (37% of the national production) [33], [34]. In fact, the SCADA system is used by the power station operators in order to supervise the good working of two production slices of the TPP. This centralized supervision permits to operators, since the control room to control facilities in their domain of exploitation and to treat, in real time, the different types of incidents (Figure 3).

Fig. 3. Control room of the TPP

The electric energy production in the TPP of Rades is based on a set of energies transformations using water as energy support. This water must have a high quality in order to guarantee the security of the installation and to improve the production groups' performances. It is therefore necessary to apply a rigorous treatment of the raw water and a stern control of its quality.

Figure 4 shows the architecture of the SCADA system of a TPP.

The pretreatment is constituted of two filtration chains each including a sand filter and an active coal filter. Thereafter, the water passes by the filtration chain then introduced in the inverse osmosis station and thereafter in the demineralization station (Figure 5).

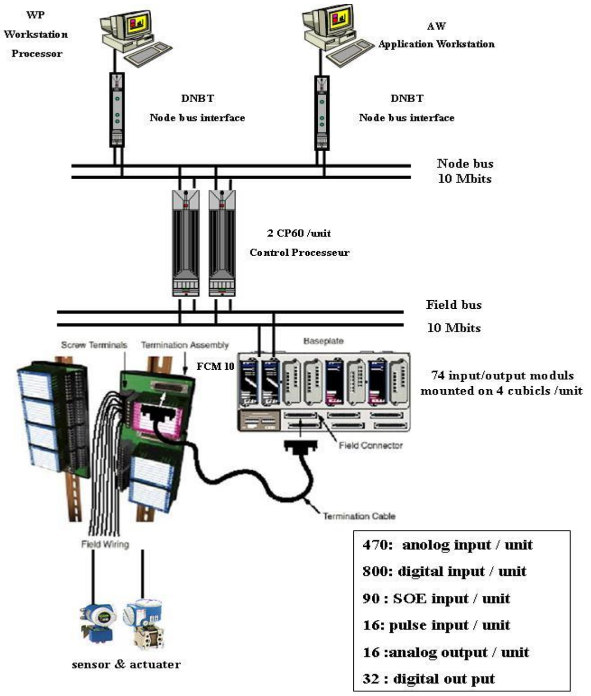


Fig. 4. SCADA system of the TPP

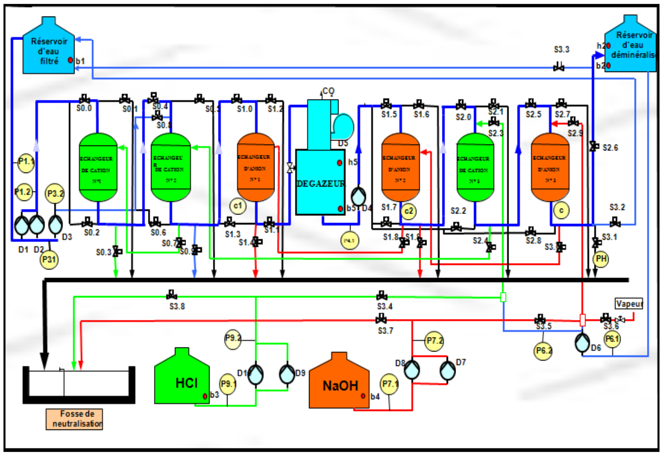


Fig. 5. Demineralization station of the TPP

B. Results of analysis

According to the first phase of the methodology established, a FA using the SADT method has been done (Figure 6). This analysis is a very interesting stage because it permits to describe the electric energy process of the TPP.

The second phase of the proposed methodology consists on the use of the SA-RT method for the analysis of a control-command application of the TPP.

Figure 7 shows the CFD of the water-steam station of the TPP.

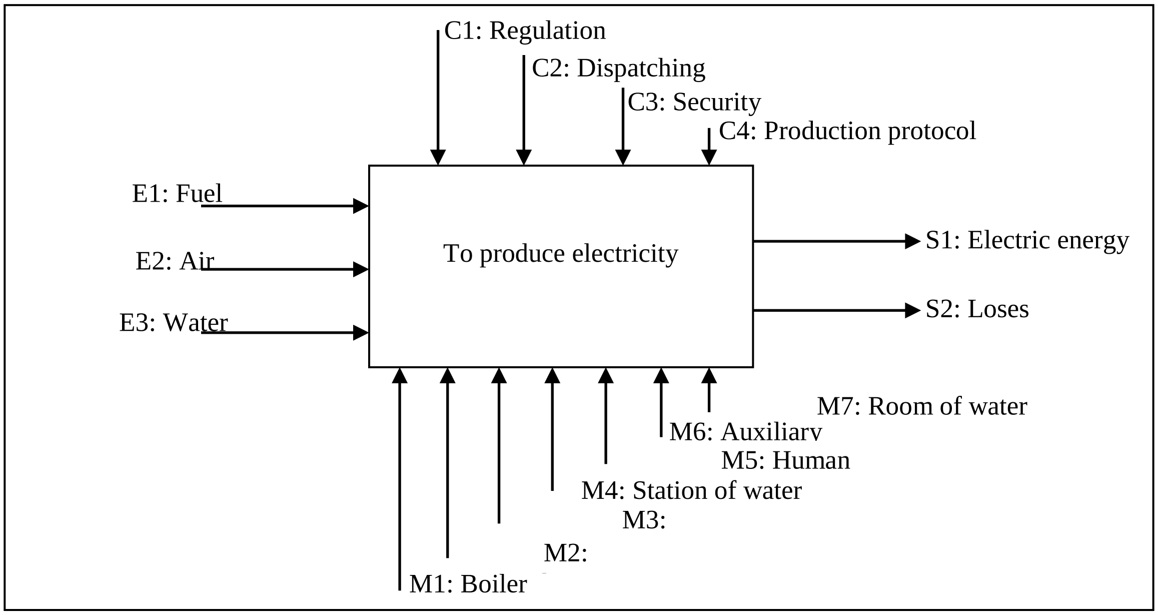


Fig. 6. Node A-0 of the SADT model of a TPP

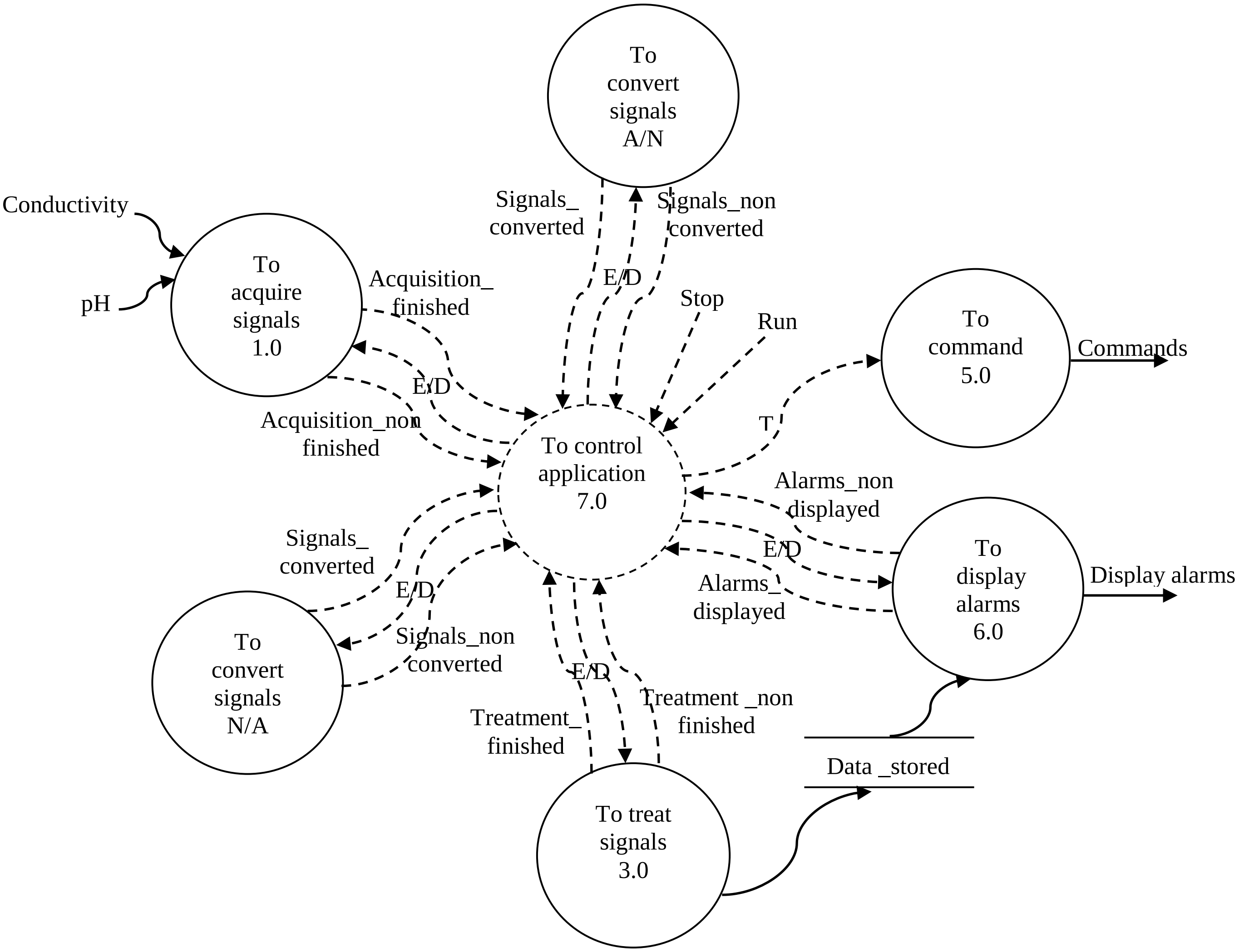


Fig. 7. Control Flow Diagram of the SCADA system of the water-steam station

The CD of the SA-RT model is constituted of one functional process « To pilot the control-command application 0 » and the terminators. It defines perfectly the interface between the designer and the client, that is, to provide or generate data in order to display these data on the tabular of the control-command application.

The DFD of the SA-RT model constitutes the first decomposition of the process presented in the CD. Then, we can identify the initial functional processes of the control-command application: acquirement process; treatment process; Human-Machine-Interfaces process.

The CFD of the SA-RT model includes the control aspect to the DFD elaborated. In fact, the implementation of the monitoring process at the level of preliminary diagram can express the sequence execution of the functional processes.

Figure 8 shows the state-transition diagram of the SA-RT model of a control-command application.

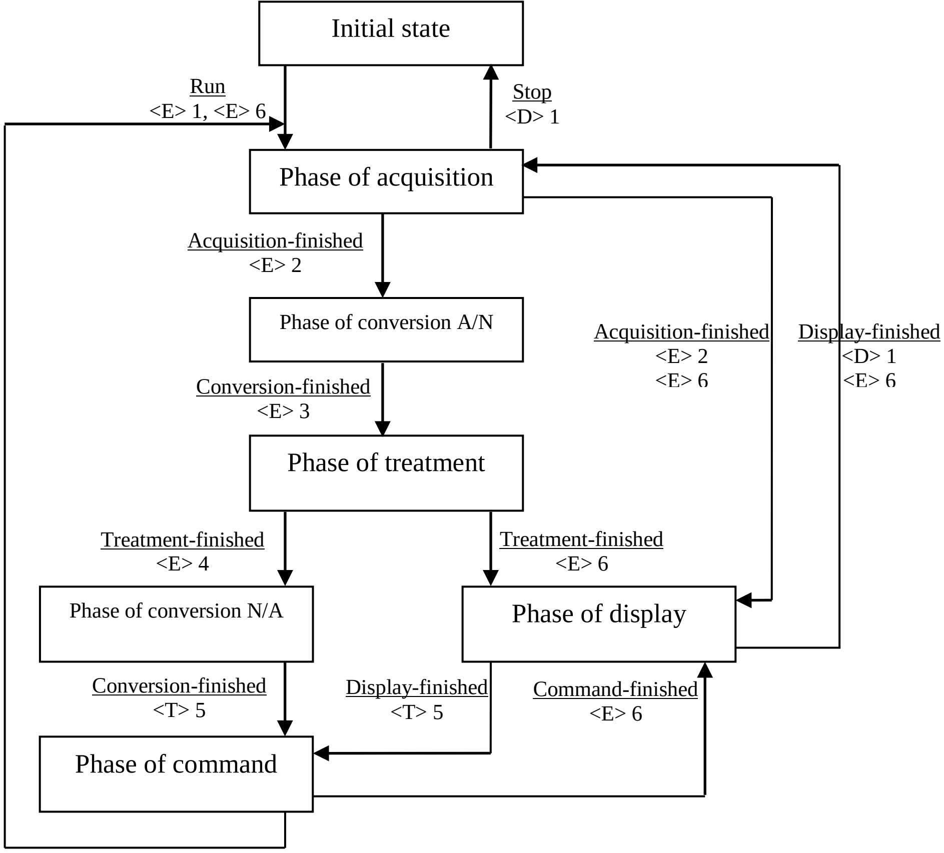


Fig. 8. State/Transition diagram of a SCADA system of the water-steam station

The potential uses for the SA-RT model are the design of the monitoring display and the diagnosis display. For the design of a monitoring display, the preliminary DFD of the SA-RT model presents an overall view of the control-command application. Indeed, information relative to each process represented through this level should appear in the monitoring display.

For the design of hierarchical diagnosis display, each DFD of the SA-RT model constitutes a vision at a given abstraction level. So, each of these DFDs gives a less or more detailed vision. In function of the objectives defined by the designer for each display, a particular DFD can supply the required information.

Finally, this application of the SA-RT method on the SCADA system of a TPP shows briefly the interests of the FA the design of supervisory systems.

C. *Results of supervision*

The last phase of the methodology proposed consists on the supervision of the control-command application. For the example of the water-steam cycle, we study the interfacing of the different signals: pH and conductivity of the ball furnace of the TPP. The application is declined in six stages: 1) choosing the site of the signal (FBM module); 2) programming both AIN and CIN blocks for the supervision of the signals pH (4 to 20 mA) and conductivity (alarm); 3) testing both AIN and CIN blocks by injection of current and by short circuit; 4) passing the cable between the sampling room and the SCADA room; 5) connecting the signals in the two modules 10FBM215 and 10FBM325; 6) conceiving a new tabular for the general vision of the sampling room.

Figure 9 shows the display of the sampling room containing the chemical analysis parameters of the water-steam cycle.

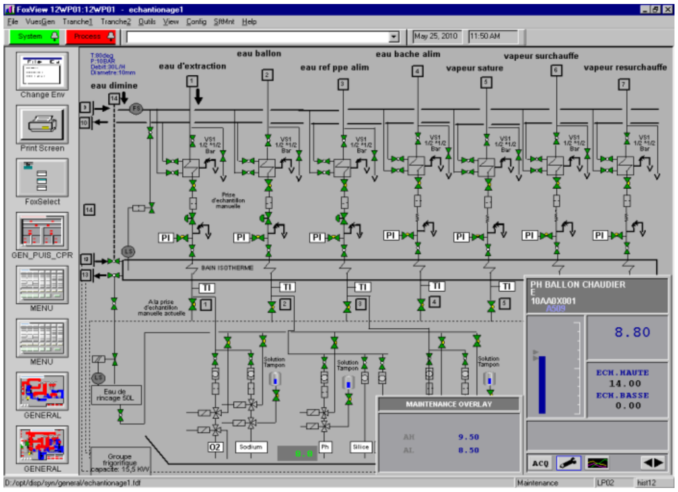


Fig. 9. Display of the chemical analysis parameters

In the first step is used the FoxView software which represents the interfacing operator to visualize the synoptic with a menu bar to activate the main functions of the SCADA system. Then, in the second step is used the ICC (Integrated Control Configurator) to create and to configure programs residing in the CP (Control Process).

To access ICC, we need from the menu bar Config choose term Control_Cfg-CIO_STN_Cfg and Config_station_name. It is used to configure the various blocks AIN, CIN, CALC, etc. Indeed, during the programming of a new signal it must be identified: the signal label; the compound and the address signal. For simulation is used FoxSelect software that permitting to reach the various elements of the hierarchy of data base of the CP.

The last step of this application is to improve the synoptic the pH meter and the conductivity meter and the design of a new synoptic of the sampling room containing these two data from the Fox Draw software. It provides tools to design displays for monitoring, alarming and process control.

To access FoxDraw we can choose function Config-FoxDraw from the menu bar. Included with FoxDraw is a large library of graphical components ready to be integrated and configured in displays. Thus, we proceeded as follows: creating of pH-meter blocks; operating test of the pH-meter; display configuration of different alarms; configuring of the overleay pH and testing the overleay pH- meter.

VI. Conclusion

In order to understand our world we need to think that almost everything can be regarded as a system, where signals must be monitored, controlled and ultimately supervised. In fact, the supervision of the production systems should by nature try to guarantee the observability, the controllability and, most important, the system stability.

In this paper, we presented a methodology of analysis and supervision of the production systems. We have suggested three main directions to achieve that methodology: (1) functional analysis of the production systems; (2) analysis of the control-command applications; and (3) supervision of the control-command applications in a SCADA environment. Two interesting functional analysis methods SADT and SA-RT are used on the one hand and a SCADA system for monitoring, control and fault tolerance, on the other hand. In fact, allowing the running of the production equipment to be understood, these techniques permit designers to decide the good information to display through the supervisory interfaces devoted to each type of supervisory task (monitoring, diagnosis, action...). In addition, functional analysis techniques might be a good assist to design support systems such as alarm filtering systems.

Staring from this case study of the analysis and supervision of the production systems discussed in this paper, work is in progress to develop a functional analysis and real time for different control-command applications in different production systems.

References

[1] M. Lambert, B. Riera and G. Martel, Application of functional analysis techniques to supervisory systems, Reliability Engineering and System Safety 64, 1999, pp. 209-224.

[2] D.J. Hatley and I.A. Pirbhai, Stratégies de spécification des systèmes temps réel (SA-RT), Masson, Paris, France, 1991.

[3] F. Cottet, Systèmes temps réel de contrôle - commande, Dunod, Paris, 2005.

[4] M.N. Lakhoua, SCADA applications in thermal power plants, International Journal of the Physical Sciences, Academic Journals, vol.5, N°7, 2010, pp. 1175-1182.

[5] M.N. Lakhoua, Surveillance of pumps vibrations using a SCADA, Control Engineering and Applied Informatics, Romanian Society of Control Engineering and Technical Informatics, vol.12, N°1, 2010.

[6] Y.L. Kaszubowski, R.S. Rosso, A. Leal, E. Harbs and M.S. Hounsell, Finite Automata as an Information Model for Manufacturing Execution System and Supervisory Control Integration, Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing, Bucharest, Romania, May 23-25, 2012.

[7] E.I. Gergely, Dependability Analysis of PLC I/O Systems Used in Critical Industrial Applications, Sudies in Computational Intelligence, 417, Springer, 2013, pp. 201-217.

[8] S. Reynard, O. Gomis, F. Bellmunt, A. Sudrià, O. Boix, and I. Benítez, Flexible manufacturing cell SCADA system for educational purposes, Comput Appl Eng Educ 16, 2008, pp. 21-30.

[9] R. Yenitepe, Design and implementation of a SCADA controlled MTMPS as a mechatronics training unit. Comput Appl Eng Educ 20, 2012, pp. 247-54.

[10] Z. Aydogmus and O. Aydogmus, A web based remote access laboratory using SCADA, IEEE Trans Educ, 52, 2009, pp. 126–132.

[11] P. Ponsa, R. Vilanova , A. Pérez, and B. Andonovski , SCADA design in automation systems, 3rd Conference on Human System Interactions (HSI), 2010, pp. 695 – 700.

[12] D. Gezer, H.O. Unver , Y. Tascioglu , K. Celebioglu and S. Aradag, Design and simulation of a SCADA system using SysML and Simulink, International Conference on Renewable Energy Research and Applications (ICRERA), 2013, pp. 1058-1062.

[13] W. Dong and S. Xian-li, The boiler design of remote monitoring system based on the SCADA, Chinese Automation Congress (CAC), 2013, pp. 864- 869.

[14] A. Morosan and F. Sisak, A SCADA system designed for making more efficient production in flexible manufacturing system, IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI), 2012, pp. 409- 413.

[15] S. Chun-Lien and C. Ya-Chin , A SCADA system reliability evaluation considering performance requirement, International Conference on Power System Technology, 2004, vol.1, pp. 574-579.

[16] H. Guobing, M. Guoqiang, Z. Jianmin and D. Jianling, Design for an embedded SCADA system, International Conference on Electrical and Control Engineering (ICECE), 2011, pp. 5819-5821.

[17] D.T. Ross, Structured Analysis (SA): A language for communicating ideas, IEEE Transaction on Software Engineering, 3(1), 1977, pp. 16-34.

[18] IEEE 1320.1-1998. IEEE Standard for Functional Modeling Language-Syntax and Semantics for IDEF0, IEEE, 1998.

[19] M.N. Lakhoua, Systemic analysis of an industrial system: case study of a grain silo, Arabian Journal for Science and Engineering, ISSN: 1319-8025, vol.38, 2013, pp. 1243-1254.

[20] T. DeMarco, Structured Analysis and System Specification, Prentice-Hall Inc., New Jersey, 1979.

[21] V. Benard, L. Cauffriez and D. Renaux, The Safe-SADT method for aiding designers to choose and improve dependable architectures for complex automated systems, Reliability Engineering & System Safety, 93(2), 2008, pp. 179-196.

[22] M. Lauras, J. Lamothe, and H. Pingaud, Une méthode orientée processus pour le pilotage par la performance des systèmes industriels, Journal Européen des Systèmes Automatisés, 41(1), 2007, pp.71-100.

[23] D.A. Marca, SADT/IDEF0 for Augmenting UML, Agile and Usability Engineering Methods, Marca, David., 2012, Software and Data Technologies, pp. 38-55.

[24] F. Jimenez, M. Courvoisier, A. Garcia, G. Munoz , N. Harchani, M. Al-Mohamed and D. Esteve, Tools and models for systems design and synthesis of MEMS based on asynchronous circuits, IEEE International Conference on Industrial Technology, vol.1, 2000, pp. 64- 69.

[25] I. Kuuluvainen and V. Ylitolva, The Bubble Operating System: a control extension to data flow diagram, IEEE Transactions on Consumer Electronics, 1991, Vol.37, Issue: 3, pp. 642- 650.

[26] A. Flo, M. Kjaernes and A. Skomedal, A bridge from structured analysis (SA/RT) to specification and description language (SDL), Eighth International Conference on Software Engineering for Telecommunication Systems and Services, 1992, pp. 93- 97.

[27] N.Sahraoui, Applying specification methods to complex systems, IEEE International Conference on Systems, Man, and Cybernetics, vol.5, 1997, pp. 4488- 4491.

[28] L. Urbain and B. Tondu, Robot controller specification using SART approach, Proceedings of the Third IEEE Conference on Control Applications, 1994, pp. 303-308.

[29] S. Lihua and J.A. Keane, Algorithmic aspects of hierarchical verification for SA/RT models, IEEE International Conference on Computational Cybernetics and Simulation, vol.3., 1997, pp. 2252- 2257.

[30] R.B. France and J. Bruel, Using Integrated Formal and Informal Modeling Techniques to Analyze Software Requirements: A Petri-Net/SART Case Study, 1996.

[31] M. Ben Hammouda, M.N. Lakhoua and L. El Amraoui, Dependability evaluation and supervision in thermal power plants, International Journal of Electrical and Computer Engineering, vol. 5, N°5: October 2015.

[32] M.N. Lakhoua and H.Laadhari, Supervision of the natural gas station using a SCADA System, Journal of Electrical and Electronics Engineering, vol.6, n°1, May 2013.

[33] M. Ben Hamouda and M.N. Lakhoua, Methodology of Operating Safety and Supervision of a Production System, CISTEM, IEEE, 3-6 Nov. 2014, Tunisia.

[34] R. Glaa and M.N. Lakhoua, Methodology of Analysis and Design of a SCADA System, CISTEM, IEEE, 3-6 Nov. 2014, Tunisia.

Effective Teaching Methods and Proposed Web Libraries for Designing Animated Course Content: A Review

Rajesh Kumar Kaushal
Department of Computer Applications
Chitkara University, CU
Rajpura, India

Dr. Surya Narayan Panda
Chitkara University Research and Innovation Network
Chitkara University, CU
Rajpura, India

***Abstract*—The primary aim of education system is to improve cognitive and computational skills in students. It cannot be achieved by just using the latest technology. This goal can only be achieved through effective teaching methods in combination with effective technology. Lot of researchers have offered effective teaching methods and published their findings in the past. Most of them offered teaching through animations, puzzles, games and storyline. This research paper focuses on identifying effective teaching methods offered by researchers and their findings by reviewing last few years articles published in renowned journals and conferences. Another aim of this paper is to propose ideas to make teaching tools more effective that can help students to understand difficult concepts deeply, improve cognitive and computational skills and retain knowledge for longer times. These ideas will serve as future research directions in this area. Another aim of this research paper is to introduce latest web libraries that can help educators to design animated courses.**

Keywords—cognitive; web education; dynamic teaching tool; animation libraries

I. Introduction

The traditional way of teaching was class-room learning which involves face to face teacher student interaction, learning through books, studying through teachers notes and appearing for final examination to obtain scores as high as possible. In traditional learning there was minor involvement of audio and video aids. The entire focus was on reading text. Presently teaching aids has changed revolutionary. In present era, educational aids involve traditional teaching style with addition to extensive use of audio and video aids. It all happens due to rapid growth of technology and its availability at low cost. The evolvement of audio video aids helps students to understand concepts better than earlier. But this teaching technology is less effective to improve cognitive and computational skills in students. One possible solution to this issue is that more time and efforts should be devoted in designing effective teaching methods rather than technology itself. Effective teaching methods and effective technology only can help students to learn deeply with higher engagement. Several teaching institutions and researchers has offered effective teaching methods that focuses on teaching through animations, by solving subject related puzzles, by playing games, quiz and by storyline. To achieve this they first redesigned the course into visual components (animations, puzzles, games etc.) and then offered it through web platform, mobile platforms and through offline applications. This research paper focuses on identifying effective teaching methods offered by researchers and their findings by reviewing last few year articles published in renowned journals. Another aim of this paper is to propose ideas to make teaching more effective that can help students to understand difficult concepts deeply, improve cognitive and computational skills and retain knowledge for longer times.

This paper first review relevant literature that depicts what efforts researchers have already made in the past to obtain better teaching results by using interactive teaching methods. Thereafter it discusses identified research problems and then this article will propose some useful web libraries to make interactive and dynamic web animations. At last article will close the discussion by suggesting future directions.

II. Effective Teaching Methods and Findings Based on Published Articles

In September 2011, Antonis, Daradoumis, Papadakis and Simos presented an evaluation methodological framework that could assess the learning methodology used [8]. It could also assess some of the educational and technical issues involved, and the solutions chosen to provide an easier to use learning environment to enhance the learning experience. In this article methodological framework was offered for the evaluation of distance learning. The entire focus was on three main evaluation parameters listed below [8]:

1) *Information and support should be provided to learner not only at the beginning but also during studies.*

2) *To evaluate learner's performance (final exam and continuous evaluation).*

3) *To evaluate learner's satisfaction (Enjoyment, Compensation, Benefits, Content)*

The course content was offered using animations, presentations and audio/video files that were uploaded on LAMS (Learning Activity Management System) and this framework also emphasis on student-teacher interaction for success. The LAMS tool was a popular learning system allowing authoring, monitoring and sharing. Researcher obtained positive satisfactory results in all research questions. To obtain results pre-questionnaire, post-questionnaire and final result analysis was used. In Fig.1 we are showing results of end-term questionnaire only.

TABLE I. OBJECTIVES & CORRESPONDING RESULTS [8]

Results (End-Term Questionnaire)		
Questionnaire Type	***Evaluation Framework Axis***	***Results***
End-Term	Information & support provided to learners	In a nutshell 55% students were very satisfied and 5% students felt disappointed.
End-Term	Learner's Performance	75% students believed that their overall performance improved and 25% students believed that either they should put more efforts or in other words they were not satisfied with their overall performance.
End-Term	Learner's Satisfaction	It is divided into several aspects like enjoyment, compensation, benefits, suitability of content, adequacy, applicability etc. and researcher got significant positive results in every aspect like 76% students enjoyed the course and 90% of students believed that material was adequate.

In 2011, Williams & Dugan claimed that presenting course material according to student's preferences and pace can help students to understand subject more effectively [6]. Even though there were three main objectives of the study but two of them were major as listed below:

1) Online Learning using GOAL can be as effective as classroom learning.

2) Adapting the presentation style to the preferred learning style of the student will enhance the student's learning experience.

To achieve above objectives researcher offered a project named GOAL (Guided on Demand Adaptive Learning) [6]. It allowed students to set the course pace and provides additional details/explanations whenever needed. The study was conducted on the subject named "Digital Logic Design". The course was designed into various activities like animated gates or circuits, puzzles and problems. Three topics were targeted namely Boolean algebra, sequential systems and addition. To evaluate the effectiveness of this approach, researcher collected data of three semesters and found that GOAL approach could achieve learning in less time as compared to class-room learning. According to the results shown in the article, average time to teach sequential system in classroom in spring 2010, fall 2010 and spring 2011 were 62.2, 50 and 67 minutes respectively. After teaching with GOAL average time required was 28.4, 19.3 and 25.8 respectively. The same effect was observed in rest of the topics also.

In December 2012, Hwang, Wu and Chen suggested that learning achievements and flow-experience could be enhanced by playing games and by web based problem solving [1]. In this research article the objectives were to study impact of online game approach on students learning, students flow experience and learning attitudes towards science learning. To find answers to these research questions they did an experimental study on 50 elementary school students who participated voluntarily. The study was conducted on butterfly ecology subject. Students were divided into two group's namely experimental group and control group. Experimental group studied through web based contents and control group studied through learning sheets and keyword search on internet. A board game was designed for experimental study. Students could move to the next location on board after throwing dice. On moving to each board location an activity triggered up and students had to solve it to move further. If the student failed to solve activity/problem a second time, learning system would show the correct answer. The game finished after solving all problems. Researcher got results through pre-test and pre-test questionnaire and post-test and post-test questionnaire. Mean score of experimental group was much higher than control group in learning achievements. T-test result for flow experiences and learning attitudes was also significant. The researchers, at the end of this study claimed that it might be difficult to claim that all of the findings are significant since the numbers of participants were not large and the activity period was short.

In January 2012, Rutten, van Joolingen and van der Veen tried to find the effect of simulation conditions on learning process [9]. To accomplish this task they didn't conduct any experimental study rather they reviewed relevant published articles from databases like ERIC, SCOPUS and ISI Web of Knowledge. Articles between 2001 and 2010 were reviewed. Final results were in favor of simulation conditions. The analysis also showed that simulation conditions were best applicable on laboratory usages.

In 2012, Jalal Kawash offered different teaching methodology. The study was conducted on first year computer science students. The idea was to improve learning by problem solving in combination with puzzle-based environment [12]. They targeted topics like basic set theory, graphs and trees, computer organization, databases fundamentals and programming concepts. Activities were designed like puzzles to improve critical thinking. At the end a survey was conducted to know the effectiveness of teaching mechanism through puzzle based along with problem solving. The results showed that majority of participants agreed that this approach is useful but offering the same course methodology second time gained better responses [12].

In 2013, Combefis, den SCHRIECK and Nootens designed a web platform developing algorithm design skills and they named this platform as ILPADS (Interactive Learning of Programming and Algorithm Design Skills). The website aimed to serve as working material to support teachers for their computer science courses in secondary schools [3]. They targeted students having age group 12 to 18. Every activity was decomposed into three different stages. In first stage learners were confronted to an interactive animation. It enabled them to discover algorithm and built it in their mind. In the second stage they were supposed to concretize the algorithm that they already had in their mind. They did it with an executable flowchart that could run on an instance of the problem. In the third stage learners were supposed to write code representing their algorithm. The entire ILPADS platform was designed in

HTML5, CANVAS and JavaScript. At the time of publication of this article no experimental study was conducted.

In 2013, O'Donovan, Gain & Marais suggested that well designed game based course content could increase engagement and encourage targeted behaviors among users [5]. In this research paper one of the university course names "Computer Games Development" was offered using gamification using online learning management tool named VULA. Objectives of their study are listed below [5]:

1) *Increase student engagement and motivation.*
2) *Improve lecture attendance and in-class participation.*
3) *Enhance content understanding and problem solving skills.*

Results were measured using marks obtained in the course, lecture evaluations, total lectures attended by students and questionnaire. This article mainly emphasis on few design principles that should be kept in mind while designing games and to obtain good results.

1) *Game should have special meaning for the user.*
2) *Game should be inspiring so that students feel motivated to master the topic.*
3) *Game should be autonomous (freedom of choice).*
4) *Game should have sense of discovery and visually pleasant.*
5) *Game approach should involve reward system like winning points, stars and badges.*

A questionnaire was distributed at the end of course to know students views on subject understanding, engagement and course marks. The results are shown below:

TABLE II. IMPROVED UNDERSTANDING [5]

Results (Questionnaire-Based)						
Improved Understanding						
Strongly Disagree	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***mean***	***Std Dev***
0	4	5	15	10	3,91	0,97
Improved Engagement						
Strongly Disagree	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***mean***	***Std Dev***
0	2	2	13	17	4,32	0,84
Higher Marks						
Strongly Disagree	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***mean***	***Std Dev***
1	1	11	15	6	3,71	0,91
Improved by Story and Theme						
Strongly Disagree	***Disagree***	***Neutral***	***Agree***	***Strongly Agree***	***mean***	***Std Dev***
1	2	11	16	4	3,59	0,89

The same course were also taught in 2011 and that time mean score was 70.8% with standard deviation of 10.3 but after offering the same course in 2012 with gamification the results were much better. In 2012 mean score was 74.9% with standard deviation of 8.6. The researcher also got positive significant results in average attendance which was 79.1% and was much higher than attendance in other computer science courses.

In January 2014, Brazilai and Blau did an experimental study to find the impact of scaffolding game based learning on learning achievements, perceived learning and game experiences [2]. In this research article the objectives were to study impact of game based learning with addition to external scaffolding on learner's ability, perception of learning, flow and enjoyment in the game. Another objective was to find correlation between learning achievements, perceived learning, flow and enjoyment. To conduct the study they used "My Money" website that includes set of online games and study materials that helped students to develop financial and mathematical skills. The study was conducted on elementary school students with average age group of 8-12 years. Participants were divided into three groups. One group would only play the game and this group was named as "Play only". Second group would study first (scaffolding) and then play the game. This group was named as "Study and Play". Third group would play first and then study (external scaffolding). This group was named as "Play and Study". The results are shown in Fig.3.

TABLE III. IMPACT ON FORMAL LEARNING ACHIEVEMNT [2]

Pre-Game and Post-Game Solving Results: Learning Achievement				
	Pre-game score		***Post-game score***	
Conditions	***M***	***SD***	***M***	***SD***
Play Only	2.62	1.87	2.58	1.89
Study and Play	2.72	2.03	3.47	1.98
Play and Study	2.83	2.06	2.55	1.85

In second condition "Studying First (external scaffolding) and then play" got better post results.

TABLE IV. IMPACT ON FORMAL LEARNING ACHIEVEMNT [2]

Pre-Game and Post-Game Solving Results: Perceived Learning, Flow, Enjoyment						
	Perceived Learning		***Flow***		***Enjoyment***	
Conditions	***M***	***SD***	***M***	***SD***	***M***	***SD***
Play Only	3.51	1.47	3.91	1.23	4.67	1.30
Study and Play	2.54	1.47	3.86	1.71	4.80	1.29
Play and Study	3.07	1.50	3.88	1.46	4.69	1.21

Adding the external scaffold before the game lowered learner's perceptions of how much they had learned from the game.

TABLE V. RELATIONSHIP BETWEEN LEARNING ACHIEVEMENT, PERCEIVED LEARNING , FLOW AND ENJOYMENT [2]

Correlation between problem solving assessment, perceived learning and enjoyment				
Variable	***1***	***2***	***3***	***4***
Problem Solving Assesment	-	-.11	-.01	-.05
Perceuved Learning		-	$.61^{***}$	$.50^{***}$
Flow			-	$.73^{***}$
Enjoyment				-

The table shows positive correlation between perceived learning and flow, perceived learning and enjoyment and between flow and enjoyment.

In 2014 researchers like Lee, Bahmani, Kwan, Laferte, Charters, Horvath and Fanny luor et al. worked on seven principles listed below to see their influence on novice programmers [4].

1) *Debugging first*
2) *Game-oriented*
3) *Fallible goals*
4) *Instruction*
5) *Scaffold*
6) *Help*
7) *Gender inclusiveness*

They presented a principled approach to teach programming using a debugging game called GIDGET, which was created using a unique set of seven design principles. The experimental study was conducted during the summer camp with a median age of 13.5 years. The primary purpose of this work was to find answers to the following research question [4].

"How does these seven principles influence the ways novice programmer learn programming concepts and solve programming problem"

Researcher tried to find answers to above stated objective by studying various barriers [4]. The article showed noticeable improvement in algorithm design barriers (19%) v/s learning phase barriers (51%). Two of the learning phase barriers improved by nearly 90% compared to the best algorithm design barrier improvement of 41%. Total 15 teams improved on learning phase barriers and on the other hand only 10 teams improved on algorithm design barriers. Teams especially struggled with composition barriers, encountering them frequently but demonstrating only 7% improvement – the least amount of improvement out of all barrier types. One interesting fact was that the algorithm design barrier did not greatly improve with instruction and practice. Algorithm design concepts needed more thorough explanations and help. Similar improvement was observed in both males (64%) and females (58%) which seemed to be encouraging [4].

In 2014, Edgcomb & Vahid stated that interactive web content like animations, responsive questions and interactive exercises can enhance students learning as compared to static text and drawings written in electronic textbooks [7]. The study was conducted on C++ programming. The primary objective of the study was "*To compare the lesson effectiveness of electronic textbooks having static content versus interactive web native content*". The researcher got positive significant results in all areas. Results are shown in Fig.6.

TABLE VI. PRE-LESSON RESULTS [7]

Pre-Lesson Result Based on Content Type	
Content Type	***AVG Pre-Lesson Correct Answers***
Static Web Content	1.8
Interactive Web Content	2.2

TABLE VII. POST-LESSON RESULTS [7]

Post-Lesson Result Based on Content Type	
Content Type	***AVG Post-Lesson Correct Answers***
Static Web Content	7.3
Interactive Web Content	8.6

All those participants who were assigned interactive web content improved 16% more than static web content. Participants spent average time of 17.5 minutes while using interactive web content and 9.4 minutes while using static web content [7].

In 2014, Costa, Toda, Mesquita and Brancher stated that learning through interactive games has shown positive results in many areas like marketing, education and health and now educators are introducing this concept in IT courses [10]. They focused on developing DSLEP (Data Structure Learning Platform) to aid higher education IT courses. The idea was to create activities that help students to enhance their understanding of various data structures concepts. DSLEP targeted major data structures concept like stack, queue, list, tree, search, graphs and hashing. One animated activity is designed for each concept using HTML 5 support. All activities were also offered on mobile devices for mobility purpose. Researcher also showed interest to conduct experimental study to see the effect of such tool on students understanding in future.

III. IDENTIFIED RESEARCH PROBLEMS

After reviewing literature it is found that animations used for teaching should not be static in nature rather teaching tools should offer concept like animation on demand so that students can deeply interact with tool as no one would like to learn through same animation demonstrating particular topic using same data set every-time. The learner should ask teaching tool to demonstrate a particular concept through animation by providing his/her choice of data set. The teaching tool should be smart enough to produce animation instantly. It can only happen through intelligent web scripting algorithm working behind the tool. Secondly to obtain accurate results, experimental study should be conducted in a long time span covering entire syllabus of particular stream. Moreover, every topic of entire syllabus should be taught using at-least two different styles (animations, puzzles, storyline, and games etc.).

IV. Proposed Web Libraries to Make Animated Course Content

There are several libraries that can be used to make animations. JavaScript is the most popular language available on web and it can be used for animation also. JavaScript can be fully utilized when used with HTML5 CANVAS. HTML5 CANVAS is especially introduced for making animations, web games and many more. JavaScript is difficult to handle when code becomes too large. So the alternate is using jCanvas which is built upon jQuery framework [11]. jCanvas can interact with HTML5 CANVAS and project is even easier to handle even if the code becomes too large. One another important aspect of jCanvas is that it works on layers, which is a very useful feature to make dynamic web animations. One another alternate is VELOCITY JavaScript library. It has large set of inbuilt functions to deal with animations.

V. Future Directions

After reviewing articles it has been observed that dynamic web based tool is required to teach students to enhance their subject specific knowledge. The proposed simulator won’t teach through fixed or static examples. Rather it can generate random data set and based on that it can produce new animation instantly and thus teaching same concept through endless examples. This kind of tool is feasible for practical subjects rather than pure theoretical subjects. Moreover, it has also been observed that while teaching courses through web animations, effort should be made to redesign entire course into animated form for better results. We should also like to suggest that each topic should be offered using at least two different approaches (e.g. teaching a topic through introductory animation and then through puzzle or teaching through interactive game or through storyline) wherever possible to obtain better results. Moreover teaching tool can be deployed on client/server architecture to reach maximum audience.

References

[1] Hwang, G. J., Wu, P. H., & Chen, C. C. (2012). An online game approach for improving students’ learning performance in web-based problem-solving activities. Computers & Education, 59(4), 1246-1256.

[2] Barzilai, S., & Blau, I. (2014). Scaffolding game-based learning: Impact on learning achievements, perceived learning, and game experiences. Computers & Education, 70, 65-79.

[3] Combéfis, S., den SCHRIECK, V. V., & Nootens, A. (2013). Growing Algorithmic Thinking Through Interactive Problems to Encourage Learning Programming. Olympiads in Informatics, 7, 3-13.

[4] Lee, M. J., Bahmani, F., Kwan, I., LaFerte, J., Charters, P., Horvath, A., ... & Ko, A. J. (2014, July). Principles of a debugging-first puzzle game for computing education. In Visual Languages and Human-Centric Computing (VL/HCC), 2014 IEEE Symposium on (pp. 57-64). IEEE.

[5] O'Donovan, S., Gain, J., & Marais, P. (2013, October). A case study in the gamification of a university-level games development course. In Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (pp. 242-251). ACM.

[6] Williams, R. D., & Dugan, J. B. (2011, October). Improving efficiency in engineering education through asynchronous computer-based instruction. InFrontiers in Education Conference (FIE), 2011 (pp. T4C-1). IEEE.

[7] Edgcomb, A., & Vahid, F. (2014). Effectiveness of Online Textbooks vs. Interactive Web-Native Content. In Proc. of 2001 ASEE Annual Conf. and Exposition.

[8] Antonis, K., Daradoumis, T., Papadakis, S., & Simos, C. (2011). Evaluation of the effectiveness of a web-based learning design for adult computer science courses. Education, IEEE Transactions on, 54(3), 374-380.

[9] Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. Computers & Education,58(1), 136-153.

[10] Costa, E. B., Toda, A. M., Mesquita, M. A., & Brancher, J. D. DSLEP (Data Structure Learning Platform to aid in Higher Education IT Courses).

[11] http://projects.calebevans.me/jcanvas/.

[12] Kawash, Jalal. "Engaging students by intertwining puzzle-based and problem-based learning." Proceedings of the 13th annual conference on Information technology education. ACM, 2012.

An Augmented Reality Approach to Integrate Practical Activities in E-Learning Systems

EL KABTANE Hamada
Faculty Semlalia
University Cadi Ayyad,
Marrakesh, Morocco

SADGAL Mohamed
Faculty Semlalia
University Cadi Ayyad,
Marrakesh, Morocco

EL ADNANI Mohamed
Faculty Semlalia
University Cadi Ayyad,
Marrakesh, Morocco

MOURDI Youssef
Faculty Semlalia
University Cadi Ayyad,
Marrakech, Morocco

***Abstract*—In the past, the term E-learning was mentioned to any learning method that used electronic machine for the distribution. With the evolution and the apparition of the internet, the term e learning has been evolved and referred to the online courses. There are a lot of platform which serves to distribute and manage the learning content. In some domain learners need to use some equipment and useful product for purpose completing the image built in the theoretical part by the practical activity part. However, most of those platforms suffer from a lack in tools that offer practical activities for learners. Using videos, virtual laboratories or distance control of real equipment as solutions to solve this lack were proposed but still limited. The mixed reality as new technology promised to create a virtual environment where the learner is an actor and can interact with the virtual objects. This article present an approach for developing integrated E-learning systems, helping to carry out the practical work by establishing a virtual laboratory that all tools and products can be manipulated by learners and teachers like in real practical activity, based on an augmented reality system.**

Keywords—E-learning; virtual reality; augmented reality; practical activities

I. Introduction

With the apparition of E-learning, a wide window has been opened to solve several problems such as the availability of media, team-learning and collaborative learning at different geographical positions without the need to travel.

The Virtual Learning Environment (VLE), not only provides rich teaching models and learning content, but also helps improve the ability of students to analyze problems and explore new concepts [1]. With the VLE, learning content has become interactive and the learner can make a self-evolution during his studies, allowing him to know his progress in a simple, sophisticated and precise way. The teacher, in turn, can follow each learner and controls his level since the platform.

However, E-learning suffer from a lack, especially in some domains, like Engineering, where learners need materials and useful products for a good training and keep learners' motivation and creativity.

In education, learners' creativity can be fostered in six different facets [2][3]:

1) *Self-reflective learning*
2) *Independent learning*
3) *Curiosity and motivation*
4) *Multi-perspective thinking*
5) *Reach for original ideas*
6) *Learning by doing*

A small study indicates that especially the facets 2,5, and 6 might be fostered insufficiently in education[4].So the use of laboratories in Education, – e.g. in material sciences, offer the opportunity to implement experiential and research. Within these laboratories, the learners have the chance to realize their own experiments and build their experiences in their learning processes. According to [5], there is a tiny enough open space for learners to work creatively with the course content and get in contact with real technical equipment. This is due to the availability and the cost of equipment, limited equipment resources, the availability of the environment of the experimentation and safety of the learners against the dangers they may come from experimentation (nuclear, chemical, etc.).

Many tools have been proposed as a technological breakthrough that has the power to facilitate the implementation of practical activities in education. Like the PeTEX project done by universities in Dortmund (Germany), Palermo (Italy) and Stockholm (Sweden) – implemented an opportunity to do experiential learning by using real laboratory equipment without being physically in the laboratory but having access via the internet [6]. Same for [7] who proposed "Tele-TP" as a solution that involves making a distance manipulation by remote control real instruments and synchronous telemetry. The use of real instruments and equipment consists maintenance all the time and its maintenance may require significant human and material resources.

The Augmented Reality and Virtual Reality are promised to be a new technology to develop Virtual Environments where the learner interacts with virtual objects. There are many other classes of Augmented Reality (AR) and Virtual Reality (VR) applications, such as medical visualization, entertainment, advertising, maintenance and repair, annotation, robot path planning [8]. Research and Application of Mixed Reality (MR) technology in education have enriched the form of teaching and learning in the current strategy in education[9].The marriage between AR/VR; the MR in VLE allows performing experiments that are up to now expensive and sometimes impossible in the real world.

An architecture integrating virtual learning environments in an adaptive e-learning system is proposed. eMouss@ide is a self-correcting adaptive learning system, offering learning scenarios based on the learners' learning styles and has the ability to self-correct based on traces and learners' feedback. The authors propose an upgrade to eMouss@ide platform that ensures the distance practical activities. The model uses the Augmented Reality (AR) techniques to present a complete course with practical aspect. The realization of such a system aspires to the completion of some activities in a virtual way like the practical work in distance learning courses or in high-risk environments (experiments in the presence of chemical or nuclear products,...) and it can economize on tools.

The rest of the paper is organized as follows: the second section presents a state of art; adaptive E-learning, Virtual Learning Environment (VLE) and Mixed Reality (MR) and Experimentation in E-learning. The Third section describes the proposal. In the fourth part outlines the results and the feedback of teachers and learners. Ultimately, the fifth section is dedicated to the conclusion.

II. Literature Review

A. Adaptation of E-learning

An E-learning system is a communication platform that permits learners to reach several learning tools, like document sharing systems, discussion boards, content repositories and assessments, without limitations of time and place[10]. Students have individual learning style preferences including: visual learners (V: they Learn from videos, pictures, graphs, charts and flow diagrams and schemas), auditory (A: they learn from speeches, they are a good listeners and they prefer to talk about what they learn) and kinesthetic (K: they learn from touch, hearing, smell, taste, sight also they tend to have the possibility to move, do experiments, go on excursions to live the experience, to examine and to manipulate the material) [11].

The internet represents an important source of information for learners, considering the number of information which offer on a specific domain that the learner can exploit freely, in one hand that make the utilization of internet in E-learning is very profitable [12].Especially after the apparition of the Web 2.0 and his tools such as podcasts, wikis and blogs, the diffusion of the information becomes easier and the e-learning has become more popular, but in other hand it can be a bad reference, because not any information is not credible also the large volume of information with a lack of supervision led the learner to a knowledge overlapping. Therefore, the organization of information according to the learner's needs is so important; here it comes the Adaptive Hypermedia Systems (AHS). Adaptive Hypermedia was presented as one of solution[13], especially where learners use an e-learning system without assistance from a physical tutor (teacher) who can ensure the adaptivity in the classroom. It offers a new functionality to the hypermedia systems by giving the users the possibility to consult smartly the content of the hyperspace. AHSs' objective is to adapt the user interface and the content and the navigational type [14]. The hypermedia system is adapted according to the user's knowledge, experience, knowledge background, preferences and his/her intentions [15][16].

There are a lot of adaptive educational hypermedia systems developed since 1996 [17] such as InterBook [18],ELM-ART [19], 2L670 [20], TANGOW[21], AHA! [22], INSPIRE [23] etc. The general architecture which they base these systems can be simplified to a learner model, a content model and an adaptation strategy, according to [24] (Fig. 1).

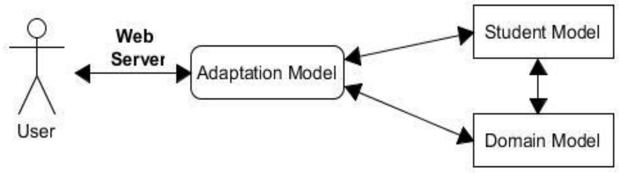


Fig. 1. The general architecture of an adaptive educational system

B. V irtual Learning Environment (VLE) and Mixed Reality (MR)

1) Virtual Learning Environment

It was noticed that the existence of the internet and the multimedia tools could be used to construct and build a new environment for education, called the Virtual Learning Environment (VLE). In this environment, learners and teachers are not obliged to be present at the same time or in a specified location (Classroom), unlike in the regular classroom environment, even if the teacher uses advanced technologies, still the limitation of the learners who must adapt to the location and time constraints of the course.

The Virtual Learning Environment (VLE) not only provides rich templates and various teaching learning content, but also helps improve the ability to analyze problems and explore new concepts of the learner. Integrated with immersive benefits, interactive and "imaginational" it built a shareable virtual learning space that all learners involved in the virtual community can access.

The VLE has the tools that administrators, teachers and learners need to realize their tasks. The students are given tests at specified times, and answers to the questions are transmitted to the teacher automatically [25]. With the VLE, learning content has become interactive and the learner can make a self-evolution during his/her studies, allowing, the learner, to follow his/her progress in the E-learning in a simple, sophisticated and precise way. The teacher, in turn, can follow each learner and controls his level since the platform. The VLE allows performing experiments that are up to now expensive and sometimes impossible in the real world [26].

Three of the most used open source VLE are DotLRN[1], Moodle[2] and Sakai[3]. DotLRN as an “adopted enterprise-class open source software for supporting e-learning and digital communities” for “supporting e-learning and digital communities”, Moodle defines itself as an “Open source software package designed using sound pedagogical principles, to help educators create effective online learning communities” and Sakai as an “Open source software freely available, feature-rich technology solution for learning, teaching, research and collaboration”.

2) *Augmented reality and Virtual reality*

On the Reality-Virtuality continuum by Milgram [27] (Fig.2), the AR is one part of the area of Mixed Reality. The AR system consists of adding virtual objects in a real-world environment. The Augmented Reality is an emerging technology with which a person can see more than others see, hear more than others hear and perhaps even touch, smell and taste things that others cannot[28]. According to [29],there is three characteristics must be integrated into an AR interface: the combination of the real and the virtual, the feasibility in three dimensions and real time interaction.

Several remarkable exists regarding the definition of Virtual Reality. This paper uses a definition of VR proposed by [30]; VR is defined as the use of a computer-generated 3D environment – called a “virtual environment” (VE) – that one can navigate and interact with, resulting in real-time simulation of one or more of the user's five senses. “Navigate” refers to the ability to move around and explore the VE, and “interact” refers to the ability to select and move objects within the VE.

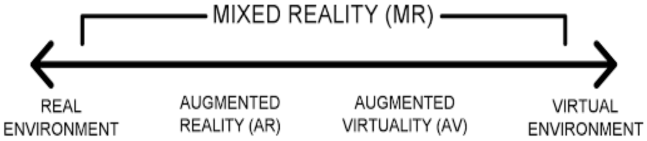


Fig. 2. Virtuality continuum (Milgram)

Augmented reality and virtual reality share some common characteristics such as interaction, immersion and navigation [31][32]. These characteristics can be derived from Azuma’s AR properties. As AR and VR technologies continue to advance, the possibilities for using AR and VR within several sectors growing and they have already found their place in many different domains like medicine, advertising, military, entertainment, manufacturing, education, as well as many others[33].

According to [34], several researchers have suggested that learners can improve their knowledge and enhance their motivation to learn with virtual and augmented reality. The usage of AR in education began since [35] created the Magic Book, where he uses a normal book as the main interface for objects and user can read the text normally (Fig. 3).

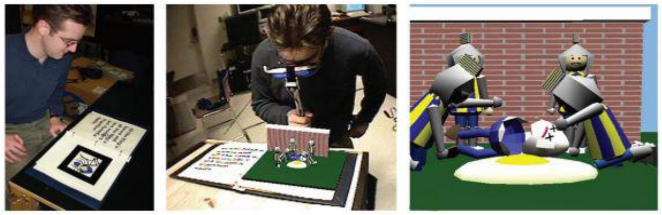


Fig. 3. Using the MagicBook interface to move between reality and virtual reality [37]

C. *Actual systems for practical activities in E-learning*

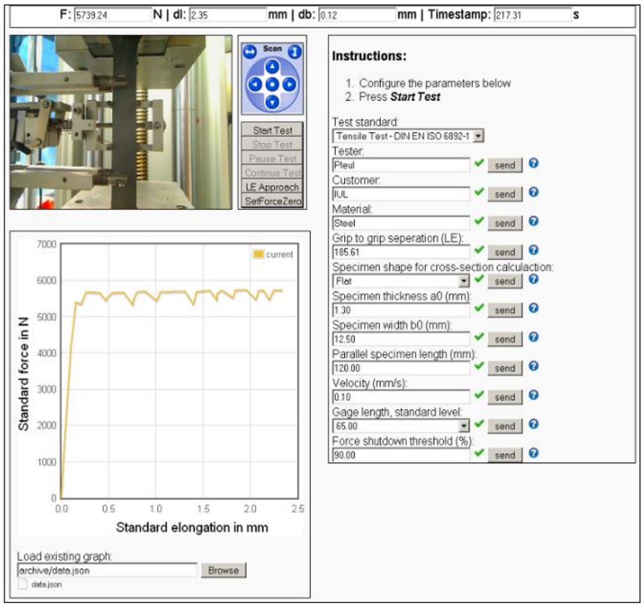
Klob stats and explained the meaning of the “learning” by: "Learning is the process whereby knowledge is created through the transformation of experience" [36]. "The use of laboratories is essential for the education in engineering and science related fields at a high qualitative level. Laboratories allow the application and testing of theoretical knowledge in practical learning situations. Active working with experiments and problem solving does help learners to acquire applicable knowledge that can be used in practical situations. That is why courses in the sciences and engineering incorporate laboratory experimentation as an essential part of educating students" [37].

At present, several solutions are implemented for research and educational activity:

As solution, the use of videos that were filmed and showed the unfolding of the practical activity, allow students to see the sequences in motion and to listen to the narration. The major limit of this solution is the luck of interactivity, the learner still passive, cannot interact with the objects [29] and just watch the video.

The use of real equipment remotely, [6] proposed PeTEX project (Fig.4) and [7] proposed “Tele-TP”. They have the same principle; offer the opportunity to do experimentation and use real instruments without being present physically in the laboratory, the learner just need access via the internet. E.g. Engineering students once they graduated will work with real technical equipment to solve real problems; those systems offer this contact with real materials remotely via internet. The main limits of this solution are the learner or the user needs internet to connect to system and if it happened and the user lose internet, there is risk that he lose his unsaved data. The use of real equipment always costly and the installation of the system is an additional investment in the purchase. The probability that the equipment break down is high, thus the laboratory equipment may need a periodic calibration, also require a lot of maintenance and it is necessary to prevent any distance contact by the users until the technicians finish the maintenance or the calibration. The maintenance may consists significant human and material resources and often takes time to restore the functioning of the system.

[1] DotLRN website: http://dotlrn.org
[2] Moodle website: http://moodle.org
[3] Sakai website: http://sakaiproject.org



F: 5739.24 N | dl: 2.35 mm | db: 0.12 mm | Timestamp: 217.31 s

Instructions:

1. Configure the parameters below
2. Press ***Start Test***

Test standard: Tensile Test - DIN EN ISO 6892-1
Tester: Pleul
Customer: IUL
Material: Steel
Grip to grip seperation (LE): 185.61
Specimen shape for cross-section calculaction: Flat
Specimen thickness a0 (mm): 1.30
Specimen width b0 (mm): 12.50
Parallel specimen length (mm): 120.00
Velocity (mm/s): 0.10
Gage length, standard level: 65.00
Force shutdown threshold (%): 90.00

Load existing graph: archive/data.json

Fig. 4. Interface of tele-operated Tensile Test [6]

Another solution proposed aimed to resolve the limits of the previous solution, is by the use of Virtual Laboratories. VMSLab-G [38] use virtual reality approaches to describe chemical experiments at both human and molecular level (Fig.5). The user can walk-by and enter the rooms. In this way, one can get in contact with the experimental setup, use the various components and follow a given protocol by driving the mouse and activating the relevant sensors. This solution solve the problem of using real material because all the equipment are virtual also the user is safe, but the main limits of this solution are the user needs internet, for connecting to system. If the connection is lost, the user may lose the unsaved data. The lack of virtual equipment setup experience and of hands on debugging experience and trouble shooting. Those virtual laboratories have a license limit to how many users can run simultaneously. In addition, the user still confined to using the keyboard, the mouse and sometimes joysticks [39].

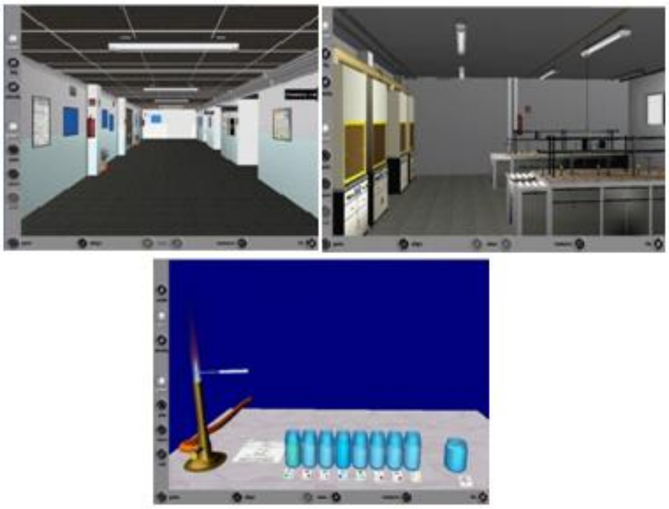


Fig. 5. The main hall of VMSLab-G and an experimentation using flame spectroscopy experiment [40]

III. Proposal

By introducing AR in eMouss@ide system, the authors ensure the existence of the practical activity part.

A. eMouss@ide System

eMouss@ide[40], is a self-correcting adaptive learning system, offering learning scenarios based on the learners' learning styles and has the ability to self-correct based on traces and learners' feedback.

This system allows the teacher to index educational resources, to design, to edit and view learning scenarios as well as the learner can pass an MBTI test to identify his learning style, search for educational resources and learning scenarios according to his learning style, study the adapted courses to his profile, and evolve courses' adaptation offered at his learning mode.

The learning style detection of the learner is through the MBTI test, the learner passes from first inclusion in the platform.

1) The MBTI test

This test is a psychological test to determine the personality type of the user who passed the test, in this case it is the learner, allowing each individual to know his manner of perceiving the world and his way of acting; so it allows the learner to know his weaknesses, his potential and his 'energy sources'. The concept of psychological types proposed by Carl Jung [41] and then developed by Katherine Briggs and her daughter Isabel Myers into a practical self-assessment tool called the Myers-Briggs Type Indicator (MBTI). The MBTI has four categories and each category represents two opposite poles (Tab. 1).

TABLE I. The Dimensions Proposed by the Mbti

Dimensions	**Preferences**	
Orientation of energy	**E** Extraversion	**I** Introversion
Collection of information	**S** Sensing	**N** Intuition
Decision making	**T** Thinking	**F** Feeling
Mode of action	**J** Judging	**P** Perception

The psychological types are defined by a combination of four letters, these four letters used to designate 16 different psychological types.

Determining personality type is by selecting the letter from the dominant part of the dimensions (orientation of energy (E / I), Collection of information (S / N), Decision Making (T / F), Mode of action (J / P)).

2) Limits of the eMouss@ide system

The eMouss@ide is an adaptation system of courses for the learner's profile according to his personality style using the MBTI test and self-correct based on the learners' comments.

The theory remains inadequate in some matter, so learners need equipment to realize practical activity to complete the image built in the course.

eMouss@ide and most of the e-learning platforms are limited to the level of existence of experiments and practical activity management.

B. *eMouss@ide's architecture:*

The platform eMouss@ide is composed of two parts Client and Server (Fig. 6); the client part contains two main users: the teacher has as role indexing educational resources and design scenarios. The second user is the learner how can take courses adapted to his personality after its determination by an MBTI test on the first registration of the learner in the system.

The two actors in the system have an appropriate interface depending on the users' tasks.

For the 'Server' part includes the warehouses, the database and the modules' part:

- Modules
  - Module "Indexing educational resources" used by teachers to technically and pedagogically indexing the pedagogical resources used in the learning scenarios.
  - Module "Scripting" used by teachers to create and index pedagogical scenarios.
  - Module "pedagogical relationship" used by the system to connect the learning styles with teaching strategies.
  - Module "Test learning style" used by the learners to determine their learning style.
  - Module "Learning" used by learners to take courses.
  - Module "Self-correction" used by the system to better adapt the course to the learner's learning style by using comments of learners at the end of the course.
- Warehouses and the database
  - The learner's database contains the information relating to learners (name, level, learning style ...).
  - The warehouse "Educational Resources ERs" contains educational resources indexed technically and pedagogically.
  - The "scenarios" warehouse contains the scenarios indexed technically and pedagogically by the teacher.

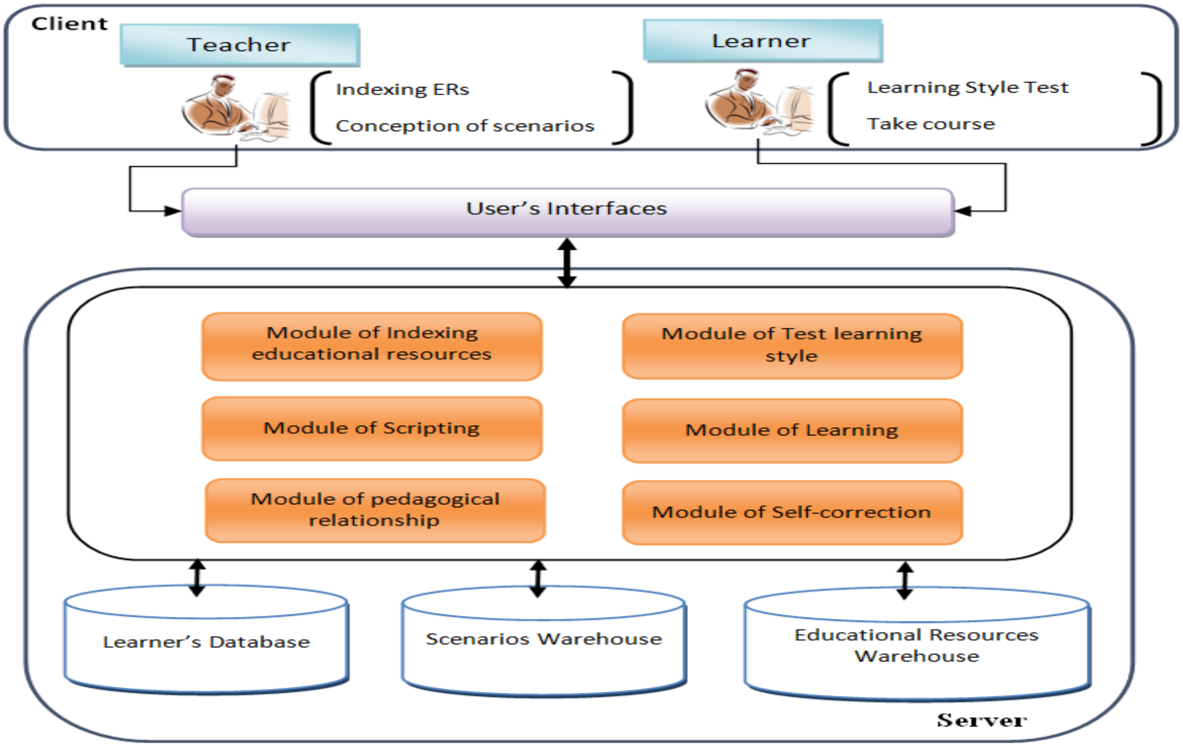


Fig. 6. General architecture of eMouss@ide [41]

C. *General Architecture of the modified system*

By using AR, the users can interact with virtual objects like if they are real objects in front of them.

The proposed upgrade to eMouss@ide platform ensures the distance practical activities using virtual reality and augmented reality, and manage the adaptation of those practical activities to the profile of each learner.

After the integration of the virtual practical activities solution, the eMouss@ide system architecture becomes like figured in Fig.7.

D. *General Architecture of the solution*

In the solution presented in Fig. 8, there are three actors: the designers, the teachers and the learners. The designers and the teachers are principle users in the design module; this module consists to create and to test the practical activities. The learners are principle user in the exploitation module, where the learners realize the saved practical activities in the "practical activities database" through the virtual environment (Fig. 9). The communication between the system and the users pass through a web interface.

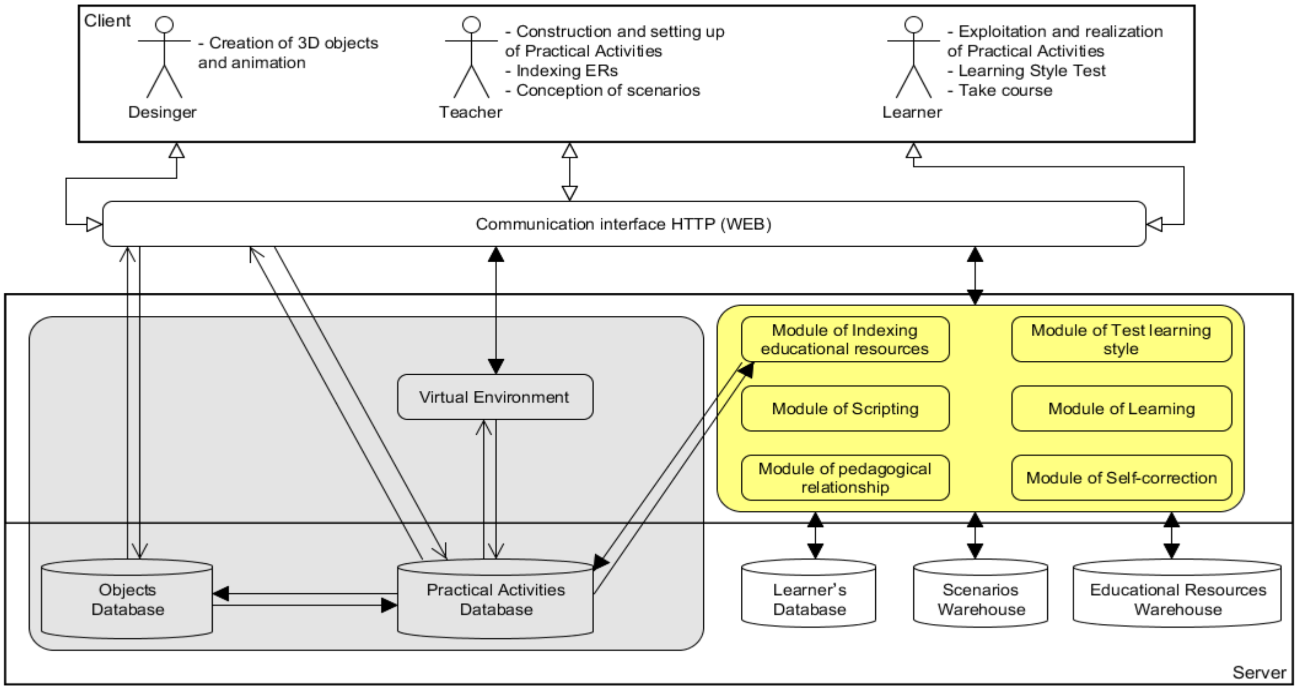

Client
- Creation of 3D objects and animation
Desinger
- Construction and setting up of Practical Activities
- Indexing ERs
- Conception of scenarios
Teacher
- Exploitation and realization of Practical Activities
- Learning Style Test
- Take course
Learner
Communication interface HTTP (WEB)
Virtual Environment
Module of Indexing educational resources
Module of Test learning style
Module of Scripting
Module of Learning
Module of pedagogical relationship
Module of Self-correction
Objects Database
Practical Activities Database
Learner's Database
Scenarios Warehouse
Educational Resources Warehouse
Server

Fig. 7. General Architecture of the modified system

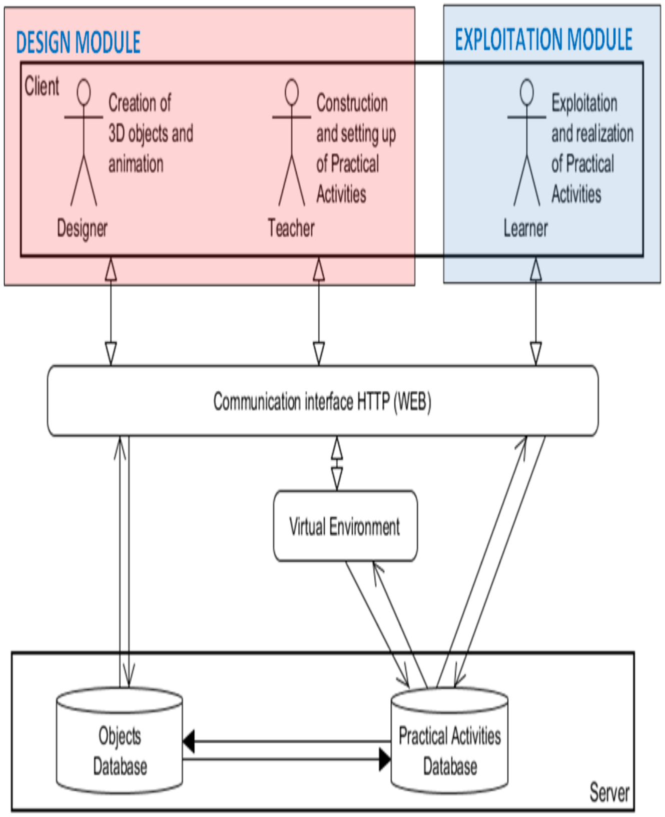

DESIGN MODULE
EXPLOITATION MODULE
Client
Creation of 3D objects and animation
Designer
Construction and setting up of Practical Activities
Teacher
Exploitation and realization of Practical Activities
Learner
Communication interface HTTP (WEB)
Virtual Environment
Objects Database
Practical Activities Database
Server

Fig. 8. General architecture of the solution

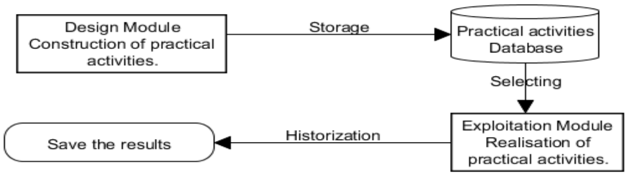

Design Module Construction of practical activities.
Storage
Practical activities Database
Selecting
Exploitation Module Realisation of practical activities.
Historization
Save the results

Fig. 9. General module of the system

1) Practical Design Module

In the design module, the teacher write a detailed description for the proposed practical activity and contacts the designer whore covers in his turn the description of the practical activity and verifies the existent objects in the database and creates the virtual equipment and there animation and storage them in the database (Fig. 10). The designer chooses the equipment to use in the practical activity from the virtual laboratory (Fig. 11) that imports the objects from "objects database". He prepares the practical activities and contacts the teacher to confirm the practical activity and storage it in "practical activities database". In addition, the teacher can modify the space of experimentation (e.g. initial positioning of the experimentation equipment) and tests the practical activity in the Virtual Environment before that the learners passes it. After that, he proceeds to create the quizzes to evaluate the learners' knowledge regarding this course. The other task consists to follow up the progress of learners based on participation in the session, delivered exercises also on realization and reports of practical activities and on the results of quizzes. Based on the results of monitoring learners, the teacher can contact the learners who appear weak to solve the problems they suffer.

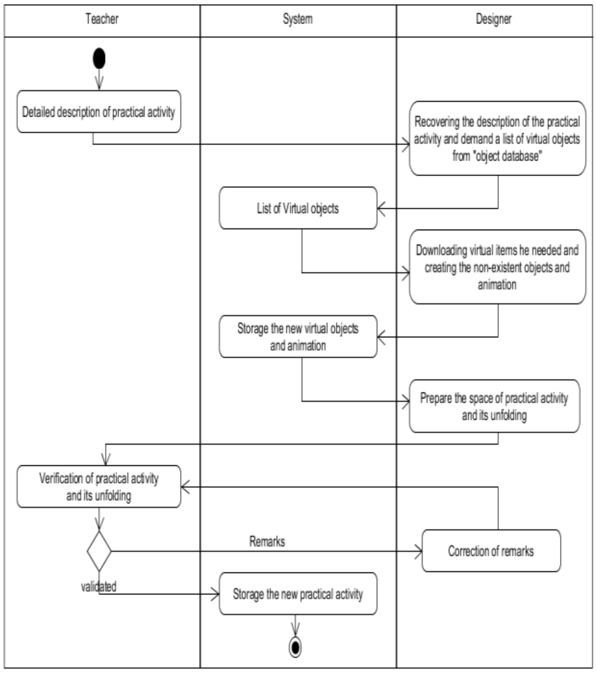

Teacher
System
Designer
Detailed description of practical activity
Recovering the description of the practical activity and demand a list of virtual objects from "object database"
List of Virtual objects
Downloading virtual items he needed and creating the non-existent objects and animation
Storage the new virtual objects and animation
Prepare the space of practical activity and its unfolding
Verification of practical activity and its unfolding
Remarks
Correction of remarks
validated
Storage the new practical activity

Fig. 10. Activity diagram of practical activities' creation

Software:

The Staging of the practical activity starts after that the designer receives the description of the practical activity and verifies the existent 3D object, by creating the non-existent virtual equipment and there animation using 3ds Max [42] or Cinema 4D [43],which are a 3D modeling software, it offers a complete solution for modeling, animation, simulation and rendering for game designers and film, as well as computer graphics and stores them in the database. After that, the designer proceeds to the screenwriting of the practical activity using Openspace3D software, which is a free and Open Source development platform for interactive real time 3D projects, it offers a solution creating a whole interactive 3D scene, with great graphical quality, without writing any code [44], that based on the principle of drag and drop of "Plugit".

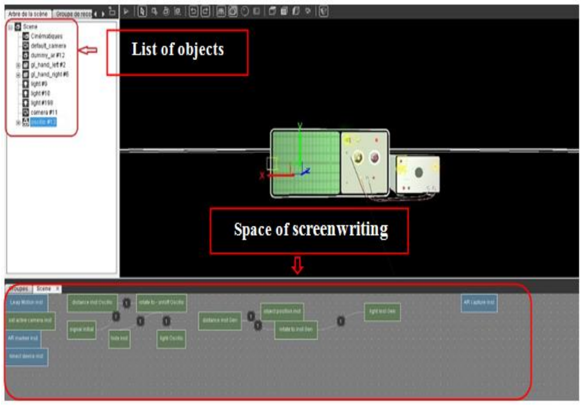


Fig. 11. Scene and list of equipment and scenario of interaction

The creation phase:

The practical works are created in OpenSpace3D. It used to create a virtual environment (or world) which the designer use as a scene to import the 3D objects from the "object database" and put them in the scene. Using a camera with its "Plugit" and combine it with the "Plugit" of augmented reality, the designer ensures the environment and the display of the virtual equipment. The designer have to make sure of the user's (the learner or the teacher) interaction with virtual environment, to supply it the designer and the other users must use a leap motion that help to detect the hand gesture. Therefore, the designer use the Leap motion "Plugit" and configures the set of hand gesture e.g. the punch to select and take the objects, the palm (open hand) to release the selected object, the index finger to click on a button.

2) Exploitation Module

The registered learner, have access to courses according to his formation and his level of education. Therefore, he can consult courses adapted to his profile and ask for teacher's assistance any time he needs it. After he understood the course, the learner proceed to the proposed exercises to fix the concepts learned in the course. Then the learner passes to the practical activity part to better clarify the lesson. Therefore, the "Module of indexing Educational Resources" of the eMouss@ide system shows the existing practical activities depending on his profile and level. After that he chooses his practical activity, the platform prepares the virtual environment where the practical activity takes place and begins the realization then writes a report of the results obtained and delivers it to the teacher. At the end, the learner passes a quiz that contains questions about the course and practical activity and its results.

a) working environment

Hardware:

To do the practical activities the learner needs a computer to display the virtual environment, a camera to generate the augmented reality's space and a Leap Motion to ensure the learner hands' gestures.

Software:

After the validation of the practical activity between the teacher and the designer, the learner finds all the objects needed to practice the practical activity in the scene of Openspace3D. The learner realize his practical activities in Openspace3D in the "Player mode" which is a passive window that the learner don't have the privilege to modify anything in the screenwriting scene.

b) Interaction modes

There are three ways to ensure the interaction of the user with the virtual objects:

- The first way to interact with the virtual objects is by using markers which are the reference point between the reel space and the augmented space. They are predetermined physical object detectable with the camera when the application is running. E.g. As figured in Fig. 12(a) a blue box present the marker of the virtual device and on the user's finger a blue sticker after running the application the camera detects the markers and converts them to virtual objects (Fig. 12-b). The interaction using markers make the precision of the virtual objects' location in the space easier but still limited to the configuration of the marker, the user needs to take a marker all the time in his hand to select an object, to do another action he needs to change the marker.

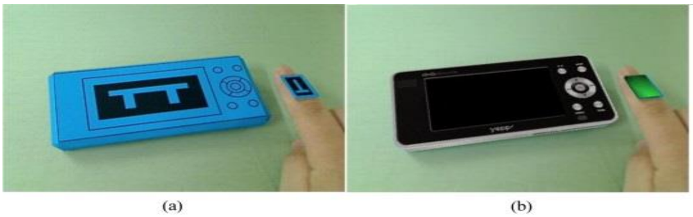


Fig. 12. Interaction using markers [45]

- Second way, the chosen one in this solution, is using a Leap motion, that contains sensors and detects the user's hands and there gestures. The interaction with the virtual objects in the augmented space is done using a virtual hand that appears when the user's hand detected and copies all the gestures of real hand. It provides seamless interaction in AR environments so the user can interact (select, take,…) with the virtual objects (Fig. 13a,b). The interaction is markerless and the configuration of the hand gestures is easier that makes the virtual objects' manipulation (catching, taking, turning, ...) much fluid. Therefore, this way needs sensors (Leap motion) that's mean some additional costs.

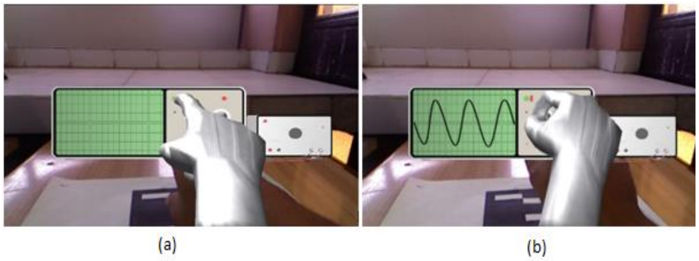


Fig. 13. a- User detection and apparition of experimentation tools b- Interaction with the oscilloscope and the generator

- The third way is using hand position in 3D environment by Stereo Camera [46]. The human hand is detected from input video image by use of the skin color model and image segmentation used in Chun's approach [47]. In the second phase, the 3D positions of the hand such as fingertip, the center of the palm and the center of the marker are evaluated using disparity map of the stereo-vision. In the third phase, the user can interact with a virtual object by detecting collision between the human hand and the object.

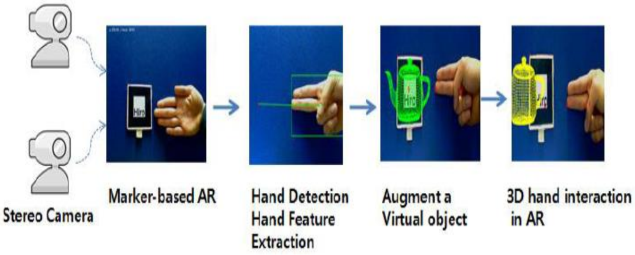


Fig. 14. Natural hand interaction[46]

c) Procedure

The learner starts by downloading the practical activity's instruction, reads it and he can ask for assistance help. Next, the learner begins the practical activity by following the instructions and takes notes to write a report. At the end, the learner saves the report and submits it to the teacher (Fig. 15).

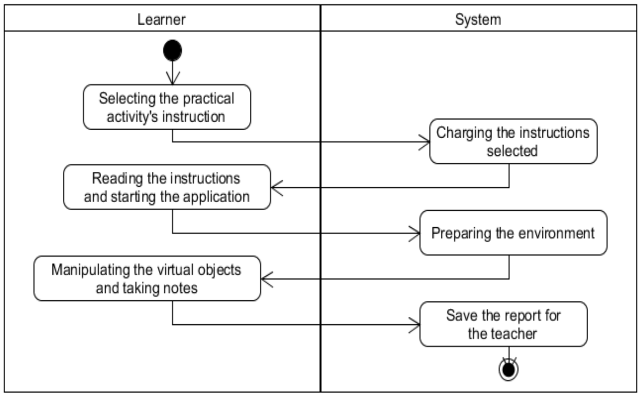

Learner
System
Selecting the practical activity's instruction
Charging the instructions selected
Reading the instructions and starting the application
Preparing the environment
Manipulating the virtual objects and taking notes
Save the report for the teacher

Fig. 15. Activity diagram of the practical activities' procedure

IV. Experimentation

A. Protocol

1) Population

The objective of this research is to measure the effect of the use of the MR in the VLE. To do this, it was proposed to a group of 10 students with different learning styles a scenario of practical activity using a virtual oscilloscope. It is noteworthy that none of the learners had any experience with this learning environment.

2) Sequencing of the practical activity

The execution of the experiment is done individually with each learner has a computer with the virtual environment, a camera and Leap motion connected to the computer. This practical activity named activity 1 as divided into 2 parts. The first part is an introduction to "Oscilloscope" and the second part is to experiment by taking measurements and render the generator voltage. Before launching the practical activity, the learner downloads the instruction manual detailing the required task. They had an unlimited access to the system (Fig.16).

At the beginning, the activity space was empty. Upon the detection of a user, the virtual oscilloscope appeared in the activity environment (Fig. 13a) where the learner started the first phase of the practical activity and began to recognize the elements of the oscilloscope by selecting any element to see its description. For example, if the learner had selected the setting button, it will display all information about this potentiometer. The same thing would happen for various buttons and keys on the oscilloscope. Here, the learner can repeat, as many times as he wants until he understands all the different components and their functions.

For the second part of this practical activity, the learner must manipulate and interact with various buttons of the oscilloscope and the generator in order to make the required measurements. The learner started by connecting the generator with the oscilloscope using cable (Fig. 13b), then he turned on the generator and the oscilloscope and proceeded to adjust the position of the signal to make it in the center of the screen. After that, he began to vary the voltage of the signal generator and observed the changes in the oscilloscope screen and started taking the measurements for report writing that had to be saved and submitted (Fig. 16).

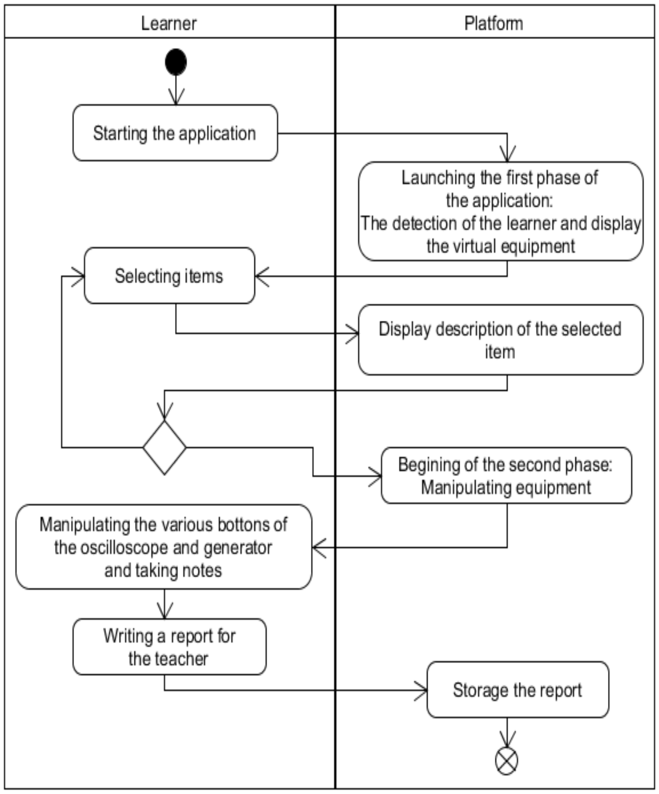

Learner
Platform
Starting the application
Launching the first phase of the application: The detection of the learner and display the virtual equipment
Selecting items
Display description of the selected item
Begining of the second phase: Manipulating equipment
Manipulating the various bottons of the oscilloscope and generator and taking notes
Writing a report for the teacher
Storage the report

Fig. 16. Activity diagram of first activity's proceeding

3) Collected data

Once the learners complete the experiment, the behavioral data are collected: report of the activity, time of realization of the experiment, the number of requests for assistance and help, number of consultation of instructions and number of errors (when the learner does not select the right element to perform an action).

B. Result

1) Report of the activity

The report delivered by the learners has two parts: Questions section on the components of the oscilloscope (the first part of the practical activity), second part the measurements and calculations done during the experiment. After correcting the reports, 100% of learners validated the first part of the activity, which was a questionnaire containing inquiries about the components of the device and its operation. Same for the second part of the activity 100% of the learners were able to complete the experimentation and realize the required calculations.

2) Time of realization

Time of realization of the experiment varied from one learner to another, analytical comparisons indicate that the procedure duration for realization is significantly shorter when the learner requires less the assistance of the technician or the teacher and doesn’t opt to the utilization or that of practical activity instructions and of course doesn’t make many mistakes.

3) Number of requests for assistance

Number of requests for assistance or help varies from one student to another; analytical comparisons indicate that the number of request for assistance depends on comprehension of the learner to the new concept of the experimentation and the manipulation of tools and to the tasks to realize.

4) Number of consultation of instructions

This number varies from one learner to another due to the lack of student’s concentration during the first reading, which influence the global time needed for the experimentation achievement.

5) Number of errors

The errors are classified into two categories: the first one, regroup the cases where learners select the wrong element in the machine to execute an action. The second one, covers the problems caused by the inability of the camera and the leap motion to detect the right gestural made by the learner.

6) Detailled results

The Tab. shows the students' result including the score of the quiz, the realization of the experimentation, the timing spent in the experimentation, the number of errors, the number of help request and the number of instruction’s consultation. In general, all the students passed the practical activity and answered the quiz and the general score mean was 100% and 97%, respectively.

First of all, the first question is about the organization and clearness of the experimentations in the platform, as a result, the students’ satisfaction was 4.7 (very good). This result highlights the importance of the organization and the display of the experimentation, so it is easy for the learner to find the experimentation that he wants. For the evaluation of the Q2, the students have been very positive, by a score mean of 4.3, that the guidelines are appropriate. In addition, the students strongly agreed (with a score mean of 4.2) that the system improves the personal effort (Q3).Furthermore, the students were satisfied about the system (Q4) and the high score mean 4.4 proved that. Same for the students’ need for this system (Q5), the majority choose that they need the system (score mean 4.6). Beside for the comparison between this system and the traditional system (Q6), the students considered that this system is very good by a mean score 4.1.

As a summary, most of the students agreed that this system is able to meet the learning and teaching objective and specially the practical activity part. Furthermore, all the students agreed that this system is an effective tool for learning and teaching domains that need a practical activity part. This is because all the participants’ students have validated the exam by a score mean of 97% (Tab.2).

TABLE II. RESULTS OF STUDENTS

Learners	Reports		Time of realization (sec)	Number of requests for assistance	Number of consultation of instructions	Number of errors
	Quiz	Realization of experimentation				
Learner1	100%	100%	453	3	4	4
Learner2	100%	100%	498	3	4	7
Learner3	100%	100%	486	3	4	5
Learner4	100%	100%	587	3	4	5
Learner5	100%	100%	610	4	3	4
Learner6	80%	100%	658	5	5	6
Learner7	90%	100%	672	5	5	6
Learner8	100%	100%	521	4	4	3
Learner9	100%	100%	324	4	4	5
Learner10	100%	100%	395	3	4	6
TOTAL (score mean)	**97%**	**100%**	**5204 (520,4)**	**37 (3,7)**	**41 (4,1)**	**51 (5,1)**

TABLE III. RESULTSOOF QUESTIONNAIRE ANALYSIS

Topic	Poor (1)	Fair (2)	Good (3)	Very good (4)	Excellent (5)	Score Mean $\bar{X} = \frac{\sum_{i=1}^{n} x_i f_i}{n}$	Standard deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{X})^2 fi}{n}}$
Q1. The experiments are clear and organized	0	0	0	3	7	4.7	0.597
Q2. The guidelines are appropriate	0	0	2	3	5	4.3	0,781
Q3. The system improves the personal effort	0	0	2	4	4	4.2	0.748
Q4. Overall reaction to the described product	0	0	1	4	5	4.4	0.663
Q5. Your need for this system	0	0	0	4	6	4.6	0.489
Q6. This system comparing to the traditional system	0	0	3	3	4	4.1	0.86

V. CONCLUSION AND FUTURE WORKS

In this paper, the authors have presented a solution for the lack of practical activities in the E-learning platform without using remote control of real equipment. This solution was introduced in the existing adaptive learning system eMouss@ide and can be integrated in other learning systems.

The teacher presents a detailed description on the practical activity, including equipment needs and the functional of each of the equipment to the designer who supports the creation of the environment of the practical activity and ensures the interaction between the user (teacher or learner) and the augmented world. By using AR technology and Leap motion, the learner can get an interaction with virtual objects and can realize the practical activities, which reduces the risks of overload and failure of materials.

Thus, for the future work, we will try to eliminate the Leap motion and any other type of sensors to ensure the interaction with the AR's objects. The determination of user hand's gestures will be detected by using a just a camera so the learner just needs his computer and a simple webcam without any other additional costs. Also we will try to implement this solution as module of practical activity in other E-learning system and platform.

REFERENCES

[1] P. Dillenbourg, D. K. Schneider, and P. Synteta, “Virtual learning environments,” in 3rd Hellenic Conference “Information & Communication Technologies in Education,” 2002, pp. 3–18.

[2] T. Haertel and I. Jahnke, “Kreativitätsförderung in der Hochschullehre: ein 6-Stufen-Modell für alle Fächer?!,” Zeitschrift für Hochschulentwicklung, vol. 6, pp. 238–245, 2011.

[3] D. May, C. Terkowsky, T. Haertel, and C. Pleul, “Using E-Portfolios to support experiential learning and open the use of tele-operated laboratories for mobile devices,” in 2012 9th International Conference on Remote Engineering and Virtual Instrumentation, REV 2012, 2012.

[4] C. Terkowsky and T. Haertel, “Where have all the inventors gone? Fostering creativity in engineering education with remote lab learning environments,” in IEEE Global Engineering Education Conference, EDUCON, 2013, pp. 345–351.

[5] H. G. Bruchmüller, Labordidaktik für {Hochschulen<p>Eine} Einführung zum Praxisorientierten {Projekt-Labor}. Leuchtturm Verlag, 2001.

[6] C. Terkowsky, C. Pleul, A. E. Tekkaya, and I. Jahnke, “PeTEX -- Platform for eLearning and Telemetric Experimentation,” in Praxiseinblicke -- Forschendes Lernen in den Ingenieurwissenschaften, vol. 1, 2011, pp. 28–31.

[7] L. Arnaud, M. Cécile, and P. Patrick, “Télé-TP : Premiers pas vers une modélisation,” Technol. l’Information la Commun. dans les Enseign. d’ingénieurs dans l’industrie, pp. 203–221, 2002.

[8] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic, “Augmented reality technologies, systems and applications,” Multimed. Tools Appl., vol. 51, pp. 341–377, 2011.

[9] Z. Pan, A. D. Cheok, H. Yang, J. Zhu, and J. Shi, “Virtual reality and mixed reality for virtual learning environments,” Computers & Graphics, vol. 30. pp. 20–28, 2006.

[10] E. W. T. Ngai, J. K. L. Poon, and Y. H. C. Chan, “Empirical examination of the adoption of WebCT using TAM,” Comput. Educ., vol. 48, pp. 250–267, 2007.

[11] E. A. Wehrwein, H. L. Lujan, and S. E. DiCarlo, “Gender differences in learning style preferences among undergraduate physiology students.,” Adv. Physiol. Educ., vol. 31, pp. 153–157, 2007.

[12] E. Kanninen, “Learning styles and e-learning,” 2008.

[13] P. Brusilovsky, “Adaptive hypermedia,” User Model. User-adapt. Interact., vol. 11, pp. 87–110, 2001.

[14] M. Athanasios, K. Theodoros, and B. Kostas, “Design & Development of a Dynamic Hypermedia Educational System,” J. Inf. Technol. Impact, vol. 2, no. 3, pp. 105–116, 2001.

[15] N. Henze, “Adaptive Hyperbooks: Adaptation for Project-Based Learning Resources,” Hannover, Germany, 2000.

[16] P. Brusilovsky, “Methods and techniques of adaptive hypermedia,” User Modeling and User-Adapted Interaction, vol. 6. pp. 87–129, 1996.

[17] P. Brusilovsky, “Adaptive Educational Hypermedia,” in Proceeding of Tenth International PEG Conference, 2001, pp. 8–12.

[18] P. Brusilovsky, J. Eklund, and E. Schwarz, “Web-based education for all: A tool for developing adaptive courseware,” in Computer Networks and ISDN Systems, 1998, pp. 291–300.

[19] P. Brusilovsky, E. Schwarz, and G. Weber, “ELM-ART: An Intelligent Tutoring System on World Wide Web,” in Intelligent Tutoring Systems, 1996, pp. 261–269.

[20] P. M. E. De Bra, “Teaching hypertext and hypermedia through the web,” J. Univers. Comput. Sci., vol. 2, pp. 797–804, 1996.

[21] R. M. Carro, E. Pulido, and P. Rodríguez, “TANGOW : Task-based Adaptive learNer Guidance On the WWW,” Second Work. Adapt. Syst. User Model. World Wide Web, pp. 49–57, 1999.

[22] P. De Bra and L. Calvi, “AHA! An open Adaptive Hypermedia Architecture,” New Review of Hypermedia and Multimedia, vol. 4. pp. 115–139, 1998.

[23] K. A. Papanikolaou, M. Grigoriadou, H. Kornilakis, and G. D. Magoulas, “Personalizing the interaction in a web-based educational hypermedia system: The case of INSPIRE,” User Model. User-Adapted Interact., vol. 13, pp. 213–267, 2003.

[24] E. Triantafillou, A. Pomportsis, and S. Demetriadis, “The design and the formative evaluation of an adaptive educational system based on cognitive styles,” Computers & Education, vol. 41. pp. 87–103, 2003.

[25] A. Kumar, R. Pakala, and R. K. Ragade, “The Virtual Learning Environment system,” in Frontiers in Education Conference, 1998. FIE ’98. 28th Annual, 1998, vol. 2, pp. 711–716 vol.2.

[26] E. Klopfer and K. Squire, “Environmental detectives-the development of an augmented reality platform for environmental simulations,” Educ. Technol. Res. Dev., vol. 56, pp. 203–228, 2008.

[27] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, “Mixed Reality (MR) Reality-Virtuality (RV) Continuum,” Syst. Res., vol. 2351, pp. 282–292, 1994.

[28] D. W. F. Van Krevelen and R. Poelman, “A survey of augmented reality technologies, applications and limitations,” … J. Virtual Real., vol. 9, pp. 1–20, 2010.

[29] R. Azuma and R. Azuma, “A survey of augmented reality,” Presence Teleoperators Virtual Environ., vol. 6, pp. 355–385, 1997.

[30] G. Burdea and P. Coiffet, “Virtual reality technology,” in Presence: Teleoperators & Virtual Environments, 2003, vol. 12, pp. 663–664.

[31] K. Bokyung, “Investigation on the relationships among media characteristics, presence, flow, and learning effects in augmented reality based learning,” in Multimedia and E-Content Trends: Implications for Academia, 2009, pp. 21–37.

[32] M. Dunleavy, C. Dede, and R. Mitchell, “Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning,” J. Sci. Educ. Technol., vol. 18, pp. 7–22, 2009.

[33] A. B. Craig, Understanding Augmented Reality. 2013.

[34] G. Chang, P. Morreale, and P. Medicherla, “Applications of augmented reality systems in education,” Technol. Teach. Educ., vol. 2010, pp. 1380–1385, 2010.

[35] M. Billinghurst, H. Kato, and I. Poupyrev, “The MagicBook - Moving seamlessly between reality and virtuality,” IEEE Comput. Graph. Appl., vol. 21, pp. 6–8, 2001.

[36] D. A. Kolb, “Experiential learning: Experience as the source of learning and development,” J. Organ. Behav., vol. 8, pp. 359–360, 1984.

[37] M. E. Auer and A. Pester, “Toolkit for Distributes Online-Lab Kits,” Adv. Remote Lab. e-learning Exp., vol. 6, pp. 285–296, 2007.

[38] O. Gervasi, A. Riganelli, L. Pacifici, and A. Laganà, “VMSLab-G: A virtual laboratory prototype for molecular science on the Grid,” Futur. Gener. Comput. Syst., vol. 20, pp. 717–726, 2004.

[39] H. EL Kabtane, Y. Mourdi, M. EL Adnani, and M. Sadgal, “The integration of augmented reality in the virtual learning environment for practical activities,” in Electrical and Information Technologies (ICEIT), 2015 International Conference on, 2015, pp. 363–368.

[40] A. Ben Bouna, “Vers un système d’enseignement à distance adaptatif aux styles d'apprentissage des apprenants,” Cadi Ayyad faculté des sciences Semlalia, 2012.

[41] C. Jung, “Psychological Types: The collected works of CG Jung,(vol. 6),” Princet. Univ. Press. Princeton, NJ, USA (1921/ …, 1971.

[42] AUTODESK, “3ds Max,” 2015. [Online]. Available: http://www.autodesk.fr/products/3ds-max/overview. [Accessed: 09-Sep-2015].

[43] maxon, “Cinema 4D,” 2015. [Online]. Available: http://www.maxon.net/fr/products/cinema-4d-studio.html. [Accessed: 09-Sep-2015].

[44] I-maginer, “Openspace3D,” 2015. [Online]. Available: http://www.openspace3d.com/. [Accessed: 09-Sep-2015].

[45] P. Hyungjun, J. Ho-Kyun, and P. Sang-Jin, “Tangible AR interaction based on fingertip touch using small-sized nonsquare markers,” J. Comput. Des. Eng., vol. 1, no. 4, pp. 289–297.

[46] J. Chun and S. Lee, “A Vision-based 3D Hand Interaction for Marker-based AR,” Int. J. Multimed. Ubiquitous Eng., vol. 7, no. 3, pp. 51–58.

[47] J. Chun and S. Lee, “Dynamic Manipulation of a Virtual Object in Marker-less AR system Based on Both Human Hands,” Trans. Internet Inf. Syst., vol. 4, no. 4, pp. 618–632, 2010.

Designing of Hydraulically Balanced Water Distribution Network Based on GIS and EPANET

RASOOLI Ahmadullah
Faculty of Engineering
Graduate University of The Ryukyus
Okinawa, Japan

KANG Dongshik
Information Engineering
Graduate University of The Ryukyus
Okinawa, Japan

***Abstract*—The main objectives of this paper are, designing and balancing of Water Distribution Network (WDN) based on loops hydraulically balanced method as well as using Geographical Information System (GIS) methodology with the contribution of EPANET. GIS methodology is used to ensure WDN's integrity and skeletonized a proper and functional WDN by using Network Analyst utilizing the geometric network and topology network by hierarchical geo-databases. The problem is to make WDN hydraulically balanced by applying WDN balancing method. For that reason, we have analyzed water flows in each pipe and performed the iterations process on loops, in order to make the algebraic summation of head loss"h_f" around any closed loop zero, in case, the summation of pipe flows must be equal to the flow amount entering or leaving the system through each node. At each iteration, reasonable changes occurred at pipes flow until the head loss has become very small or fixed zero as (optimizes correction) by using excel sheet solver. Since this method is confirmed to be effective, simulations were done by using GIS and EPANET water distribution platform. As a result, we accomplished hydraulically balanced WDN. Finally, we have analyzed and simulated hydraulics parameters for the targeted area in Kabul city. Thus, we successfully determined the hydraulics state of parameters around the network as a positive result. It is worth mentioning that, Hardy-cross method is being used for approaching more precise optimized correction and consequences concerning hydraulically-balanced and optimal WDN. This method can be done for complex loops WDN as well; the advantage of the method is simple math and self-correction. Managers and engineers who work in the field of water supply this methodology has been recommended as the more advantageous workflow in planning water distribution pattern.**

Keywords*—*Geographical Information System (GIS); Water Distribution Network (WDN); Hydraulics; EPANET

I. INTRODUCTION

To manage and control WDN we need to create Geo-database and knowledge-base in order to store water background data layers with features in ArcGIS and manage WDN. Therefore, GIS is comprehensive and multifunctional computer-based software being used in water transmission and distribution systems in modern and systematic water supply. However, it is the best application to manage, manipulate and maintain geospatial data and to develop and sustain asset management for today's water utilities in worldwide. Though for the targeted area there was no previous data available on water supply, no distribution lines, and service connection information as well as with no service population and sewerage system network the entire situation is unmapped. We have produced three hierarchical Geo-databases separately [1]. The Geo-database structures indicate main, geometric network and topology Geo-databases consisted of feature data sets. Water supply background data (vector data and raster data) collected from various source that working for shoulder to shoulder for Kabul city water supply extension. We designed a proper WDN created in GIS then imported to EPANET to be analyzed and simulated in order to approach the objectives and successful consequences.

The commonly network has been contained of physical and non-physical components and features such as pipes, nodes and reservoirs with pumps and valves-types and non-physical describes the behavior and operational aspects of a distribution system. Since, GIS project scenarios imported to EPANET in (.inp) format in order to carry out simulation and find various WDN's parameters state.

In this paper, we have considered two closed-loops of the network using two fundamental hydraulic principles such as continuity and energy conservation equations. The statement of this valuable method is first the sum of pipe flows into and out of a node equals the flow entering or leaving the system through each node applied to all pipes. And second, the algebraic sum of pressure drops around a closed loop must be zero we applied it to all the nodes. Some of the given parameters include water demand at each node, diameter of pipe and pipe length as well as pressure at the first node and pipe roughness. We need to find corrected water flows at pipes, this happens by reducing head loss around loops, finally we will also get the pressure at all nodes.

II. STUDY SITE

Water supply conditions in Kabul city are serious, and water availability will be the most critical constraint to the development of the city. There are lack and shortages of water not only for irrigation but also for domestic; the entire current water supply in Kabul depends exclusively on local ground-water resources. However, surface water transferring is needed to be extended and developed from Shatoot dam located on Maidan River and also Gulbahar dam located on Panjshir river through a proper treatment plant process in order to supply potable water to city's population. However, water supply is being extended and developed in the last decay in order to meet present-day population demand; water consumption is increased due to the rise of population and economic development in the city. Therefore, the local water resources are not sufficient for current demand to cover the whole users. Local groundwater needs to be recharged naturally or artificially.

Population prediction is one of the necessary factors for designing water supply systems. Therefore, population should be estimated precisely to continuously supply increasing water demand for the community. Population projection and growth rate in percentage from 2002 until 2032 is shown in "Fig. 1". The estimated potential of Kabul groundwater is approximately 44 million (m^3) per year according to the current water study. The estimated groundwater potentials are as presented in "TABLE1".

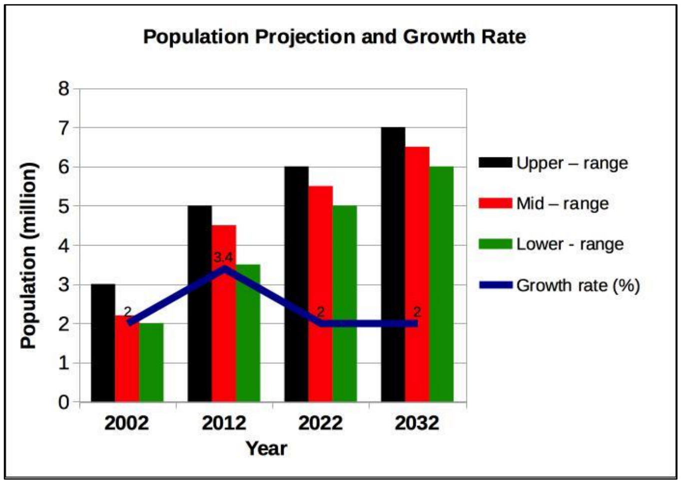


Fig. 1. Population projection and growth rate in percentage for Kabul city based on upper, middle and lower range of statistics data

TABLE I. AVAILABLE POTENTIAL OF GROUND-WATER OF KABUL CITY

Aquifers' name	Water resources availability
Logar	24.64
Allaudin and upper Kabul	12.48
Afshar	3.65
Lower Kabul	3.65
Total	**44.42**

So that this is the targeted WDN located in eastern part of the city, called (Khoshal Khan Mena) district five on the left side of "Fig. 2". And here is the reference regarding Kabul city water supply system that has evaluated since 1992-94 civil war [1].

III. METHODOLOGY

GIS is the best application as knowledge-base and spatial database to manage, manipulate and analyze geospatial data to develop and sustain asset management for today's water utilities. Geographical Information System can be used as a key tool for making WDN. The process of creating a systematic and functioning water distribution network has been addressed in our paper under the topic of Designing an Optimal Water Distribution Network Using GIS and EPANET, Kabul city [2].

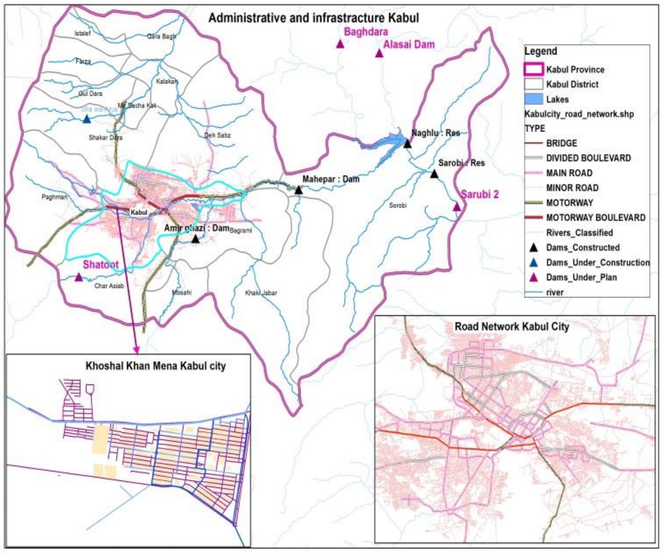


Fig. 2. Infrastructure dam sites of Kabul province and city road network

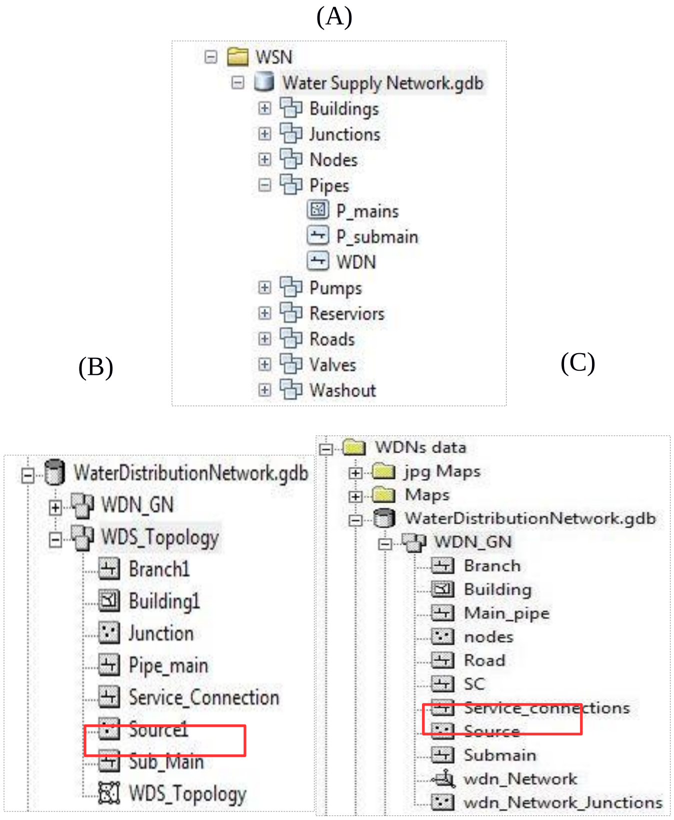
A. Geo-database Structures of WDN in ArcGIS

In this step, an actual Geo-database structure has been created contained necessary input data such as background data as raster and vector (e.g. both existing and future urban planning with public utility pipelines, population, buildings, topographical survey, edges and junction points etc. For making an optimal WDN the Geo-database model is a generic model for geographic features and attributes that support a wide variety of object relations and behavior. Improvement, manage, and control of water supply systems (WSS) is essential trough Geo-database and knowledge-base ArcGIS-based utilizing necessary data on water supply even for other utilities, these includes water supply background data (e.g. urban development plan, buildings and stellate image of the community), population, roads and pipes (e.g. water mains, sub-mains, and branches). With other WSN parameters which are required for WSN, it can be seen bellow in the hierarchical Geo-databases' structures that have been produced for Kabul city water supply as shown in "Fig. 3".

Regarding the bellow hierarchical Geo-database structures kindly refer to our journal paper, Geo-database and Knowledge Base in Arc GIS are organized and contains the required data (raster and vector) as a national level for Kabul basin [3].

B. Input Data

Input data contains satellite images that need to be extracted and vector data such as Infrastructure information, existing and planned public utility pipelines. With diameter and length, customer's information with coordinates, population, edges and nodes points [4]. And other necessary WDN's background related data for a correct affordable improvement control and maintain of the system.



- WSN
  - Water Supply Network.gdb
    - Buildings
    - Junctions
    - Nodes
    - Pipes
      - P_mains
      - P_submain
      - WDN
    - Pumps
    - Reserviors
    - Roads
    - Valves
    - Washout

- WaterDistributionNetwork.gdb
  - WDN_GN
  - WDS_Topology
    - Branch1
    - Building1
    - Junction
    - Pipe_main
    - Service_Connection
    - Source1
    - Sub_Main
    - WDS_Topology

- WDNs data
  - jpg Maps
  - Maps
  - WaterDistributionNetwork.gdb
    - WDN_GN
      - Branch
      - Building
      - Main_pipe
      - nodes
      - Road
      - SC
      - Service_connections
      - Source
      - Submain
      - wdn_Network
      - wdn_Network_Junctions

Fig. 3. Prepared Geo-database structures for Kabul city water Distribution network. GDB-main, GDB-Topology and GDB-GN for WDN respectively

C. Network Analysis

a) Topology Analysis

Topology Geo-database has been created from existing features due to clarify edges and nodes layers connectivity by topology analysis in order to make ensure concerning data integrity and containment with its adjacency and coincidence. An example of topology analysis is shown in the "Fig. 4".

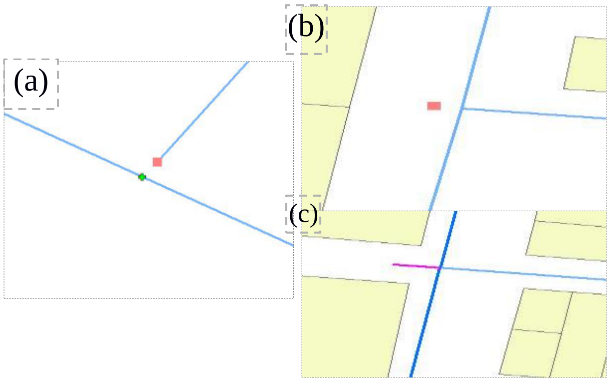


Fig. 4. An example of topology analysis validated data and fixed errors, figures represent extend, trim and the node need to be replaced

Topology in ArcGIS fixes errors in data and it is the process to describe and maintain special relationships of map features.

b) Geometric Network

This utility network can also be used for other public utilities and services such as electrical, gas pipe lines, sewer and storm networks, telecommunication transmission and distribution networks either for loops and branches systems. These components can be modeled and analyzed by the help of network analyzes. But need a sophisticated GIS application to analyze and model an optimal WDN. Basically, GN is performed in Arc-Catalog tree. Once a geometric network is modeled, it is possible to benefit from performing various network analyses. For instance, in "Fig. 5", (a) we found the shortest path between (P1) and (P2), while (b) represents downstream of it's related node.

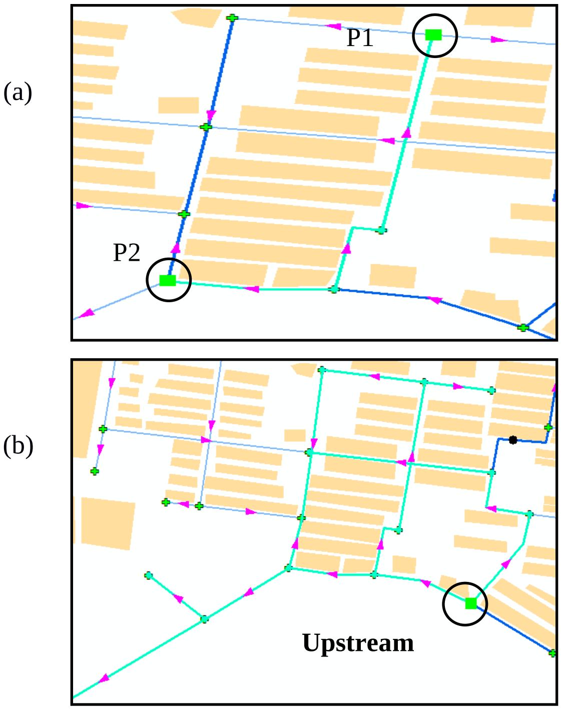


Fig. 5. (a) Defines an example of geometric network short path between (P1) and (P2), while (b) represents downstream of related node

c) Distribution Network Skeletonization

GIS application provides functions for development and preparation of accurate spatial digital information as input into data for the network design optimization model, which included network layout, connectivity, pipe characteristics and cost, pressure gradients, demand patterns, cost analysis, network routing and allocation, and effective color graphic display of results [5].

The following example of network layout was skeletonized in order to cover a part of western Kabul city in shown in "Fig. 6".

Attribute table of the network elements are consisted of junctions and pipes. An example is shown in the "TABLE2 and TABLE3".

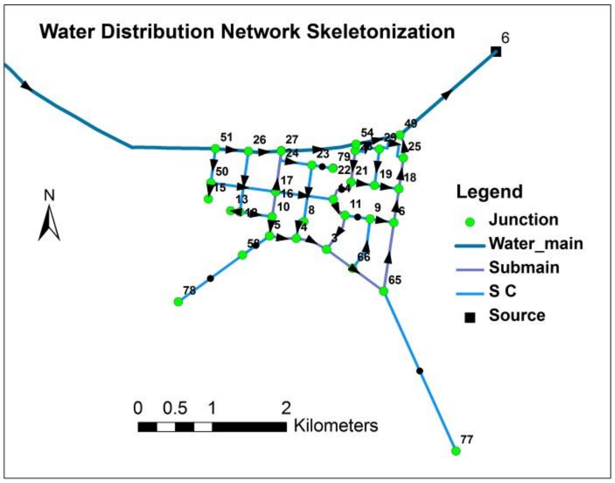
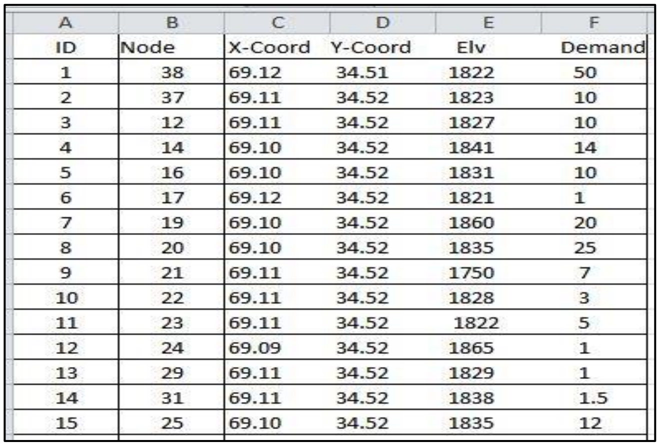


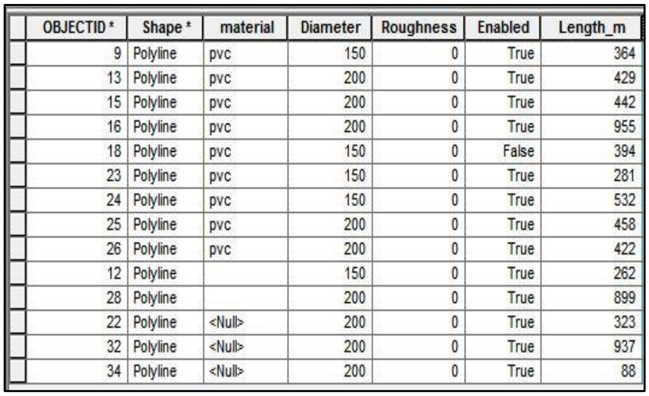
Fig. 6. It is illustrating water distribution network flow paths and WDN Skeletonization

TABLE II. ATTRIBUTE TABLE OF JUNCTION



A	B	C	D	E	F
ID	Node	X-Coord	Y-Coord	Elv	Demand
1	38	69.12	34.51	1822	50
2	37	69.11	34.52	1823	10
3	12	69.11	34.52	1827	10
4	14	69.10	34.52	1841	14
5	16	69.10	34.52	1831	10
6	17	69.12	34.52	1821	1
7	19	69.10	34.52	1860	20
8	20	69.10	34.52	1835	25
9	21	69.11	34.52	1750	7
10	22	69.11	34.52	1828	3
11	23	69.11	34.52	1822	5
12	24	69.09	34.52	1865	1
13	29	69.11	34.52	1829	1
14	31	69.11	34.52	1838	1.5
15	25	69.10	34.52	1835	12

TABLE III. ATTRIBUTE TABLE OF PIPES



OBJECTID *	Shape *	material	Diameter	Roughness	Enabled	Length_m
9	Polyline	pvc	150	0	True	364
13	Polyline	pvc	200	0	True	429
15	Polyline	pvc	200	0	True	442
16	Polyline	pvc	200	0	True	955
18	Polyline	pvc	150	0	False	394
23	Polyline	pvc	150	0	True	281
24	Polyline	pvc	150	0	True	532
25	Polyline	pvc	200	0	True	458
26	Polyline	pvc	200	0	True	422
12	Polyline		150	0	True	262
28	Polyline		200	0	True	899
22	Polyline	<Null>	200	0	True	323
32	Polyline	<Null>	200	0	True	937
34	Polyline	<Null>	200	0	True	88

d) Set of Geo-Database Model in GIS

The model illustrates, the methodology of creating WDN through GIS software, it is a set of Geo-database model objects with relations and behaviors for WDN made of object classes, layers includes (polygons, lines and points) as shown in “Fig. 7” .

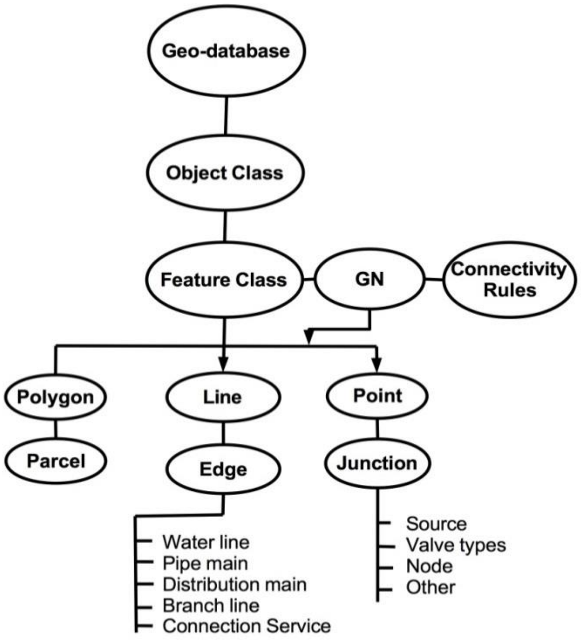


Fig. 7. A set of Geo-database model objects with relations and behaviors for water distribution network through GIS

IV. HYDRAULICS ANALYSIS

A. Statements of the Method

- Continuity Formula

The sum of pipe flows amount into and out of a node equals to the flow amount entering or leaving the system through each node. Therefore, we have had the following equation:

$$Q = Q_1 + Q_2 \qquad (1)$$

Where, Q = Total inflow, $Q_1 + Q_2$= Total outflow

- Energy Conservation Formula

The algebraic Summation of head loss “h_f” around any closed loop is zero. Therefore,

$$\sum h_f(Loop) = 0 \rightarrow \sum K(Q + \Delta Q)^n = 0 \qquad (2)$$

Where,

Q= Actual inflow, ΔQ= Correction

K= Head loss coefficient, n= Flow exponent.

The general relationship must be maintained between discharges and head-losses for each pipe in loops by the following formula:

$$h_f = K.Q^n \qquad (3)$$

These pressure heads loss equations have commonly been using by EPANET hydraulic solutions.

- *Darcy-Weisbach Equation*

$$K = \frac{8\text{fL}}{g\Pi^2 D^5}, \text{ n= 2} \qquad (4)$$

- *Exponential friction Equation (Hazen-William)*

$$K = \frac{10.67}{C^{1.85} D^{4.87}}, \text{ n= 1.85} \qquad (5)$$

$$\Delta Q = \frac{-\sum h}{2\sum h/Q} \qquad (6)$$

Where,
C= Equivalent resistance,
D= Internal pipe diameter
L= Pipe length,
g= Gravitational acceleration, f= Friction factor.

V. Method Procedure and Approach

D. Divided the network into loops (e.g. loop1, loop2, etc)

E. For each loops done the fallowing steps

1) Assumed flow, flow direction in pipes, flow direction for loops whether positive (clockwise) or negative (counter-clockwise) applying continuity equation at each node. Estimated pipe flows are connected with iteration until head losses in the clockwise direction are equal to the counter-clockwise direction in each loop.

2) Need to calculate equivalent resistance "K” for each pipe based on given parameters as demand at each node, pipe diameter and pipe length, temperature with pipe material.

3) Calculate$h_f = K.Q^n$ for each pipe. Retain sign from “(1)” and compute sum for loops “h_f’.

4) Compute ❘ h_f/Q ❘ for each pipe and sum for each loop$\sum$ ❘ h_f/Q ❘.

5) Calculate correction by the fallowing formula

$$\Delta Q = -\sum h_f / (n \sum \mid h_f/Q \mid).$$

6) Applying correction to Q_{new}= Q+ΔQ

7) Repeat step” (3) to (6)” until Δ become very small.

8) Ultimately solve for pressure at each node using energy method.

The above method is known as Hardy-cross method in hydraulics and this method is applicable to closed loop systems [6]. Hardy-Cross (1885-1951), who was a professor of civil engineering at the university of Illinois, Urbana-Champaign, presented in 1936 a method for the analysis of looped pipe network with specified inflow and outflows (fair et al., 1981) [7], [8].

We have made the effort to solve the problems addressed in the “Fig. 8”.

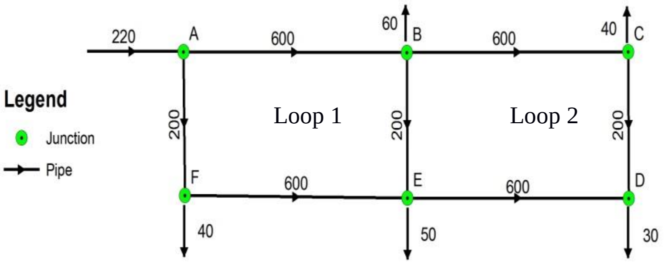


Fig. 8. Two closed-loops WDN contains of pipes (p) and nodes or junction (ju) with given pipes length, estimated flow and demand at each node as well as water flows and water flows direction at each pipe

Pipe roughness size equals to 0.06mm and pressure heads elevation at point (A) is given 70m other characteristics of the network are shown in as followings“Table4”.

TABLE IV. Pipe Length in Meter and Diameter in Mm

Pipe	AB	BC	CD	DE	EF	AF	BE
Length (m)	600	600	200	600	600	200	200
Diameter (mm)	250	150	100	150	150	150	100

In the next table elevation of each pipe node has been estimated as followings shown in “Table5”.

TABLE V. Pipe Nodes Elevation in Meter

Nodes	A	B	C	D	E	F
Elevation (m)	30	25	20	20	22	25
Demand	0	60	40	30	50	40

Therefore, based on the above-given hydraulics parameters we are going to determine unknown parameters in the network such as optimized flow in pipes, head losses applying energy equation and finding the unknown pressure at nodes as well as making loops hydraulically balanced. Thus, performing Iterations process for optimizes correction as followings.

VI. Iterations Process

Computing f, k and h (m), h/Q for each pipe and then finding summation of h(m), h/Q around each loop shown in “Table6”.

To know corrected flow (Q_{new}) for each pipe around loops need to calculate Q (L/s) + ΔQ. The correction can be found as follows for instance for we found for the first loop:

$$\Delta Q = \frac{-\sum h}{2\sum h/Q} = \frac{-(-33.91)}{2\text{x}1191.82} = 0.0143 = 14.23\text{L}/s.$$

TABLE VI. ITERATIONS PROCESS

	pipe	Q (L/s) (Q1+Q2)	Re(x 105)	f	K	hf(m)	hf/Q (m/m3/s)
1st iteration	AB	120	5.41	0.0157	797	11.48	95.64
	BE	10	1.31	0.0205	33877	3.39	338.77
	EF	-60	4.51	0.0172	11229.1	-40.42	673.75
	FA	-100	5.63	0.0162	336.6	-8.36	83.66
						Σ -33.91	1191.82
	BC	50	3.76	0.0174	11359.7	28.4	567.98
	CD	10	1.13	0.0205	33877	3.39	338.77
	DE	-20	1.5	0.0189	12338.9	-4.94	246.78
	EB	-24.23	2.73	0.0189	31232.9	-18.34	756.77
						Σ -8.51	Σ 1910.3
2nd iteration	AB	134.23	5.41	0.0156	791.9	14.27	106.3
	BE	26.46	1.13	0.0188	31067.7	21.75	822.05
	EF	-45.77	4.51	0.0175	11424.9	-23.93	522.92
	FA	-85.77	5.63	0.0164	846.9	-6.23	72.64
						Σ 5.86	Σ 1523.92
	BC	47.77	3.76	0.0174	11359.7	25.93	542.7
	CD	7.77	1.13	0.0205	33877	2.05	263.3
	DE	-22.23	1.5	0.0189	12338.9	-6.1	274.3
	EB	-24.54	2.73	0.0189	31232.9	-18.81	766.5
						Σ 3.07	Σ 1846.7
3rd iteration	AB	132.31	5.41	0.0157	797	13.95	105.45
	BE	25.37	1.31	0.0205	33877	21.8	859.46
	EF	-47.69	4.51	0.0172	11229.1	-25.54	535.53
	FA	-87.69	5.63	0.0162	836.6	-6.43	73.36
						Σ 3.78	Σ 1573.81
	BC	47.77	3.76	0.0174	11359.7	25.93	542.68
	CD	7.77	1.13	0.0205	33877	2.05	263.31
	DE	-22.23	1.5	0.0189	12338.9	-6.1	274.26
	EB	-23.34	2.73	0.0189	31232.9	-17.01	745.29
						Σ 4.09	Σ 1825.5

The corrected flows obtained after iterations process and have been shown in the “Table7”. f is not calculated in some iteration so that K is similar, for more precision it can be changed. Therefore, now it is possible to find unknown pressure heads at nodes though the final values of head losses like in “Table8”.

TABLE VII. THE CORRECTED FLOWS

	pipe	Q (L/s)	Corrected Q (L/s)
1st iteration	AB	120	134.23
	BE	10	24.23
	EF	-60	-45.77
	FA	-100	-85.77
	BC	50	47.77
	CD	10	7.77
	DE	-20	-22.23
	EB	-24.23	-26.46
2nd iteration	AB	134.23	132.31
	BE	26.46	24.54
	EF	-45.77	-47.69
	FA	-85.77	-87.69
	BC	47.77	46.94
	CD	7.77	6.94
	DE	-22.23	-23.06
	EB	-24.54	-25.37
3rd iteration	AB	132.31	131.11
	BE	25.37	23.34
	EF	-47.69	-48.89
	FA	-87.69	-88.89
	BC	47.77	46.5
	CD	7.77	6.5
	DE	-22.23	-23.5
	EB	-23.34	-24.61

TABLE VIII. PRESSURE HEADS AT EACH NODE IN (M)

Node	**Pressure heads in mater**
A	70-30=40
B	70-25-13.7=31.3
C	70-20-24.74-13.7=11.56
D	70-20-13.7-24.74-1.52=10.04
E	70-22-6.59-26.67=14.74
F	70-25-6.59=38.41

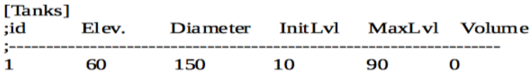
The result also compared in EPANET software that carried out in 10 trails with no error, the simulation is shown above. Therefore, we propose EPANET for analyzing a big and complex WDN. By using EPANET we are able to analyze WDN precisely and avoid time-consuming.

VII. EPANET ANALYSIS

This paper moreover presents analysis through EPANET too for designing an optimal and hydraulically balanced water distribution network. EPANET is a free Windows computer program developed by the U.S. Environmental Protection Agency (EPA). EPANET performs simulations of hydraulic and water quality behavior within pressurized pipe networks, such as a city water supply system. A network can consist of pipes, pipe junctions, pumps, valves, storage tanks, and reservoirs [9]. For our case in two closed loops pipe network we put necessary basic description of the network placed in a simple text file format so I address how to import a network to EPANET, to import a text file it must be contained in a list of node ID's with their coordinates as well as a list of link ID's and their connecting nodes.

Note that only junctions and pipes are represented. Other network elements, such as reservoirs and pumps, can either be imported as junctions or pipes and converted later on or simply be added in later on. The user is responsible for transferring any data generated from a CAD or GIS package into a text file with the format shown below [9]. EPANET tracks the flow of water in each pipe, the pressure in each node, and the height of water in each tank or reservoir during a simulation period consisting of multiple time steps. Consequently, how to organize project's scenario looks like bellow:

```
[Junctions]
;id            Elev.   Demand
;--------------------------------
1       30          0
2       25          60
3       20          40
4       20          30
5       22          50
6       25          40
[Coordinates]
;id    x-coord           Y-coord
;---------------------------------------------
1     -11264937.393   5371052.628
2     -11264472.173   5371051.188
3     -11264011.274   5371052.628
4     -11264009.834   5370817.857
5     -11264470.747   5370818.241
6     -11264934.549   5370822.302
[Pipes]
;id     Nodel1  Nodel2  Length  Diam   Roughness
;--------------------------------------------------------------
1       1        2       600     250    0.006
2       2        3       600     150    0.006
3       3        4       200     100    0.006
4       1        6       200     150    0.006
5       6        5       600     150    0.006
6       2        5       200     100    0.006
7       5        4       600     100    0.006
```



```
[Tanks]
;id        Elev.      Diameter    InitLvl    MaxLvl    Volume
;---------------------------------------------------------------------
1          60         150         10         90        0
```

In addition, other network elements, like pumps and reservoirs can either be imported as junctions or pipes. The principle of EPANET network analysis is based on the continuity equation and energy conservation theory. The purpose of the bellow network is to supply water at adequate pressure and flow. The following two closed network illustrates assumed flows at each pipe and demand at nodes before balancing loops shown in “Fig. 9”.

The result of analyzed and obtained by EPANET platform effectively as well as the status of the hydraulics parameters has been determined and shows the network is in a very good condition. Hydraulic status is balanced after 10 trails as as shown in“Fig. 10”.

EPANET can use any one of the three popular forms of the head-loss formula the Hazen-Williams formula, the Darcy-Weisbach formula mentioned above, or the Chezy-Manning formula.

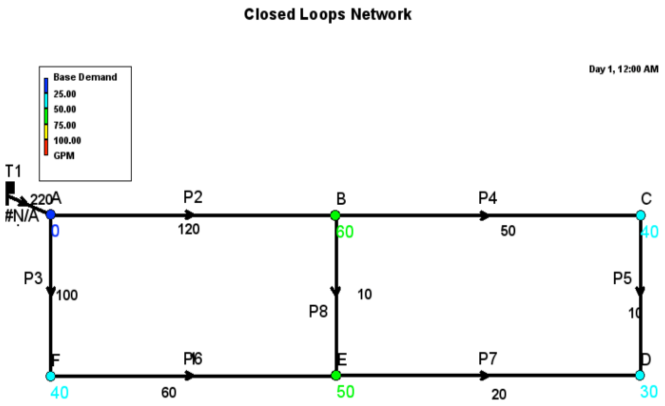


Fig. 9. Before hydraulic analysis

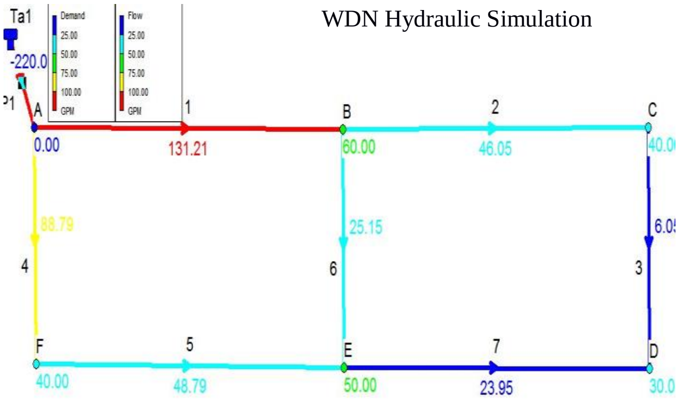


Fig. 10. After analysis and optimizes correction of balancing loops through EPANET platform

In order to, analyze EPANET needs the input files such as nodes and pipe description as input data, see [10, 11, 12] for the input and output data. Determined the hydraulics parameters of the targeted area by the use of EPANET, so that the simulation is shown in “Fig. 11”.

Finally, the flow chart of the method relations is shown in “Fig. 12”. This will help readers to understand the methodology easily.

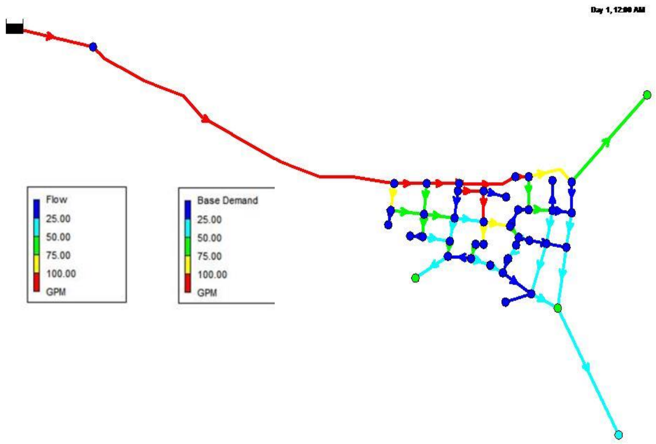


Fig. 11. Water Distribution Network of the targeted area

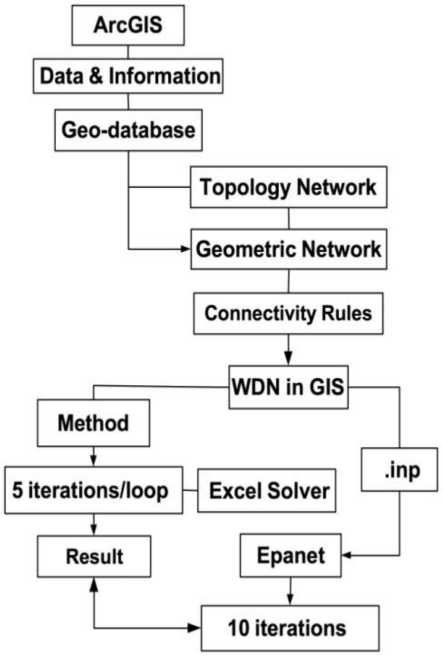


Fig. 12. Flow chart of the method

VIII. Conclusion

To conclude, the implications of applying this work are to create WDN in GIS by performing GIS with the cooperation of EPANET methodologies. In addition, analyzed loops network according to continuity and energy conservation formulas due to the determination of unknown discharges, flows and pressure at nodes. Hardy-Cross method is used due to solve and analyze closed loops network for flow continuity and head-loss in order to balance the network as the context of this paper. The result of the balanced method was effective finally compared to EPANET hydraulic status simulation which was precise. Ultimately accomplished proper hydraulically balanced loops WDN by performing several iterations empirically to find the corrected flows around the network until head loss around each loop became zero. Hence, as a context these principles have been applied to each pipe and each node in closed-loops pipe network until we have gotten corrected flows and successful result and concisely found the required parameters such as the pressure of nodes, discharges, and water flows and flows direction in the network. This solution technique for loops balancing can be optimized correction and hydraulically balanced method as well.

IX. Future Work

In the practical side, we need to implement it in Afghanistan specially in Kabul city by the support of Ministry of Energy and Water (MEW). In the theoretical side, we will extend the optimization algorithm through a comprehensive research for the WDN. This work hopes to be a good step toward further understanding this important issue.

Acknowledgment

This work presents a part of a Master's Degree research work carried out by Rasooli Ahmadullah under the supervision of Professor Kang Dongshik.

I would like to thank Japan International Cooperation Agency (JICA), PEACE Project, the Project for the Promotion and Enhancement of the Afghan Capacity for Effective Development (PEACE), for letting me fulfill my dream of being a student at University of The Ryukyus and providing me the necessary fund for the paper. Also, my grateful thanks are extended to Engineering Faculty staff, University of The Ryukyus, for their cooperation during my academic life.

Finally, I am thankful to my family for their kind support, love, and encouragement. I dedicate this work to them.

References

[1] P.G. Nembrini,1 P. Jansen,2 J.F. Pinera,2 O. Bernard,2 R. Luff,3 M. Weber,4 and M.J. Elliot4, Kabul Water Supply October 2002, Occasional paper No. 7. Evolution since the 1992-94 Civil War.

[2] RASOOLI Ahmadullah and KANG Dongshik. 'Designing an Optimal Water Distribution Network Using GIS and EPANET, Kabul city'. The Institute of Electronics, Information and Communication Engineers (IEICE) September 2015 Sendai, Japan.

[3] RASOOLI Ahmadullah, KANG Dongshik, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015. UK, London.

[4] Martínez-Solano, F. Javier; Pérez-García, R. & Iglesias-Rey, P.L, VALENCIA. Water Mains Creation Using GIS. 5th AGILE Conference on Geographical Information Science, Palma (Balearic Islands, Spain) April 25th- 27th 2002.

[5] John W. Labadie Margaret T. Herzog Civil Engineer / GIS Coordinator. Optimal Design of Water Distribution Networks with GIS. Dept. of Civil Engineering Colorado State University, Fort Collins, Colorado 80523-1372.

[6] A. RASOOLI Ahmadullah, B. KANG Dongshik, Construction of hydraulically balanced water distribution network. ICIIBMS, OIST. Journal of Information and Communication Engineering (JICE), 1 (1): 26-29, 2015 ISSN 2186-9162, Applied Science and Computer Science Publications.

[7] A Saminu, Abubakar, Nasiru, L Sagir, Design of NDA Water Distribution Network Using EPANET. International Journal of Emerging Science and Engineering (IJESE) ISSN, 2319–6378, Volume-1, Issue-9, July 2013.

[8] Prabhata K. Swamee Ashok K. Sharma, Design of water supply pipe networks. 2008 John Wiley & Sons, Inc.

[9] Lewis A. Rossman, The Epanet 2 Users Manual. United state environmental protection agency. National Risk Management Research Laboratory Cincinnati, OH 45268.

[10] Sahita I Waikhom, Darshan J, Mehta Optimization of Limbayat Zone Water Distribution System Using EPANET. International Research Journal of Engineering and Technology (IRJET) India, Volume: 02 Issue: 04 | July-2015.

[11] Terry Henshaw and Ify L. Nwaogazie. Civil Engineering. Improving water distribution network performance: A comparative analysis, PENCIL Publication of Physical Sciences and Engineering, Nigeria. Vol. 1(2):21-33.

[12] Darshan Mehta, Krunal Lakhani, Divy Patel, Govind Patel, Study of water distribution network using Epanet. International Journal of Advanced Research Engineering , Science & Management (IJARESM).

Corporate Responsibility in Combating Online Misinformation

Fadi Safieddine
Department of Management
Information Systems
College of Business
American University of the Middle
East
Egaila, Kuwait

Wassim Masri
Department of Technology
College of Engineering and
Technology
American University of the Middle
East Egaila, Kuwait

Pardis Pourghomi
Department of Technology
College of Engineering and
Technology
American University of the Middle
East Egaila, Kuwait

***Abstract*—In the age of mass information and misinformation, the corporate duty of developers of browsers, social media, and search engines are falling short of the minimum standards of responsibility. The tools and technologies are already available to combat misinformation online but the desire to integrate these tools has not taken enough priority to warrant action. This paper presents an effective and practical method based on technologies already available that could be used for browsers and social media websites that would help combat misinformation presented in the form of photo evidence, video evidence, or textual evidence the authors have termed as the "Right-click Authenticate" every browser and social media website should have.**

Keywords—Misinformation; Social Media; Browsers; Search Engines; Corporate; Ethical; Responsibility

I. Introduction

When it comes to authenticating information on social media, the responsibility seems to fall on the web users. An oligopoly of browsers, social media, and news agencies under the banner of freedom of expression have left users with a choice between being passive viewers allowed to either like, comment, or having to undergo long process of online searching in order to authenticate that piece of information. This apparently ethical approach to online freedom of expression seems to have backfired. It has allowed totally unprofessional content; developers bombard predominantly passive web content consumers with news, facts, and stories that cannot be easily challenged. Web users, and specifically online social media users, gradually filter pages, news agencies, or even friends whom they disagree with their political, theological, and/or ethical predispositions. This has become a hotbed for conspiracy theories who are capitalizing on creating parallel realities to an audience unable to verify what is presented to them. Blocked away from the alternative facts or views, these web users become easy plucking to creating alterative realities that are alien from the truth.

Ethical and social responsibilities of browsers and social media dictate that they make it possible for any individual to authenticate the validity of the information presented.

Our goal is to provide a concept for minimizing the spread of misinformation across social media that is technically feasible and is accepted by all parties involved. The purpose of this paper is thus to provide a study describing the impact of misinformation on societies as well as lesson's learned from the ways in which Wikipedia manages its contents' validity. This is followed by the proposal of a solution, which the authors believe would introduce a simple and effective approach towards authenticity validation of information in social media. In so doing, we focus on the following research question:

In what forms is misinformation being shared online and how can browsers and social media websites combat the spread of them?

To this end, the paper will provide an evaluation of contents' validation in the world of online information sharing and study the relative weaknesses of the current situation in order to propose our approach.

The forms of misinformation that have been identified in this paper are three:

1) Fake, edited, or misrepresented images: This happens when an image is presented as a fact to event, place or story that is untrue or inaccurate.

2) Fake, edited, or misrepresented videos: This happens when a video is presented as a fact to event, place or story that is untrue or inaccurate.

3) Fake, edited, or misrepresented texts: This happens when a story, quote, or news is presented as facts of events or places that are untrue or inaccurate.

These forms of misinformation can appear as political, medical, scientific, theological, race, gender, and others. Some of this misinformation can be dangerous and can have serious consequences. Misleading news in particular is dangerous as it can be used to instigate hatred, racism, theological intolerance, DIY medical treatments, and crime.

The problem with misinformation is that it spreads fast; it has short-term effect that can have long-term consequences. This paper proposes a method that would allow online users quickly, in as little as one click, to check the authenticity of the information. A one-click check can provide the means to discredit the source before it spreads or at the least make the job of unprofessional content developer much harder.

II. LITERATURE REVIEW

Social media cannot be underestimated as a major source of information for the masses. Online social networks have many benefits as a medium for fast, widespread information dissemination. They provide fast access to large scale news data, sometimes even before the mass media as in the case of the announcement of death of Michael Jackson [1]. They also serve as a medium to collectively achieve a social goal. While the ease of information propagation in social networks can be very beneficial, it can also have disruptive effects.

A. Text as form of misinformation

The research of the literature suggests two types of textual misinformation. One type observed during the shootings at Fort Hood, Texas, when a soldier inside the base sent out messages via Twitter as the event unfolded. Incorrect reports of multiple shooters and shooting locations quickly spread through the social network and reached the mass media where it was reported on television broadcasts [2]. This type could be best termed as breaking news and this type of misinformation would be extremely hard to validate as the events play out. Another example of this type is the spread of misinformation on swine flu on Twitter [3]. The spread of misinformation in this case reached a very large scale causing panic in the population. In the wake of the devastating attacks in Paris that left at least 129 people dead and hundreds more injured, social media sites like Twitter, Facebook and Instagram were flooded with updates on the coordinated attack. However, not all of that information was correct. More than 10.7 million tweets were posted about Paris between Friday and Saturday, NBC news reported at least 9 posts that contained misinformation [4].

The other type of textual misinformation shared online is in the form of reports and reflection on news. The difference noted here has to do with the time period which allows some element of reflection and spreading of rumors. This represents hotbed for conspiracy theorist. For example, following the disappearance of Malaysia Airlines flight MH370, NBC news highlighted various false reports spreading on social media which alleged that the plane had made a safe landing [5].

Social media sites can be a more convenient way for people to source their information but it has been proven many times as inaccurate [6].

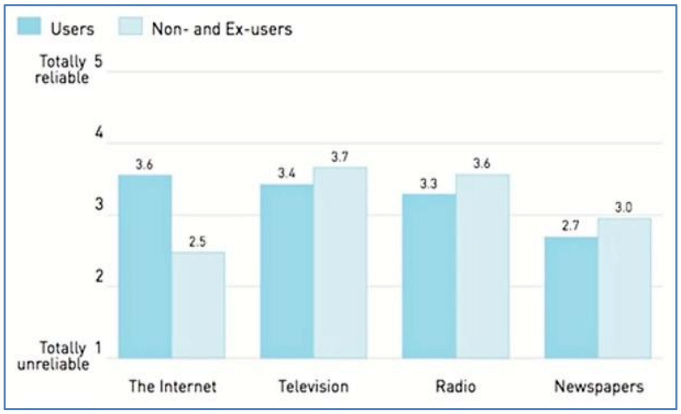


Fig. 1. Reliability of Information by Internet Users and Non-Users [3]

While the WEF data [7] showed the rapid spread of false information as a key trend for 2014, the 2013 Oxford Internet Survey [8] found that trust in the reliability of online information among British internet users has changed very little in the past 10 years. More worrying is that the same survey found a trend where web users identified the Internet as the most reliable source of information over television and radio (with a score of 3.6 on average, where 1 is unreliable and 5 is totally reliable) [5].

B. Images as form of misinformation

Various Web-based sharing and community services such as Facebook, Flickr, and YouTube have made a vast and rapidly growing amount of multimedia content available online. Uploaded by individual participants, these pools of content includes varied types of images accompanied by details such as source of image, any editing tools used, date creation, and/or descriptive textual information [9-10]. There is potential in this information that can help confirm the reality of the images users find online. Below, describes a few recent examples of image meta-data misuse on social media. It is not always possible for social media users to make sure that they are looking at an official organization’s page when viewing online images describing a story or news. One suggestion on how to deal with unverified information on social mean suggests that users should simply be aware of that when sharing the image with anyone, and weigh the verification of the unconfirmed information into their decision making [11].

1) Examples of misinformation by means of images

Misuse of images as a form of evidence can have more profound effect in getting misinformation believable. This form of misinformation has existed far before the arrival of the Internet.

Just as propaganda was not born alongside the Internet, images have been edited well before the advent of Twitter and Photoshop. As an example, the Soviet Union regularly erased disgraced political leaders from photos, even if the results appear woefully amateurish compared to the advanced photo-editing techniques [12].

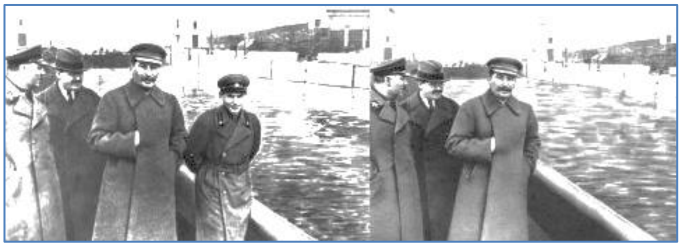


Fig. 2. Nicolai Yezhov and former Soviet leader Joseph Stalin during the 1930s [6]

Technology has advanced significantly since and with it the means for content development. Disseminating false information with image evidence has become much easier for unprofessional content developers with the likes of image editing tools.

A Canadian Sikh's bathroom selfie has gone viral after someone photoshopped the image and posted it on social media claiming that he was one of the terrorists behind the attacks in Paris [4].

The problem when images are used as form of misinformation is that they become harder to discredit. In this paper we shall look at a particular example of “Giant Skeleton Unearthed” [13].

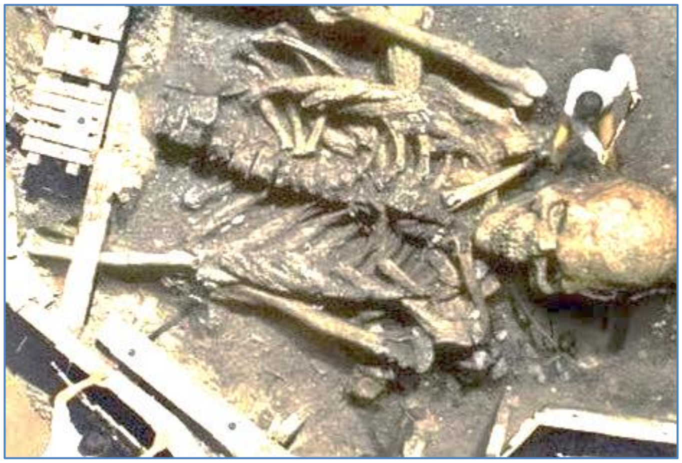


Fig. 3. Giant Skeleton discovery [13]

Initially reported in 2004 by unprofessional bloggers in India it reached the news media. The Voice newspaper editor, who was first to report it, has since posted a retraction and an apology [13], however the image refused to go away and is still being shared almost 12 years later [14][15][16]. In fact, when searching for matches of this image and where it is being shared, the team unearthed over 800 occurrences. The image which was initially reported as a discovery in India linked to Mahabharata, a Hindu epic story from 200 B.C, has been recycled multiple times and always rebranded with new story of discovery in Portugal, El Salvador, Malaysia, South Africa, Greece, Dominican Republic, Egypt, Kenya, and Saudi Arabia. A key success factor in the wide spreading of misinformation via image is that the users ‘wanting to believe’ the image because it would confirm something they want to believe in which could be political, scientific, health (cures), prejudicial, or religious belief. This represents the framework that has the appearance of credibility.

2) *Image meta-data*

As digital photography becomes more prevalent, the number of digital images that are stored on photo-sharing sites is increasing dramatically, and the number of images will make it increasingly difficult to authenticate what users are looking at. Image Metadata provides some vital information that could only be available from the original image or subsequent editing of that image. Image metadata is sometimes copied by online photo-sharing sites and made available for views [17]. Image metadata could also assist in search, and is useful in retrieving desired images from a large collection of images [18]. For instance an image meta-data file can be a XML file containing meta-data from Flickr or other social media tools for all the retrieved photos (e.g. photo title, photo description, photo id, tags, Creative Common license type, number of posted comments, the URL link of the photo location, the photo owner’s name, user id, the number of times the photo has been displayed, etc.) [19]. However social media websites have been accused of removing metadata information from images as a part of the social media processing of images, and thus rendering the job of image forensics harder at validating images [20].

In 2013, the International Press Telecommunications Council (IPTC) published a study in the British Journal of Photography into the use of images by social media websites, finding that some of the most predominant ones, such as Facebook, Twitter and even Flickr, remove photographers’ metadata from images they host. The IPTC has tested 15 social media websites, looking at how image sharing, through upload and download, affects the integrity of embedded metadata as defined by the IPTC standards and the Exif standards. The results show that Facebook and Flickr are some of the worst offenders, with most of the metadata removed from the original files uploaded. Twitter has also been found to remove Exif and IPTC metadata from its files. Google+, however retained all types of metadata even when the pictures are embedded or downloaded from the social media site [20].

Although social networks are the main source of news for many people today, they are not considered reliable due to the concerns mentioned above. Clearly, in order for social networks to serve as a more reliable platform for disseminating critical information, it is necessary to have tools that limit or help in combating the effect of misinformation [21].

C. *Wikipedia information management*

Wikipedia has become an important source of information online, with more than 37 million articles published by November 2015 in more than 250 different languages. A study in 2005 [22] published by Nature concluded that the scientific articles published in Wikipedia and edited mainly by anonymous contributors came close to the level of accuracy of those published in Encyclopedia Britannica which are provided by renowned scholars and scientists. A key component in how Wikipedia manages to reduce misinformation and increase reliability of information is through their validation process and referencing.

Wikipedia still does not receive the same appraisal in academia due to the lack of the peer-review process and sometimes to the lack of proper referencing. Whether or not to trust information online will be largely debatable [23] and no consensus has been yet reached about this issue. Still Wikipedia has gained a huge success due to its strict authenticity of the information process. The system is mainly based on a letter scheme which reflects the quality of each article. The quality classes available are stub, start, class C, class B, Good Article, and Featured Article [24].

Good Articles are articles that passed the Good Articles criteria but not enough to make it to the Featured Articles category. Good articles should usually be [26] well written; verifiable, i.e. contain a list of all references, all inline citations are from reliable sources, and should contain no copyright violations or plagiarism; broad in its coverage, i.e. address main aspects and stay focused on the topic; neutral, stable, and illustrated. Featured Articles [27] are the best articles published on Wikipedia, distinguished by professional standards or writing, and sourcing. Featured Articles should be well-written; comprehensive, i.e. should neglect no major facts or details; well-researched, where claims are verifiable against high-quality reliable sources and supported by inline citations; non-biased, stable, follow the Wikipedia style guidelines, contain enough media, and stay focused on the main topic without going into unnecessary details. Obviously, articles in other categories are less demanding in terms of quality, such as Class B articles which are mostly complete and without major problems, but require some further work to reach the Good articles category; Class C articles which are substantial, but still missing important content or contain much irrelevant material; Start articles which are developing, but quite incomplete with missing adequate reliable sources; and Stub articles which provide only a description of the topic. In the image below Figure 4, the evolution of an article in Wikipedia ("Atom") is presented. It demonstrates how an article's profile can develop through levels and time. For instance, it took the article "Atom" around 6 years to get from a Stub article to a Featured Article.

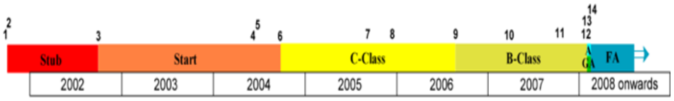


Fig. 4. Development of the Article "Atom" through levels [24]

The quality assessments of articles from stub to class B are mainly performed by members of WikiProjects [26], which are projects allowing a group of contributors to work together as team to improve a certain topic area in Wikipedia. Good Articles and Featured Articles, on the other hand, are assessed by selecting an external panel. Candidates for these panels are nominated on nomination lists and then judged against well-defined criteria [24]. Judgments usually demand a certain consensus to be reached for an article to be categorized, and some projects have even assessment teams.

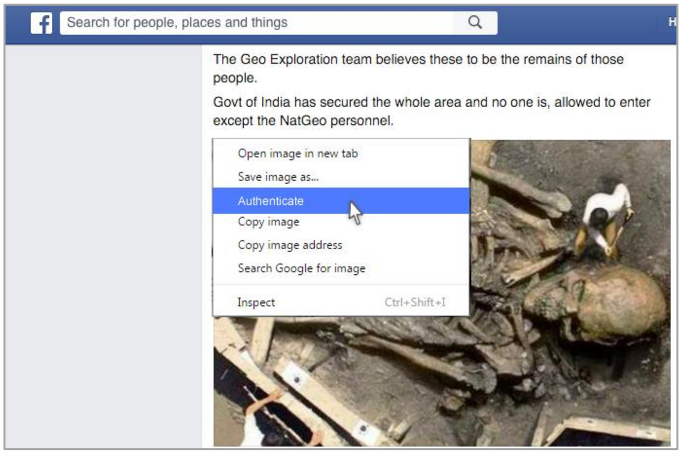
This thorough process, which is not immune to faults, has proven important and reasonably effective tool in Wikipedia fight to combat misinformation on its website. However, this process is deemed lengthy and time-consuming, making it more suitable for verifying the authenticity of static information rather than dynamic information such as news that might need a less rigorous but rapid method to verify it.

D. Current publications on combating misinformation

It came as no surprise that there is very limited literature on this subject. One work by Budak et al. [25] presented a network algorithm that could be tested in case of two competing campaigns that would test the accuracy of the information. The paper, theoretical, relies on the design of the system itself and input of 'influential' people to counter 'bad' campaign and limit misinformation. This could potentially be useful during time-sensitive political campaigns or breaking news events. The paper, however, does not suggest how the method in which 'good' campaign can participate in countering 'bad' campaign. No other academic papers could be found on this topic. The Observer-France 24 posted a guideline for verifying photos online, a process involves some 15 steps and thus confirming the challenges that web users have when authenticating images online [5][36].

III. Approach

The literature review on textual and image misinformation has demonstrated the issues and problems surrounding misinformation online and specifically on social media. However, the tools that can be used to review, rank, and identify misinformation are already found online but may have not been used combined together in a format that would help users in their quest for authentication check. The proposal here is a conceptualization of a quick and easy process that could be used to combat misinformation online. It should and could start with a right-click 'Authenticate' option as shown in Figure 5.


The Geo Exploration team believes these to be the remains of those people.

Govt of India has secured the whole area and no one is, allowed to enter except the NatGeo personnel.

Fig. 5. Conceptualizing a right-click 'Authenticate' option

A. Images checks

Reverse image search [37] using Google Images [29], available via Chrome desktop browser as an add-on, is one tool that is underutilized. This is a completely different search engine to Google image keyword search that returns images based on the web user keywords using their standard search page. This search requires user to upload an image or copy the image's web address to search for matches to that actual image online. The results reveal the sources and dates of the first appearances of that image online and content which appeared with that image. The Google Images search is refined to detect even modifications of the image including color tones changes, cropping, and writing, yet still be able to link it to the original image. Finding the earliest sources of an image is the first step to validating the image origin or the stories associated with it. Second layer is to validate any meta-data linked with the questioned image. Original, image metadata could reveal the device that was used to take the image, the creation date, what changes and on which parts of the image these meta-changes have taken place. Meta-data may also help detect if any image editing tools have been used [31] [32]. Finally, an editorial feedback in a similar format to how Wikipedia operates authentication of information, could be linked to an image. Image editorial feedback maybe combined with explanation based on the origin, date, meta-data, where it appears online, or article that dismisses or confirms that image. Finally a crowdsourcing of feedback could be added as final confirmation. These four sections could be identified as: Image Match, Image Metadata, Editorial, and Feedback respectively. The solution would be to bundle these four sections into one single right click option, see Figure 6.

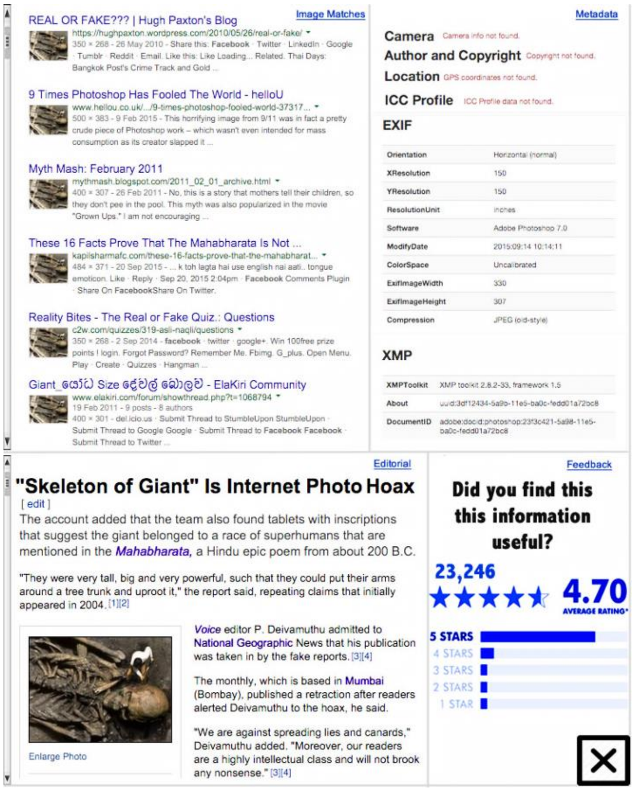


Image Matches

REAL OR FAKE??? | Hugh Paxton's Blog
https://hughpaxton.wordpress.com/2010/05/26/real-or-fake/
350 × 268 - 26 May 2010 - Share this: Facebook · Twitter · LinkedIn · Google · Tumblr · Reddit · Email. Like this: Like Loading... Related. Thai Days: Bangkok Post's Crime Track and Gold ...

9 Times Photoshop Has Fooled The World - helloU
www.hellou.co.uk/.../9-times-photoshop-fooled-world-37317...
500 × 383 - 9 Feb 2015 - This horrifying image from 9/11 was in fact a pretty crude piece of Photoshop work – which wasn't even intended for mass consumption as its creator slapped it ...

Myth Mash: February 2011
mythmash.blogspot.com/2011_02_01_archive.html
400 × 307 - 26 Feb 2011 - No, this is a story that mothers tell their children, so they don't pee in the pool. This myth was also popularized in the movie "Grown Ups." I am not encouraging ...

These 16 Facts Prove That The Mahabharata Is Not ...
kapilsharmafc.com/these-16-facts-prove-that-the-mahabharat...
484 × 371 - 20 Sep 2015 - ... k toh lagta hai use english nai aati.. tongue emoticon. Like · Reply · Sep 20, 2015 2:04pm · Facebook Comments Plugin · Share On FacebookShare On Twitter.

Reality Bites - The Real or Fake Quiz.: Questions
c2w.com/quizzes/319-asli-naqli/questions
350 × 268 - 2 Sep 2014 - facebook · twitter · google+. Win 100free prize points ! login. Forgot Password? Remember Me. Fbimg. G_plus. Open Menu. Play · Create · Quizzes · Hangman ...

Giant_යෝධ Size දේවල් බොලව් - ElaKiri Community
www.elakiri.com/forum/showthread.php?t=1068794
19 Feb 2011 - 9 posts - 8 authors
400 × 301 - del.icio.us · Submit Thread to StumbleUpon StumbleUpon · Submit Thread to Google Google · Submit Thread to Facebook Facebook · Submit Thread to Twitter ...

Metadata

Camera Camera info not found.
Author and Copyright Copyright not found.
Location GPS coordinates not found.
ICC Profile ICC Profile data not found.

EXIF

Orientation	Horizontal (normal)
XResolution	150
YResolution	150
ResolutionUnit	inches
Software	Adobe Photoshop 7.0
ModifyDate	2015:09:14 10:14:11
ColorSpace	Uncalibrated
ExifImageWidth	330
ExifImageHeight	307
Compression	JPEG (old-style)

XMP

XMPToolkit	XMP toolkit 2.8.2-33, framework 1.5
About	uuid:3df12434-5a9b-11e5-ba0c-fedd01a72bc8
DocumentID	adobe:docid:photoshop:23f3c421-5a98-11e5-ba0c-fedd01a72bc8

Editorial

"Skeleton of Giant" Is Internet Photo Hoax
[edit]
The account added that the team also found tablets with inscriptions that suggest the giant belonged to a race of superhumans that are mentioned in the *Mahabharata*, a Hindu epic poem from about 200 B.C.

"They were very tall, big and very powerful, such that they could put their arms around a tree trunk and uproot it," the report said, repeating claims that initially appeared in 2004.[1][2]

Voice editor P. Deivamuthu admitted to National Geographic News that his publication was taken in by the fake reports.[3][4]

The monthly, which is based in Mumbai (Bombay), published a retraction after readers alerted Deivamuthu to the hoax, he said.

"We are against spreading lies and canards," Deivamuthu added. "Moreover, our readers are a highly intellectual class and will not brook any nonsense."[3][4]

Enlarge Photo

Feedback

Did you find this this information useful?

23,246 ★★★★☆ 4.70 AVERAGE RATING*

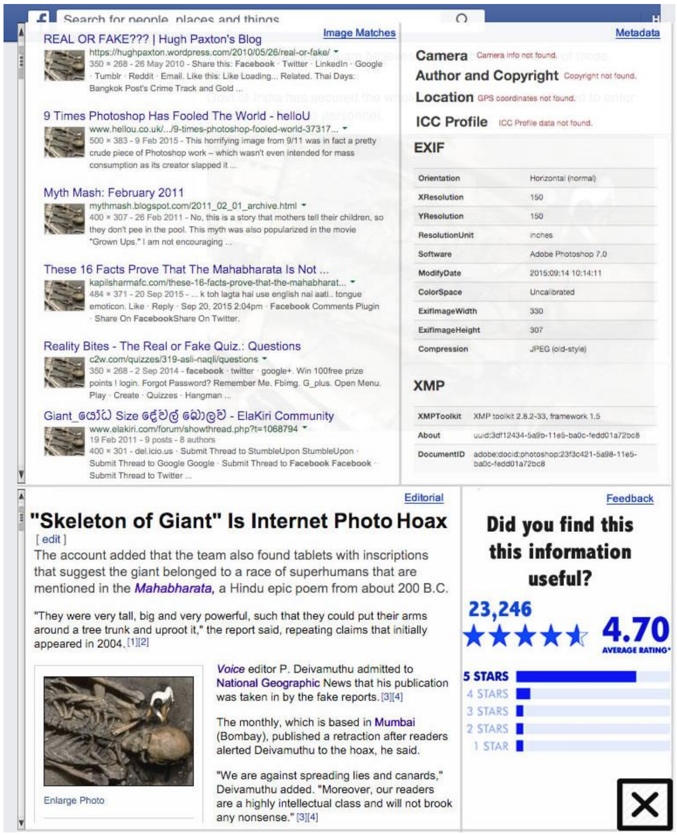
5 STARS
4 STARS
3 STARS
2 STARS
1 STAR

☒

Fig. 6. Conceptualization of the 'Authenticate' outcome as a separate page

So the right-click 'Authenticate' option would perform an image search to display early appearances, dates, and early text linked to the image; Display meta-data that shows creation dates, editing, and originality; editorial section with references; combined with crowdsourcing of feedback from visitors. Where an image is new and the authenticity of the image remains unanswered, this would be shown too although the attention could then be focused on the image metadata. Finally, using the same algorithm used for online search engines, an image that gets frequently selected as a match would get higher ranking than those images that do not get selected as a match. To demonstrate this concept, we have provided a conceptualization images in Figure 5, Figure 6, and Figure 7 of the giant skeleton identified in the literature review.

The Chrome bowser in this case would take the lead with a right click that usually allows users to perform several other options, Figure 5 can now include an 'Authenticate' option. The output could be shown in a new tab, Figure 6 or as a layer over the current display, Figure 7. The information shown in figures 6 and 7 is genuine with the exception of Feedback Section. In the case of the giant skeleton, the Image Match section returned Google Images [29] results that almost immediately questioned the authenticity of the images; the Metadata section showed Adobe Photoshop 7.0 has been used on the photo with no information of camera or author [30]; the Editorial section is taken from National Geographic [13] but could have been easily linked to Wikipedia had it developed a section to authenticate images; and finally the Feedback section could have been the crowdsourcing of feedback allowing final confirmation on the quality of the editorial.



Search for people, places and things

Image Matches

REAL OR FAKE??? | Hugh Paxton's Blog
https://hughpaxton.wordpress.com/2010/05/26/real-or-fake/
350 × 268 - 26 May 2010 - Share this: Facebook · Twitter · LinkedIn · Google · Tumblr · Reddit · Email. Like this: Like Loading... Related. Thai Days: Bangkok Post's Crime Track and Gold ...

9 Times Photoshop Has Fooled The World - helloU
www.hellou.co.uk/.../9-times-photoshop-fooled-world-37317...
500 × 383 - 9 Feb 2015 - This horrifying image from 9/11 was in fact a pretty crude piece of Photoshop work – which wasn't even intended for mass consumption as its creator slapped it ...

Myth Mash: February 2011
mythmash.blogspot.com/2011_02_01_archive.html
400 × 307 - 26 Feb 2011 - No, this is a story that mothers tell their children, so they don't pee in the pool. This myth was also popularized in the movie "Grown Ups." I am not encouraging ...

These 16 Facts Prove That The Mahabharata Is Not ...
kapilsharmafc.com/these-16-facts-prove-that-the-mahabharat...
484 × 371 - 20 Sep 2015 - ... k toh lagta hai use english nai aati.. tongue emoticon. Like · Reply · Sep 20, 2015 2:04pm · Facebook Comments Plugin · Share On FacebookShare On Twitter.

Reality Bites - The Real or Fake Quiz.: Questions
c2w.com/quizzes/319-asli-naqli/questions
350 × 268 - 2 Sep 2014 - facebook · twitter · google+. Win 100free prize points ! login. Forgot Password? Remember Me. Fbimg. G_plus. Open Menu. Play · Create · Quizzes · Hangman ...

Giant_යෝධ Size දේවල් බොලව් - ElaKiri Community
www.elakiri.com/forum/showthread.php?t=1068794
19 Feb 2011 - 9 posts - 8 authors
400 × 301 - del.icio.us · Submit Thread to StumbleUpon StumbleUpon · Submit Thread to Google Google · Submit Thread to Facebook Facebook · Submit Thread to Twitter ...

Metadata

Camera Camera info not found.
Author and Copyright Copyright not found.
Location GPS coordinates not found.
ICC Profile ICC Profile data not found.

EXIF

Orientation Horizontal (normal)
XResolution 150
YResolution 150
ResolutionUnit inches
Software Adobe Photoshop 7.0
ModifyDate 2015:09:14 10:14:11
ColorSpace Uncalibrated
ExifImageWidth 330
ExifImageHeight 307
Compression JPEG (old-style)

XMP

XMPToolkit XMP toolkit 2.8.2-33, framework 1.5
About uuid:3df12434-5a9b-11e5-ba0c-fedd01a72bc8
DocumentID adobe:docid:photoshop:23f3c421-5a98-11e5-ba0c-fedd01a72bc8

Editorial

"Skeleton of Giant" Is Internet Photo Hoax
[edit]
The account added that the team also found tablets with inscriptions that suggest the giant belonged to a race of superhumans that are mentioned in the *Mahabharata*, a Hindu epic poem from about 200 B.C.

"They were very tall, big and very powerful, such that they could put their arms around a tree trunk and uproot it," the report said, repeating claims that initially appeared in 2004.[1][2]

Voice editor P. Deivamuthu admitted to National Geographic News that his publication was taken in by the fake reports.[3][4]

The monthly, which is based in Mumbai (Bombay), published a retraction after readers alerted Deivamuthu to the hoax, he said.

"We are against spreading lies and canards," Deivamuthu added. "Moreover, our readers are a highly intellectual class and will not brook any nonsense."[3][4]

Enlarge Photo

Feedback

Did you find this this information useful?

23,246 ★★★★☆ 4.70 AVERAGE RATING*

5 STARS
4 STARS
3 STARS
2 STARS
1 STAR

☒

Fig. 7. Conceptualization of the 'Authenticate' outcome as layer over a page

B. Text checks

The option to highlight a text and search for it online is already a well-established tool on many browsers [33]. The problem with such tools is that they only search for where the text appears and provides little or no further information on its authenticity. A right-click authenticate could select that sentence (or few sentences) and make specific online search following the criteria listed above: first appearance, origin, and editorial comments. Where there is dispute of its authenticity, this is would be clearly shown. What would make this option useful is that if it can again harness crowdsourcing to link such pieces of information to other pieces of information, which may be presented in different context or different wording. Turnitin [34], a tool used predominantly in academia to check the originality of students' work is one of the tools that could be employed in this context.

IV. Research Challenges and Limitations

There are three aspects of research challenges and limitations the authors acknowledge: research sourcing, conceptualization, and implementation.

The team acknowledges that a large portion of the sources is that of online resources, but they owe much of these to the examples of sharing misinformation on social media, which could not be found or sourced from academic sources.

The second limitation has to do with the limitation of the concept. The authors acknowledge that the 'authentication' option would have little or no real impact at authenticating breaking news. For example of misinformation that spread during the shootings at Fort Hood, Texas. Soldiers inside the base sent out messages via Twitter of multiple shooters and shooting locations that were incorrect. The misinformation quickly spread through the social networks and reached the mass media where it was reported on television broadcasts [2]. The right-click authenticate would not be able in such instances to give real answers or help.

Finally challenge would be the implementation, the authors attempted to recreate a working prototype, which is still in progress and will be subject to further research. However, early on the team has reported code obstacles in the way Google search is written. Early indications suggest that an authorized account is needed with the search engine and even then the codes would only allow text search and not image search [34]. Searching via http coding seems to have been disabled by Google.

V. Further Research

As reported in the challenges, the authors will be working with developers, independent or corporate, to develop the concept into an actual working prototype. The authors also believe that this technique could eventually be extended further to authenticate videos, although complicated algorithms need to be designed and tested. At the time of publishing this paper, no accessible tool could be found to search for matches of videos based on content as opposed to titles. Many online video storage and social media allowed searching by keywords presented by the source and not by distinctive content of these videos.

VI. Conclusion

The work presented here remains at the early stage and is aimed at starting the debate on the importance and corporate responsibility of online companies and browsers towards their customers. The authors have showed that the tools and methods that could be used to authenticate text and images are available and achievable but may need cooperation between different corporations. The duty of regulatory bodies should be to set standards and timeline for these changes to come, in the format proposed or alternative formats. What cannot continue is inaction and with acceptance of improper use of a mass media at the scale of social media and the Internet.

References

[1] Coyle. J. “News of Jackson's death first spread online,” NBC News (online), June 2009. http://abcnews.go.com/Technology/story?id=7938705

[2] Heussner. K. M. “Ft. hood soldier causes stir on twitter,” ABS News (online), November 2009. http://abcnews.go.com/Technology/AheadoftheCurve/tweeting-uniform-ft-hood-soldier-stir-twitter/story?id=9042726

[3] Morozov. E.. “Swine flu: Twitter’s power to misinform,” Foreign Policy, April 2009.

[4] Whitten, S. “Misinformation Spreads on Social Media Following Paris Attacks,” NBC News (online), November 2015. http://www.nbcnews.com/tech/tech-news/misinformation-spreads-social-media-following-paris-attacks-n464291

[5] Vis, F. “To tackle the spread of misinformation online we must first understand it,” The Guardian (online), April 2014. http://www.theguardian.com/commentisfree/2014/apr/24/tackle-spread-misinformation-online

[6] Rusk, D. “How the internet misled you in 2015,” BBC News (online), December 2015, http://www.bbc.com/news/world-35051618

[7] World Economic Forum Report. “Top 10 trends of 2014:10. The rapid spread of misinformation online.” (n.d). http://reports.weforum.org/outlook-14/top-ten-trends-category-page/10-the-rapid-spread-of-misinformation-online/

[8] Dutton, W.H, Blank, G., and Gorseli, D. “Cultures of the Internet: The Internet in Britain.” Oxford Internet Survey 2013 Report: University of Oxford.

[9] Naaman, M. “Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications,” Multimedia Tools and Applications, 2012, vol. 56, issue 1, pp. 9-34.

[10] McAuley, J., and Leskovec, J. “Image labeling on a network: using social-network metadata for image classification,” In Computer Vision–ECCV, Springer Berlin Heidelberg, 2012, pp. 828-841.

[11] Brown, S. “Social Information: Gaining competitive and business advantage using social media tools,” Elsevier, 2012.

[12] Lucas, D. "Famous Pictures Magazine - Altered Images," Famous Pictures Magazine, September 2012. http://www.famouspictures.org/altered-images/#TrotskyWhatTrotsky

[13] Owen, J. “"Skeleton of Giant," Is Internet Photo Hoax, National Geographic News, December 2007. http://news.nationalgeographic.com/news/2007/12/071214-giant-skeleton.html

[14] Facebook User. “Unsolved Mysteries f d world,” February 2014, https://www.facebook.com/permalink.php?story_fbid=1383147595286815&id=1382816751986566

[15] Wildlife, E. “Giant Human Skeleton Discovered,” Youtube, June 2013, https://www.youtube.com/watch?v=azjWu6Uva8k

[16] Andrews, K. “Giants,” Pinterest, (n,d) https://www.pinterest.com/thekeithandrews/giants/

[17] Svendsen, H., & Scardino, P. “U.S. Patent No. 6,954,543,” Washington, DC: U.S. Patent and Trademark Office, 2005.

[18] Parulski, K. A., and McCoy, J. R. “U.S. Patent No. 6,629,104,” Washington, DC: U.S. Patent and Trademark Office, 2003.

[19] Ionescu, B., Popescu, A., Lupu, M., Gınsca, A. L., and Müller, H. “Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation,” MediaEval 2014 Workshop, Barcelona, Spain, October 2014.

[20] Laurent, O. “Study exposes social media sites that delete photographs’ metadata,” British Journal of Photography, March 2013. http://www.bjp-online.com/2013/03/study-exposes-social-media-sites-that-delete-photographs-metadata/

[21] Keim, M. E., and Noji, E. “Emergent use of social media: a new age of opportunity for disaster resilience,” American journal of disaster medicine, 2010, Vol. 6, Issue 1, pp. 47-54.

[22] Giles, J. "Internet encyclopaedias go head to head," Nature 438.7070, 2005, pp. 900-901.

[23] Bachi, Giacomo, et al. "Classifying trust/distrust relationships in online social networks," Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). IEEE, 2012.

[24] Wikipedia Contributors. “Wikipedia : Version 1.0 Editorial Team,” Wikipedia, The Free Encyclopedia, December 2015, https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

[25] Budak, C., Agrawal, D., and El Abbadi, A. “Limiting the spread of misinformation in social networks,” Proceedings of the 20th international conference on World Wide Web, March 2011, pp. 665-674. ACM.

[26] Wikipedia Contributors. “Wikipedia : WikiProject,” Wikipedia, The Free Encyclopedia, July 2015, https://en.wikipedia.org/wiki/Wikipedia:WikiProject

[27] Wikipedia Contributors. “Wikipedia : Good article criteria,” Wikipedia, The Free Encyclopedia, December 2015, https://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria

[28] Wikipedia Contributors. “Wikipedia: Featured article criteria,” Wikipedia, The Free Encyclopedia, October 2015, https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

[29] Google Image, (n,d), https://www.google.co.uk/imghp

[30] Metapicz, An online image metadata analyzer, (n,d), http://metapicz.com/

[31] Buchholz, F. “On the role of file system metadata in digital forensics,” Digital Investigation, vol 1, Issue 4, December 2004, pp. 298–309.

[32] Castiglione, A., Cattaneo, G., and De Santis, A. “A Forensic Analysis of Images on Online Social Networks,” IEEE Conference, 30th Nov-2nd Dec 2011, pp. 679-684. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6132891

[33] Google Chrome. “Simple=Select + Search,” Google Chrome, May 2015. https://chrome.google.com/webstore/detail/simple-select-+-search/aagminaekdpcfimcbhknlgjmpnnnmooo

[34] Dahl, S. “Turnitin®: The student perspective on using plagiarism detection software,” Active Learning in Higher Education July 2007 vol. 8 no. 2, pp. 173-191.

[35] Google Developers. “Creating Custom Search Engines programmatically,” Google Inc., December 2015, https://developers.google.com/custom-search/docs/api?hl=en

[36] Team Observers. “The Observers’ guide to verifying photos and videos on social media networks,” France24 News (online), October 2015, http://observers.france24.com/en/20151110-observers-guide-verifying-photos-videos-social-media-networks

[37] Martin, J. “How to do a reverse Google Image search on Android or iPhone,” PC Advisor (online), February 2016, http://www.pcadvisor.co.uk/how-to/internet/how-do-reverse-google-image-search-on-android-or-iphone-3634872/

A Framework for Classifying Unstructured Data of Cardiac Patients: A Supervised Learning Approach

Iqra Basharat
Department of Computer Engineering
National University of Sciences and Technology
Islamabad, Pakistan

Ali Raza Anjum
Former Specialist - Business Analysis and Planning
Mobilink
Islamabad, Pakistan

Mamuna Fatima, Usman Qamar, Shoab Ahmed Khan
Department of Computer Engineering
National University of Sciences and Technology
Islamabad, Pakistan

Abstract—**Data mining has recently emerged as an important field that helps in extracting useful knowledge from the huge amount of unstructured and apparently un-useful data. Data mining in health organization has highest potential in this area for mining the unknown patterns in the datasets and disease prediction. The amount of work done for cardiovascular patients in Pakistan is scarcely very less. In this research study, using classification approach of machine learning, we have proposed a framework to classify unstructured data of cardiac patients of the Armed Forces Institute of Cardiology (AFIC), Pakistan to four important classes. The focus of this study is to structure the unstructured medical data/reports manually, as there was no structured database available for the specific data under study. Multi-nominal Logistic Regression (LR) is used to perform multi-class classification and 10-fold cross validation is used to validate the classification models, in order to analyze the results and the performance of Logistic Regression models. The performance-measuring criterion that is used includes precision, f-measure, sensitivity, specificity, classification error, area under the curve and accuracy. This study will provide a road map for future research in the field of Bioinformatics in Pakistan.**

Keywords—bioinformatics; classification techniques; heart disease in Pakistan; heart disease prediction; multinomial classification; logistic regression

I. Introduction

Cardiac disease is one of the most serious and death causing health problem. This disease has its bad effects not only on older people but affected severely younger generation also. There are some significant risk elements of cardiac disorder like excessive amount of cholesterol, high blood pressure, hypertension, smoking and sometimes family history. Various precautionary mediations for individuals are available, involving either prescription or change in daily life routine. There exists raw data in the form of patients' history and complex reports. All these resources are key factors to extract meaningful results, for better medical diagnosis. This data can be processed and analyzed to extract valuable in-formation that helps practitioner in decision-making and cost saving. Medicine or Bioinformatics technology has flourished a lot in the past few years that we have applied in healthcare industry such as; treatment effectiveness, customer relationship management, healthcare information management fraud and abuse. However, the significant applications in data processing perhaps implicate predictive modeling [1].

Data mining enacts a substantial part in various applications of different domains such as business corporation's, e-commerce, healthcare industry, science and engineering. In the health care industry, primarily it is used to predict disease. Various diseases like Diabetes, Hepatitis, Cancer and are diagnosed using data mining or machine learning techniques [2]. According to Benko and Wilson in healthcare organizations where data mining is being practiced are performing better in meeting their long term needs. Benko and Wilson argue [3] "In healthcare, data mining is becoming increasingly popular, if not increasingly essential."

To assist the medical practitioners, intelligent information system, knowledge based system and prediction systems are being developed. Healthcare organizations have kept voluminous data of patients in the form of medical reports, patients' history, electronic test results etc. [4]. This data in its present unstructured form is complex, noisy, high dimensional and discrete [5]. Considerably there is a lot of useful knowledge buried in those records. However, the question arises that how can we mine and transform those unstructured and complex reports into practically useful information, that could assist the doctors to draw knowledgeable medical conclusions.

A. Research Motivation

The main motivation of this research is to propose a framework to extract the hidden valuable information from the unstructured records of patients in the form of medical reports and to classify the data into important classes or patients' impressions that could assist the healthcare experts to make intelligent decisions and for predictive analysis.

B. Objectives and Contribution

This study is designed at predictive analysis/classification of cardiac patients by proposing a classification framework for un-structured data. We propose to use a classification framework that emphasize on pre-processing of unstructured data of healthcare organization in the form of patients' reports. Hence, contributions of this research study are as follows:

1) We present a manual approach that extracts attributes (like age, sex, blood pressure, LVEF value, BMI, defected areas, etc.) from patients' reports in order to classify the patient's condition. This study aims to provide a framework that uses supervised learning techniques of data mining.

2) We have used the multinomial class label (namely, fair, moderate, risk and critical).

3) A classification framework is proposed that uses supervised learning techniques to classify unstructured of data patients into four classes mentioned above.

4) We use best-known machine learning technique, Logistic Regression, which is most widely being used in prediction of heart disease.

5) Comparison performance evaluation is presented based on some performance measures are explained later section 3.

II. Literature Review

Cardiac syndrome is a serious and a death-causing syndrome [6]. However, the science and Bioinformatics has developed a lot and treatment for this disease is possible and available to almost every person. The increase in no of deaths occurring because of cardiac disease all around the world has focused the attention of medical practitioners and researcher on this serious issue. There is quite a comprehensive literature available on this topic. Various useful medical applications and decision support system have been advanced to aid the medical practitioner in better medical treatment of their patients. . These systems predict the likelihood of patients getting heart disease or heart attack, etc. data mining has played an important role in this field. Here, patients historical data is used to make and develop such system where artificial intelligence and machine learning techniques are used widely. The ongoing research in this field has provided much success and opened up the doors for further improvements.

It has been noticed from the detailed survey of the literature that SVM, Logistic regression, Neural Nets and Naïve Bayes, are most widely use algorithms for heart disease prediction. To predict the survival of cardiac patients, three prediction models were built on 1000 cases of cardiac heart disease patients. Using a binary categorical variable (1 for survival and 0 for non-survival), 10-fold cross validation procedure was performed on SVM, ANN and Decision trees. This gives less biased prediction and highest classification accuracy in SVM [6].

To identify and prevent the cardiovascular disease, classification techniques are used. Authors of [7] present comparison of Artificial Neural Network, Decision Tress and RIPPER and SVM techniques. Based on accuracy measures these techniques were compared. The results of this study show that the Support Vector Machine model is the best giving 84% accuracy. Ripper, a classification algorithm based on association rules with reduced error pruning algorithm gives 81% accuracy, whereas Decision tree gives the least accuracy and sensitivity ratio among all three-classification models.

Literature reports various comparative studies on data mining, classification algorithms for predicting heart disease. One such comparative analysis was carried out using the Cleveland cardiovascular disease dataset from UCI repository with 13 attributes and 303 instances [8]. On this dataset three classification models were developed, namely, Sequential Minimal Optimization, Logistic Function and Multilayer Perceptron Function. The Accuracy of these classification techniques was determined through kappa statistics, ROC, True positive rate and F measures. All these accuracy measures show that logistic regression gives better results than other techniques. The different error rates calculated shows that the Logistic Function algorithm performs much better than other two classification algorithms. The rate of true positives and ROC Area of the point reaches the maximum accurateness in the logistic function algorithm. Even Kappa statistics and F measures give better results in the logistic function than SMO and Multilayer perceptron.

Using physiological measuring devices like Point-of-Care devices (PoC), mobile gateways and monitoring server, a remote cardiac monitoring system was designed for preventive care [9]. The system was developed to provide preventive care services to cardiac patients. By calculating the information gain of features, highly related feature subsets were selected and SVM classifier was applied to them. The proposed prediction algorithm gives 87.5 % accuracy. F. Imran Kurt et al uses a real data set from VA Medical Center from Long Beach, California and compare performances of logistic regression, decision tree, and neural networks for predicting coronary artery disease [15]. Lift charts and error rates were used to compare the performances of these classification models. Prediction of coronary artery diseases by Neural Networks yields excellent results as it gives higher accuracy while classifying the data. Logistic Regression was found to be second most accurate classed whereas, decision trees show least accuracy and highest error rate.

A prototype of intelligent heart disease prediction system (IHDPS) [10] was developed by using Naïve Bayes, Decision Trees and Neural Network. this system is capable of mining unknown patterns and associations correlated to cardiac disease. IHDPS assists medical practitioners to make intelligent decisions as it can give response to simple as well as complex' what if' queries. Further, it delivers operative and inexpensive treatment and enhances visualization and develop better understanding. IHDPS has used the CRISP-DM methodology to make three models (Naïve Bayes, Decision Trees and Neural Network) and Data Mining Extension language is used to create, train, predict and access model content. Classification Matrix and Lift Chart methods are used to check which model gave a maximum percentage of right predictions. In this research study, five mining goals were set and assessed with respect to three trained models. These are:

- Predict those patients who have chances to get heart disease based on their medical profile.

- Find out the important influences and relationships between medical inputs and medical attributes related to the predictable state.
- Find out heart patient characteristics.
- Define attribute values that discriminate nodes favoring and disfavoring the predictable conditions.

Naïve Bayes appears most efficient by answering four out of the five goals and by identifying all important medical predictors; Decision Trees answered three and its results are easier to interpret; Neural Network two and in this the correlation between attributes is hard to interpret. Intelligent heart disease prediction system is constructed using 15 features, in categorical sample data of 909 patients. The authors of the paper suggested that more features and techniques like association rules, clustering and time series can be used by prolonging it.

TABLE I. COMPARATIVE ANALYSIS OF CLASSIFICATION TECHNIQUES USED IN LITERATURE

Ref #	Data Set (Real/ Artificial)	Classification Type (Categorical/ multinomial)	Tools	Techniques						Results
				ANN	LR	DT	NB	SVM	Other	
7	Artificial	Binomial categorical	WEKA	✓		✓		✓	✓	SVM 84% accuracy
8	Artificial	Binomial categorical	WEKA	✓	✓				✓	LR is considered to be best
6	Real	Binomial categorical		✓		✓		✓		SVM achieved highest accuracy
10	Artificial	Binomial categorical		✓		✓	✓			NB gives highest accuracy

Another widely used machine learning technique, Neural Network, is used to predict heart disease, BP and Diabetics with the help of. Their dataset consists of 78 records with 13 input variables on which various experiments are conducted and the system is trained. For heart disease diagnosis, the author has suggested supervised network, which is trained using a Back Propagation algorithm. When the system is input with new data, it will find it from the trained data and generate a list of probable diseases that may be occurred by the patient. The success rate of the system to give the desired output from the inaccurate input is 100%. The results of the study show that a neural network has the extreme capability to be used as an indexing function. It is used for modeling and prediction of experimental data, this is a fast substitute to classical statistical techniques. The system can avoid human error. Thus, the system is reliable and assists medical practitioners in making accurate decisions. Certainly, neural networks cannot replace the expert mind since the expert is more reliable, but it can assist human experts by cross checking their diagnosis. Table 1 shows the comparative analysis of the research carried out in the field of classification of cardiac data.

The literature discussed above reveals that artificial neural networks (ANN) and Decision trees most widely used classification algorithms for categorical data and Logistic regression is also being used widely for prediction. There are several intelligent heart disease prediction systems, which uses different approaches and propose various models implementing Naïve Bayes and ANN.

In countries like China, India and Malaysia and in some European countries much work has been done in medical data mining and specifically in cardiac data mining [11] [12] [13] [14] on the basis of real and artificial data sets.

All the research discussed above is based on either of these countries. Besides, most of the medical data mining work discussed above focuses on either clustering, classification or association rules mining while in some [11] [12] the details of the results are not discussed and visualized properly.

In 2004, a community-based survey was carried out in an ur-ban populated major city of Pakistan (Sunita Dodani et al.). This survey-based study was estimation of the occurrences and aware-ness of different risks associated with coronary heart disease (CHD). The notable risk factor of heart disease was hypertension, obesity and a sedentary lifestyle. Lack of awareness among people was a common reason of negligence about health. The statistical results of high prevalence factors of cardiac heart disease are well presented and it was suggested to develop some guidelines to manage the coronary heart disease. However, the authors of the study did not formulate any guidelines or model to manage and prevent the increasing risk factors of this serious death causing disease.

However, there is little research seen in Pakistan in building a framework specifically for cardiac data mining based on real data obtained from some renowned cardiac hospital. In addition, there is a need of framework that unifies the data mining tasks from data preparation to data visualization and the discovery of knowledge. In this work, analysis is based on the results of machine learning techniques like clustering, correlation and logistic regression to better and complete visualization of results.

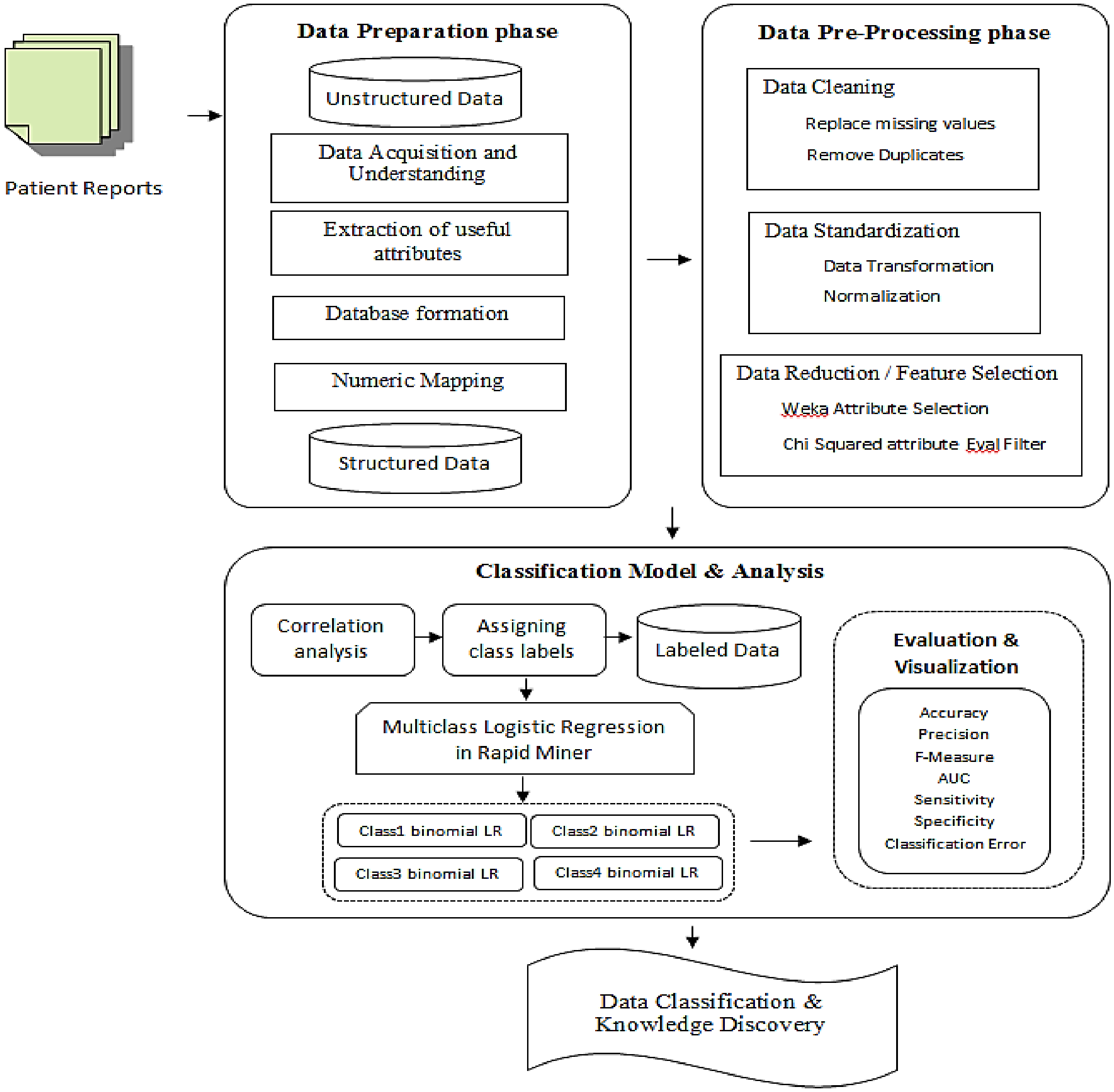


Fig. 1. Proposed Methodological Framework to Classify Unstructured Data

III. Proposed Framework

In this research study of heart disease data for predicting heart patients' condition, following a well-known CRISP-DM methodology we proposed a framework. It is used to classify the unstructured data of cardiac patients. The unstructured reports of cardiac patients were preprocessed manually and a structured database of patients' records was developed. An architectural model of our proposed classification approach is illustrated in figure 1.

It consists of four phases.

A. Data preparation

B. Data preprocessing

C. Feature selection

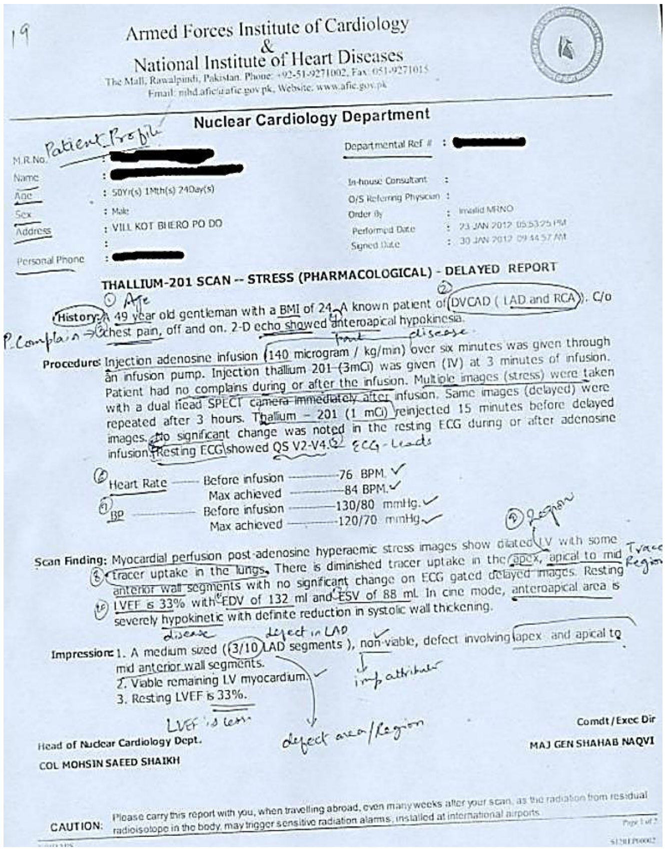
D. Classification model

A. Data preparation

The Data Preparation was carried out in five steps as discussed below.

1) Data Acquisition and Data Understanding

This research study was carried out in close collaboration with the Armed Forces Institute of Cardiology (AFIC), Pakistan. Previously, data used in this type of research study are mostly taken from an online data repository and different classification algorithms are applied on that. In this research, a real data set from AFIC was used. This collected data was unstructured historical records of 1500 patients. During manual preprocessing, data was transformed from unstructured reports to structured format and 50 plus attributes/features were identified. Figure 2 shows a sample patient's report from which attributes/features were extracted.



Armed Forces Institute of Cardiology
&
National Institute of Heart Diseases
The Mall, Rawalpindi, Pakistan. Phone: +92-51-9271002, Fax: 051-9271015
Email: nihd.afic@afic.gov.pk, Website: www.afic.gov.pk

Nuclear Cardiology Department

M.R.No.	: [redacted]	Departmental Ref #	: [redacted]
Name	: [redacted]		
Age	: 50Yr(s) 1Mth(s) 24Day(s)	In-house Consultant	:
		O/S Referring Physician	:
Sex	: Male	Order By	: Invalid MRNO
Address	: VILL KOT BHERO PO DO	Performed Date	: 23 JAN 2012 05:53:25 PM
	:	Signed Date	: 30 JAN 2012 09:44:57 AM
Personal Phone	: [redacted]		

THALLIUM-201 SCAN -- STRESS (PHARMACOLOGICAL) - DELAYED REPORT

History: A 49 year old gentleman with a BMI of 24. A known patient of DVCAD (LAD and RCA). C/o chest pain, off and on. 2-D echo showed anteroapical hypokinesia.

Procedure: Injection adenosine infusion (140 microgram / kg/min) over six minutes was given through an infusion pump. Injection thallium 201 (3mCi) was given (IV) at 3 minutes of infusion. Patient had no complains during or after the infusion. Multiple images (stress) were taken with a dual head SPECT camera immediately after infusion. Same images (delayed) were repeated after 3 hours. Thallium – 201 (1 mCi) reinjected 15 minutes before delayed images. No significant change was noted in the resting ECG during or after adenosine infusion. Resting ECG showed QS V2-V4.

Heart Rate	Before infusion	76 BPM. ✓
	Max achieved	84 BPM. ✓
BP	Before infusion	130/80 mmHg. ✓
	Max achieved	120/70 mmHg. ✓

Scan Finding: Myocardial perfusion post-adenosine hyperaemic stress images show dilated LV with some tracer uptake in the lungs. There is diminished tracer uptake in the apex, apical to mid anterior wall segments with no significant change on ECG gated delayed images. Resting LVEF is 33% with EDV of 132 ml and ESV of 88 ml. In cine mode, anteroapical area is severely hypokinetic with definite reduction in systolic wall thickening.

Impression: 1. A medium sized ((3/10 LAD segments), non-viable, defect involving apex and apical to mid anterior wall segments.
2. Viable remaining LV myocardium. ✓
3. Resting LVEF is 33%.

Head of Nuclear Cardiology Dept.
COL MOHSIN SAEED SHAIKH

Comdt / Exec Dir
MAJ GEN SHAHAB NAQVI

CAUTION: Please carry this report with you, when travelling abroad, even many weeks after your scan, as the radiation from residual radioisotope in the body may trigger sensitive radiation alarms, installed at international airports.

Page 1 of 2

Fig. 2. Cardiac Patient's Report

2) Extraction of Useful Data and Collection of Attributes

We developed a thoughtful understanding of medical reports of patients with the help of medical practitioners and cardiac spe-cialists. The stop criterion was used to determine whether the extracted attributes are mature enough and comprises all the im-portant useful attributes to proceed further or still more reports needs to be filtered out to get valuable attributes. This manual extraction of attributes is a key advantage to get the insight on the problem domain and gives a deep understanding of cardiology.

3) Database Formation

Using MS SQL, database was created to store the extracted attributes after the information retrieval and the structuring of the collected data that was extracted from the past patient reports.

4) Mapping Data to Numeric Values

Once the patients' records are moved into the database, it was quiet easy to handle inconsistency in the data. Machine learning algorithms are applied to structured data. We have used mapping tables in MS SQL Server to transform textual data to numeric data compatible with machine learning algorithms. Once we have prepared the data into structured format and created a database, we move towards the next phase, data preprocessing.

B. Data Preprocessing

The data-preprocessing phase includes data cleansing, data transformation and data reduction.

C. Data Cleansing

In the process of data cleansing, missing, identical and inconsistent data are handled. When the data is in unstructured form, it is manually cleaned and identical and missing records are removed and replaced. All data preprocessing steps are recursive in nature, they are performed in cycles or iteratively. When the data was fed into the database, it was again cleansed with the help of Rapid Miner tool. Missing values were replaced by average value of the attributes with the help of 'Replace Missing Values' operator.

D. Data Transformation

With Rapid Miner 5, a powerful tool of data mining, data was normalized using its 'Normalize' operator. Data normalization standardized the data to a range of 0 to 1 to avoid the attributes with greater values from misbalancing the attributes with smaller values in the assessment procedure [17]. Once the data is gone through the preprocessing, featured selection technique was applied in WEKA to select the important features.

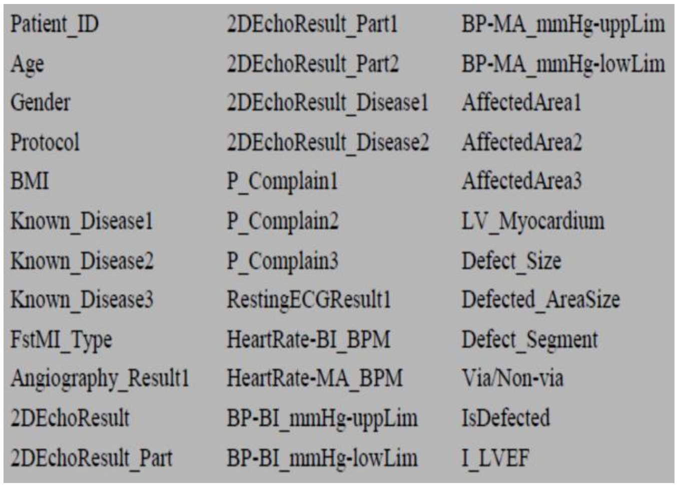
1) Data Reduction

In this step, we remove the irrelevant and redundant data.

E. Feature Selection

Initially the total number of features extracted from the patients' report was 53. Some of the features were irrelevant and redundant that needs to be removed, thus we have used feature subset selection technique. This is a common preprocessing step used in machine learning. Feature subset selection is a step in data preprocessing that helps in reducing the dimensionality and irrelevant data that further enhances the learning efficiency, increases analytical accuracy of classification models The resulted feature subset states those features that are useful in predicting class and thus produces higher classification accuracy [18] [19].

Weka 3.6.9, data mining tool was used to determine important and relevant feature subset using weka.attributeSelection.ChiSquaredAttributeEval technique. It evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. It uses weka.attributeSelection.Ranker search method with threshold 0.5. This search method ranks attributes by their individual evaluations. Figure 3 shows the top 36 selected attributes that are effective in classification accuracy.



Patient_ID
Age
Gender
Protocol
BMI
Known_Disease1
Known_Disease2
Known_Disease3
FstMI_Type
Angiography_Result1
2DEchoResult
2DEchoResult_Part

2DEchoResult_Part1
2DEchoResult_Part2
2DEchoResult_Disease1
2DEchoResult_Disease2
P_Complain1
P_Complain2
P_Complain3
RestingECGResult1
HeartRate-BI_BPM
HeartRate-MA_BPM
BP-BI_mmHg-uppLim
BP-BI_mmHg-lowLim

BP-MA_mmHg-uppLim
BP-MA_mmHg-lowLim
AffectedArea1
AffectedArea2
AffectedArea3
LV_Myocardium
Defect_Size
Defected_AreaSize
Defect_Segment
Via/Non-via
IsDefected
I_LVEF

Fig. 3. Useful Attributes Selected through Attribute Selection Technique

The modeling tools used to carry out this research work is Rapid Miner® 5.3, Microsoft SQL Server R2 2008, WEKA 3.6.9 and Microsoft Excel 2010. except at the beginning of a sentence: "Equation (1) is . . ."

IV. MODELING

Now when we have already gone through feature extraction, pre-processing and finally classification steps, this section will focus on modeling correlation, clustering and prediction model.

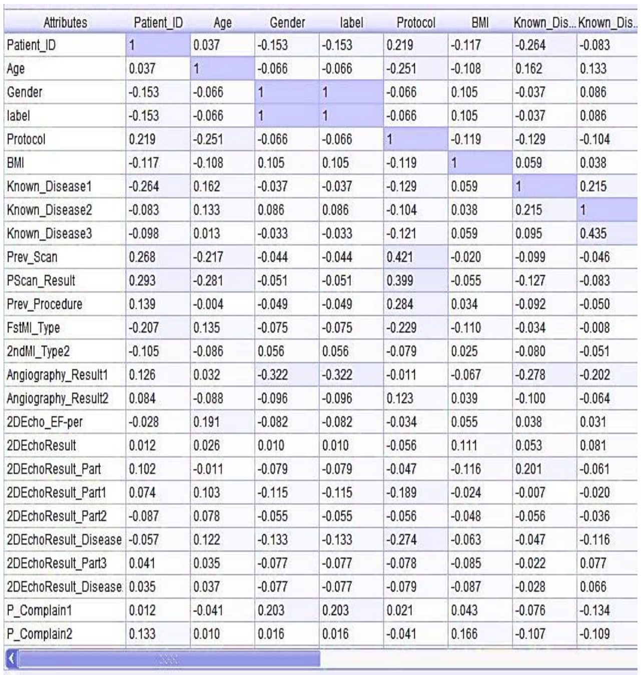
A. Correlation -- Modeling

The correlation coefficient measures the linear relationship between two attributes or columns of data. The value can range from -1 to +1 [20].

Correlation is the association between different attributes. It is a statistical measure, widely being used in data mining for finding the relationship between attributes. We have used The 'Correlation Matrix Operator' of Rapid Miner to calculate the correlation among all the attributes of our data set.

After running the process, the correlation matrix was generated as shown in Figure 4. The results of correlation show that there are no two attributes that can be statistically correlated to each other. Some attributes have a strong positive correlation whereas some have weak correlation. For example, the LVEF is an important attribute that has 0.733 weights. The attributes Via/Non-via and I_LVEF has -0.123 correlations that show the negative correlation with each other when the value of one increases, the second value decreases. In actuality, the low LVEF value indicates a critical or risky situation of cardiac patients and Via/Non-via has value 0 and 1. The 0 value indicated viable condition and 1 indicated non-viable. Similarly, the value of LVEF is negatively correlated with Age and IsDefected attributes.

To label the data for applying the logistic regression model along with intelligent guesses and these weights and relationship between attributes, the clustering results and trends shown in work [21] helped us to assign class labels to each record.



Attributes	Patient_ID	Age	Gender	label	Protocol	BMI	Known_Dis...	Known_Dis...
Patient_ID	1	0.037	-0.153	-0.153	0.219	-0.117	-0.264	-0.083
Age	0.037	1	-0.066	-0.066	-0.251	-0.108	0.162	0.133
Gender	-0.153	-0.066	1	1	-0.066	0.105	-0.037	0.086
label	-0.153	-0.066	1	1	-0.066	0.105	-0.037	0.086
Protocol	0.219	-0.251	-0.066	-0.066	1	-0.119	-0.129	-0.104
BMI	-0.117	-0.108	0.105	0.105	-0.119	1	0.059	0.038
Known_Disease1	-0.264	0.162	-0.037	-0.037	-0.129	0.059	1	0.215
Known_Disease2	-0.083	0.133	0.086	0.086	-0.104	0.038	0.215	1
Known_Disease3	-0.098	0.013	-0.033	-0.033	-0.121	0.059	0.095	0.435
Prev_Scan	0.268	-0.217	-0.044	-0.044	0.421	-0.020	-0.099	-0.046
PScan_Result	0.293	-0.281	-0.051	-0.051	0.399	-0.055	-0.127	-0.083
Prev_Procedure	0.139	-0.004	-0.049	-0.049	0.284	0.034	-0.092	-0.050
FstMI_Type	-0.207	0.135	-0.075	-0.075	-0.229	-0.110	-0.034	-0.008
2ndMI_Type2	-0.105	-0.086	0.056	0.056	-0.079	0.025	-0.080	-0.051
Angiography_Result1	0.126	0.032	-0.322	-0.322	-0.011	-0.067	-0.278	-0.202
Angiography_Result2	0.084	-0.088	-0.096	-0.096	0.123	0.039	-0.100	-0.064
2DEcho_EF-per	-0.028	0.191	-0.082	-0.082	-0.034	0.055	0.038	0.031
2DEchoResult	0.012	0.026	0.010	0.010	-0.056	0.111	0.053	0.081
2DEchoResult_Part	0.102	-0.011	-0.079	-0.079	-0.047	-0.116	0.201	-0.061
2DEchoResult_Part1	0.074	0.103	-0.115	-0.115	-0.189	-0.024	-0.007	-0.020
2DEchoResult_Part2	-0.087	0.078	-0.055	-0.055	-0.056	-0.048	-0.056	-0.036
2DEchoResult_Disease	-0.057	0.122	-0.133	-0.133	-0.274	-0.063	-0.047	-0.116
2DEchoResult_Part3	0.041	0.035	-0.077	-0.077	-0.078	-0.085	-0.022	0.077
2DEchoResult_Disease	0.035	0.037	-0.077	-0.077	-0.079	-0.087	-0.028	0.066
P_Complain1	0.012	-0.041	0.203	0.203	0.021	0.043	-0.076	-0.134
P_Complain2	0.133	0.010	0.016	0.016	-0.041	0.166	-0.107	-0.109

Fig. 4. Correlation matrix generated through RapidMiner®. Correlation coefficients are visible

We studied the correlation results, consulted the experts' views, sorted out some rules manually, and based on those rules we classified the data into four classes and allocated each row a target class label. The four classes are, Normal: patient's condition is fair, means he is normal, Moderate: patient is having a moderate type heart disease, Risk: Patients condition is risky, means having serious heart disease and Critical: patient condition is critical.

B. Logistic Regression --- Modeling

Similar to linear regression, logistic regression is a extrapolative analysis, but logistic regression implicates the likelihood of a dichotomous contingent variable, whereas the predictors can be continuous or dichotomous [22].

1) Data preparation

The major preprocessing and data preparation was done in the previous section. The problem under discussion was multi-class classification problem. However, Logistic Regression deals with binary classification problems. To perform multiclass classification, there are some simple methods for transforming multi-class problems into a set of binary problems. These techniques are known as class-binarization techniques [23].

Using this approach, the data sheet was arranged in four sets with four different class label attributes (i.e. Normal, Risk, Moderate and Critical). Let say we have dataset A, B, C and D where dataset A is Normal / not normal, dataset B is Moderate / not moderate, dataset C is Risk / not risk and dataset D is Critical / not critical. After classifying the data into four categories in previous section it has been noted that this data was not balanced. Each class has different number of patients.

2) *Process Description*

After being done with data preprocessing tasks, Logistic Re-gression was applied on the dataset in Rapid Miner. The 'Retrieve' operator, load the data set to be trained. In order to label the class attribute we have used 'Set Role' operator to set the role of 'Report-Category' attribute as class label. The attribute is to be predicted by the model. As the 'Logistic Regression' operator takes the binomial label, so we have used 'Numerical to Binomial' operator and selected the Report Category attribute.

The Logistic Regression operator is used with 'dot' kernel type and complexity constant 0.5. It receives the training dataset in the input port and runs the algorithm for logistic regression. The logistic regression model is delivered at the output port. In order to use the learned/trained model on unseen data 'Apply Model' operator is used to calculate the performance of training model using 'Performance' operator that provides some important performance measures. The results of trained model are saved in the excel sheet using 'Write Excel' operator.

TABLE II. CONFUSION MATRIX OF FOUR LOGISTIC REGRESSION MODELS

Classes		True Negative	True Positive
Normal Class	Pred. False	775	138
	Pred. True	144	443
Moderate Class	Pred. False	514	229
	Pred. True	397	360
Risk Class	Pred. False	870	80
	Pred. True	467	83
Critical Class	Pred. False	1119	104
	Pred. True	151	126

One by one, the data sheets with four class attributes were loaded into Rapid Miner and four models were generated respectively. Our concerned attribute was, "Report-Category". This was the categorical attribute of a patient, summarizing his complete impressions as concluded by the cardiologist. The categories involved are; Normal, Moderate, Risk and Critical. Along with training the data the validation of the models was done using 10-fold cross validation. To analyze the performance of a classifier, confusion matrix is obtained using 'Performance' operator. Con-fusion matrix obtained from our four logistic regression models as shown in table 2. The confusion matrix helps, we can obtain accu-racy of model with the help of some important accuracy measures. We evaluated our models based on some performance measuring criteria discussed in next section.

V. RESULTS AND ANALYSIS

A. *Performance Measuring Criteria*

Extensive effort was made in order to obtain optimal results of classification by playing-around with different values of attributes used in modeling to run the model. Large numbers of the dataset and parameter alterations were carried out to reach optimal results against each model. Models that generated graceful-predictions against our stipulated 'performance criterion' are covered below.

1) *Accuracy Measures*

To classify the heart disease using logistic regression models, the elementary phenomenon used in calculating the performance and accuracy of the classifier. Sensitivity and Specificity are used for computing the accuracy. These Sensitivity and Specificity are obtained from a confusion matrix resulted from classification model. The confusion matrix displays the number of true positive, true negative and false positive, false negative assessments. It shows the comparison of actual values in the test dataset with the predicted values in the trained model. To measure the performance of logistic regression models we have used the performance measurement criteria [8]. The accuracy measures are shown in table 3.

TABLE III. ACCURACY MEASURES [8]

Accuracy Measures	Description
Precision	Precision is the proportion of relevant documents in the results returned.
F-measure	F Measure is a way of combining recall and precision scores into a single measure of performance.
Area Under Curve (AUC)	The AUC is an estimate of the probability that a classifier will rank a randomly chosen positive instance, higher than a randomly chosen negative instance
Sensitivity	TP/(TP + FN) (Number of true positive assessment)/(Number of all positive assessment)
Specificity	TN/(TN + FP) (Number of true negative assessment)/(Number of all negative assessment)
Accuracy	(TN + TP)/(TN+TP+FN+FP) (Number of correct assessments)/Number of all assessments)

B. *Experimental Results and Discussion*

The models were evaluated on performance criterion discussed above. To evaluate the unbalanced dataset, the more pertinent measures are precision, F-measure, AUC, sensitivity and specificity.

TABLE IV. ACCURACY MEASURES FOR LOGISTIC REGRESSION MODELS

Logistic Regression Models	Precision %	F-measure %	AUC	Classification Error %	Sensitivity %	Specificity %	Overall Accuracy %
Normal Class	66.11	75.86	0.890	18.80	76.01	81.80	81.20
Moderate Class	48.50	51.65	0.576	41.73	63.47	56.72	58.27
Risk Class	13.82	28.34	0.67	35.07	65.82	64.46	64.93
Critical Class	45.49	49.70	0.810	17.00	37.14	87.77	83.00
Average	43.48	51.38	0.73	28.15	60.61	72.68	71.85

There exists a commonality in these metrics, as they are all class independent. These are calculated by confusion matrix The confusion matrix is obtained for all three models and detailed accuracy is shown in the table 4. The area under the curve (AUC) is an independent metric. It gives equal weight to both classes and the greater the value of AUC the better the classifier performance is.

Similarly, for f-measure, precision, sensitivity and specificity, larger values show better performance. Hence, the logistic regression classifier for Normal Class gives better performance. To easily understand and depict the results of the experiment graphical representation of the results are presented in the following figure 5.

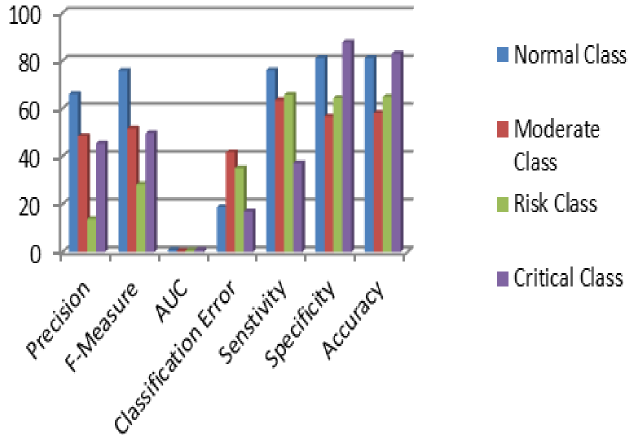


Fig. 5. Performance Measures of Four Logistic Regression Models

From table 4 it is noted that the critical class model performs best among all with 83% accuracy. Although it is showing least precision as compared to Normal and Moderate Class models, but it does not affect its accuracy. In the meantime, it has at least classification error. The second best model is Normal Class model with accuracy of 81.2 % and classification error of 8.8%. The risk Class model shows 64.93% accuracy with the lowest 35.07% classification error and least precision.

It is observed that models with high accuracy can have the least precision that shows precision is not dependent of accuracy. The model can be very precise, but inaccurate, as described above. Moreover, it can be accurate, but imprecise. Accuracy states the close relationship between measured value, standard or known value, and precision shows how close two or more measurements are to each other [24]. By computing the average accuracy of four models, we get to know the overall accuracy of logistic regression is 71.85%. However, here, we can see that for imbalanced dataset, accuracy measure is not an appropriate metric to evaluate the classifier. That is why we calculate other performance measures to better evaluate the performance of classifiers.

VI. CONCLUSION AND FUTURE WORK

The main motivation of this research was to propose a framework to extract the hidden valuable information from the unstructured medical reports of patients and to classify the data into important classes or patients' impressions that could assist the healthcare experts to make intelligent decisions and for predictive analysis.

The data set was taken from the Armed Forces Institute of Cardiology (AFIC), Pakistan. It was collected in unstructured form that was then preprocessed and maintained in a structured form in the database. We used Weka 3-6-4, a famous data-mining tool for classification. ChiSquaredAttributeEval technique was used to lessen the dimensionality and irrelevant data that increased the learning efficiency and analytical accuracy of classification models. Logistic Regression was used to predict four classes with the help of four models. To evaluate the performance of classifier models, accuracy measures were used. The aggregated results showed that when Logistic Regression was applied on four binomial models it gives a classification accuracy of 71.85 %.

The achievement of this research study is as follows:

- Useful data is extracted manually from unstructured records of patients.
- Database of heart patients is designed that could be further used for research and practical implementation of intelligent decision support systems.
- Logistic regression is used to classify the data into four classes of patients; Normal Class, Moderate Class, Fair Class and Critical Class.
- The proposed framework can be utilized for mining unstructured data in other health care centers

- The study on the whole gives new directions in the field of biomedical research in Pakistan.

The classification and predictive analysis of such data is of utmost importance nowadays. The future work of the research can be:

- These results can always be improved by improving the classification/prediction model. In the future, to improve the result by applying 'bootstrapping' technique that would balance the data and thus will give better results.
- Secondly, other important and better classification models like SVM and Artificial Neural Networks could be used to achieve high accuracy.
- The main future concern could be to design an inference engine for cardiac data that would assist the practitioner to make better decisions.
- The results of the study could be further improved by investigating other algorithms and by improving the data pre-processing techniques as well.
- The Text mining technique can be applied to mine the huge unstructured data in hospitals.
- Using the same data set to explore the reason, solution and precautionary measures of specific types of complaining diseases and problem occurring in specific type of patients.

Acknowledgment

We would like to thank and acknowledge the support and guidance of our supervisors. We are also thankful to Armed Forces Institute of Cardiology (AFIC), Pakistan for providing us the data set of heart patients and helping us to understand the research problem.

References

[1] Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare, Journal of Healthcare Information Management—Vol, 19(2), 65

[2] Parpinelli, R. S., Lopes, H. S., & Freitas, A. A. (2001, July). An ant colony based system for data mining: applications to medical data. In Proceedings of the genetic and evolutionary computation conference (GECCO-2001) (pp. 791-797).

[3] Benko, A. & Wilson, B. (2003). Online decision support gives plans an edge. Managed Healthcare Executive, 13(5), 20.

[4] AbuKhousa, E., & Campbell, P. (2012, March). Predictive data mining to support clinical decisions: An overview of heart disease prediction systems. In Innovations in Information Technology (IIT), 2012 International Conference on (pp. 267-272). IEEE.

[5] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Applied statistics, 100-108..

[6] Xing, Y., Wang, J., Zhao, Z., & Gao, Y. (2007, November). Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on (pp. 868-872). IEEE.

[7] Milan Kumari, Sunila Godara "Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction", IJCST Vol. 2, Issue 2, June

[8] Vijayarani, S., and S. Sudha. "Comparative Analysis of Classification Function Techniques for Heart Disease Prediction", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 3, May 2013

[9] Kwon, K., Hwang, H., Kang, H., Woo, K. G., & Shim, K. (2013, January). A remote cardiac monitoring system for preventive care. In Consumer Electronics (ICCE), 2013 IEEE International Conference on (pp. 197-200). IEEE.

[10] SH, M. I., & Sanap, S. A. (2013). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International J. of Healthcare & Biomedical Research, 1(3), 94-101.

[11] Avci, E., & Turkoglu, I. (2009). An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases. Expert Systems with Applications, 36(2), 2873-2878.

[12] Rajeswari, K., Vaithiyanathan, V., & Amirtharaj, P. (2011). Prediction of Risk Score for Heart Disease in India Using Machine Intelligence. In 2011 International Conference on Information and Network Technology, IACSIT Press, Singapore IPCSIT (Vol. 4).

[13] Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).

[14] Dangare, C. S., & Apte, D. S. S. (2012). A data mining approach for prediction of heart disease using neural networks. International journal of Computer Engineering & Technology (IJCET), 3(3), 30-40.

[15] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1), 366-374

[16] Dodani, S., Mistry, R., Khwaja, A., Farooqi, M., Qureshi, R., & Kazmi, K. (2004). Prevalence and awareness of risk factors and behaviours of coronary heart disease in an urban population of Karachi, the largest city of Pakistan: a community survey. *Journal of public health*, *26*(3), 245-24

[17] Han, J., Kamber, M., & Pei, J., (2006), Data mining: concepts and techniques, Morgan kaufmann.

[18] Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection. Global Journal of Computer Science and Technology, 13(3).

[19] Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. arXiv preprint arXiv:0912.3924.

[20] http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.db2.udb.doc/admin/c0006909.htm.

[21] Fatima, M.; Basharat, I; Khan, S.A; Anjum, AR., "Biomedical (cardiac) data mining: Extraction of significant patterns for predicting heart condition," Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on , vol., no., pp.1,7, 21-24 May 2014.

[22] http://www.upa.pdx.edu/IOA/newsom/pa551/lectur21.htm] (Retrieved on 23 March 2014).

[23] Fürnkranz, J. (2002). Round robin classification. The Journal of Machine Learning Research, 2, 721-747.

[24] http://www.ncsu.edu/labwrite/Experimental%20Design/accuracyprecision.htm (Retrieved on 2nd April, 2014).

Optimum Route Selection for Vehicle Navigation

Dalip
Department of Information Technology
MMEC, Maharishi Markandeshwar University
Ambala, India

Vijay Kumar
Department of Computer Science and Engineering
MMEC, Maharishi Markandeshwar University
Ambala, India

***Abstract*—The objective of Optimum Route Selection for Vehicle Navigation System (ORSVNS) article is to develop a system, which provides information about real time alternate routes to the drivers and also helps in selecting the optimal route among all the alternate routes from an origin to destination. Two types of query systems, special and general, are designed for drivers. Here, the criterion for route selection is introduced using primary and secondary road attributes. The presented methodology helps the drivers in better decision making to choose optimal route using fuzzy logic. For experimental results ORSVNS is tested over 220 km portion of Haryana state in India.**

Keywords—Vehicle Navigation; Route Selection; Fuzzy; Optimum; Navigation System

I. Introduction

Today, enhancement in advance technologies (communication, sensors, microelectronics and information technology) is technically possible to provide the real time [1] [16] information of traffic conditions to drivers. A Dynamic Route Guidance (DRG) system provides the actual route advice to drivers in real time traffic condition to reduce the traffic delays. This system is implemented in Indian road context, which are categorized in the following parts- Express Way, National Highways, State Highways, Urban Roads and Rural Roads. National Highways and Express ways are the most preferred by the most of the DRG system routes to reach the destination. As there is no Express Way in Haryana (experimental study), it is not included in the presented work. The traffic volume is increasing every year. The production of automobiles in India is in millions of [2] units as shown in Fig. 1, which is the main cause of traffic congestion. The other causes are poor road condition, poor traffic management and heterogeneous mode of transport (Heavy goods vehicle, Trunk, Bus, Mini Bus, Cab and motorcycle etc.). There are three modes of route selection- short, economical and easy to travel. The route with the shortest distance is recommended in short mode. In case of economy mode, the most economical path in terms of fuel, toll plaza and vehicle maintenance cost is recommended. In easy mode, a route is selected which is suitable for easy driving. Firstly, short, mode is most commonly used mode in all DRG systems.

An alternate route provides [3] additional capacity to service primary route traffic. Alternate routes start from one point on primary route and ends at another point as shown in Fig. 2. Some events like the traffic clogging, emergency, natural disasters, traffic incidents, snowstorm, fog, flood, road blockage, forgotten route by drivers are responsible for alternate routes.

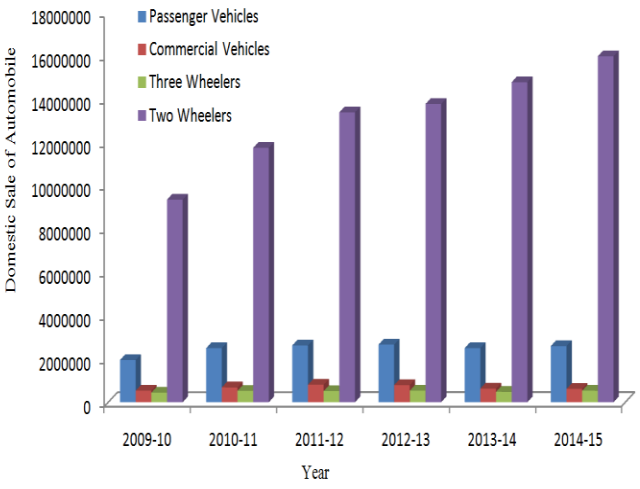


Fig. 1. Domestic Sales of Automobile in India [Source: SLAM]

Fig. 2 shows a road network, which represents three alternate routes with beginning and terminal point on the primary route. In case of blockage of selected route there is need of route diversion and reassignment, which is represented in Fig. 3. Planned events like major roadway construction and maintenance, and unplanned events like traffic incidents, emergency roadwork, adverse weather and emergency are the causes for alternate route plan. The congestion impacts of these events on road capacity are blocked travel lane(s) or road segment, reduce traffic speed and capacity.

A large number of route guidance, alternate route selection, optimum route selection and route diversion systems (Peter Bonsall *et al.* (1992), Kim *et al.* (2003), Mohamed Abdel-Aty *et al.* (2004), Shou-Ren HU *et al.* (2005), Hey Ran Kim *et al.* (2005), Alexander Aved *et al.* (2007), S.S. Keshkamat (2008), Constantinos Antonioua *et al.* (2011), Anastasia Spiliopoulou *et al.* (2013), Xingang Li *et al.* (2014)) have been developed to select the optimum route but a limited technical literature for selecting optimal route to the best of our knowledge is presented here. Peter Bonsall designed a route guidance [4] and information system on route choice for urban networks. Advanced Traveler Information Systems (ATIS) provided [5] alternate routes to driver using generalized estimating equations and binomial probit link function. Time taken to travel on diverted route is less as compared to on normal route thus it increases the probability of diversion. Real-Time Route Diversion Management System (RTRDS) was implemented to [6] create optimal route diversions using available real-time and historical traffic information.

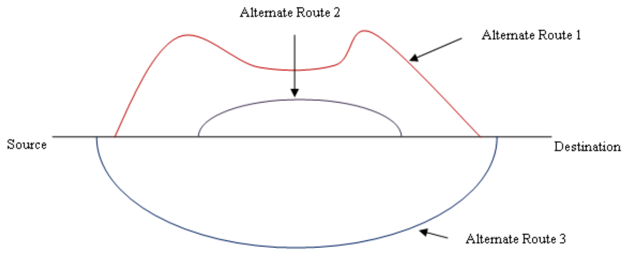


Fig. 2. Road Network

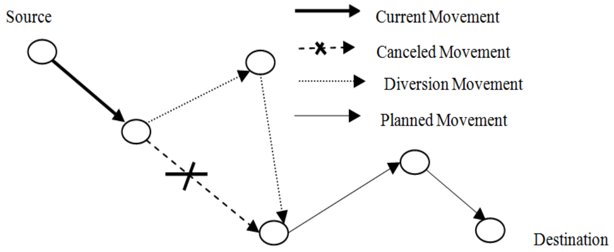


Fig. 3. Route Diversion and Reassignment

A flexible and cost-effective systematic framework based [7] on traffic assignment models was designed to analyze the value of traffic information in route diversion control scheme. With the use of online traffic management system, other dynamic route diversion model [8] is developed to capture travelers' route and their switching behavior in real time Adaptive Kalman Filtering technique is also used to present freeway O-D pattern prediction algorithm. The framework presented by Constantinos Antonioua *et al.* has discussed the impact of predictive guidance in reducing average travel time and travel time variability during incidents. The predictive Variable Message Signs (VMS) [9] shifts a significant number of vehicles from routes taking higher travel time to routes taking lower travel time, showing more consequently reduction in delays and efficient utilization of the network capacity. The contribution of onsite information [10] is to estimate the influence of information on drivers' route diversion decision and network performance accurately. To avoid the [11] recurrent motorway congestion, two route guidance policies have been developed, which is triggered by a saturated off-ramp, but is not efficient to resolve the spread out and motorway congestion problem. To resolve this, a second case based [12] on user-optimum considerations was implemented. Xingang Li *et al.* designed the diversion drivers' [13] from expressway to analyze the behavior under different conditions of network and traffic, to find the key factors influencing diversion behavior. Multi criteria methods for high [14] speed rail route selection presented by Mohd Rapik Saat *et al.* for Malaysia. These methods provide a set of tools to analyze and compare the different alternatives for a rail transport route. It is found from the above review of literature that there is a need of existing as well as implementation of new systems, which will help the system users to select more precise optimal route among all alternate routes on the basis of vehicle type. A fuzzy logic based decision making system is designed to achieve our goal which displays optimal route using experienced based fuzzy score of primary and secondary road attributes.

The rest of the paper is organized as follows. The section I describes introduction and the related work for optimal route selection for vehicle navigation system. The section II presents case study from origin (MM University, Mullana) to destination (Karnal) of Haryana state, India. The road characteristic and discussion on primary and secondary road characteristics is provided in section III. The optimal route selection criterion is discussed in section IV. Section V presents the results and discussion of implemented work. Finally section 6 shows the conclusion and future work of this paper.

II. Case Study: Analysis of Routes from MM University to Karnal

From an origin (MM University, Mullana) to destination (Karnal) route is taken with three alternate routes to present the specific problem in this case study. Table I presents route description of one original and three possible alternate routes of Haryana, India. The distance pair between the cities is shown in Table II which is used in analysis of routes from origin to destination. As described in Table II, the distances between two same cities remain zero. Entries below zero shows distance between city pairs in kilometers, miles and nautical miles but entries above zero are blank due to repetition of city pair. Traffic congestion may be more during starting and ending session of the companies and educational institutes. Thus the volume of automobile traffic is higher than the usual at that duration of time. As the driver already knows all feasible routes for a trip, there is a need of selecting optimal route and during the time of journey, before each road junction, the driver has the options to evaluate alternate routes. The driver has to take decision to choose alternate route on the basis of road attribute values. The evaluation of route is measured by number of factors. For instance, if the driver wants to reach his/her destination as soon as possible, the dominant factor would be travel time of all feasible routes from origin to destination.

In case of Route1 blockage or traffic congestion on Route1, alternate routes will be considered, but the problem in choosing more precise optimal route among all alternate routes according to vehicles type arises here. For example Route4 will not be well suitable for heavy vehicles in term of several road attributes like "bridge clearance", "bridge load", "unnecessary detours" and "congested road". The following research gaps are analyzed in this case study.

- No criteria for selecting route according to transportation mode
- No query system designed for general and special query
- No provision for selecting more precise optimal route

The solution of these research gaps and the criteria for optimum route selection are explained in the further sections.

TABLE I. ROUTE DESCRIPTION OF MM UNIVERSITY TO KARNAL

	Original/ Alternate	**Stop List**	**Via**	**Route Type**
Route1	Original	[MM University, Barara, Sahabad, Pipli, Nilokheri, Trarori, Karnal]	Barara	[Urban+ National Highway]
Route2	Alternate	[MM University, Mullana, Kalpi, Saha, Sahabad, Pipli, Nilokheri, Trarori, Karnal]	Saha	[Urban+ National Highway]
Route3	Alternate	[MM University, Dosdka, Chapper, Bhamboli, Yamuna Nagar, Diamla, Radaur, Ladwa, Indri, Rambha Karnal]	Yamuna Nagar	[Urban Only]
Route4	Alternate	[MM University, Barara, Babain, Ladwa, Indri, Rambha Karnal]	Babain	[Urban + Rural]

TABLE II. DISTANCE BETWEEN CITY PAIRS OF HARYANA STATE, INDIA

	Yamuna Nagar	Amabala Cantt	Kurukshetra	Karnal	Panipat	Hisar	Rohatak	Jind
(a) In Kilometers								
Yamuna Nagar	0							
Ambala Cantt	58	0						
Kurukshetra	45	46	0					
Karnal	66	81	37	0				
Panipat	99	120	71	36	0			
Hisar	206	190	162	156	142	0		
Rohatak	180	226	177	116	77	90	0	
Jind	149	151	103	84	68	76	59	0
(b) In Miles								
Yamuna Nagar	0							
Ambala Cantt	58	0						
Kurukshetra	28	28	0					
Karnal	41	50	23	0				
Panipat	62	74	44	22	0			
Hisar	128	118	101	97	88	0		
Rohatak	111	140	110	72	47	56	0	
Jind	92	94	64	52	42	47	36	0
(c) In Nautical Miles								
Yamuna Nagar	0							
Ambala Cantt	58	0						
Kurukshetra	24	25	0					
Karnal	36	43	20	0				
Panipat	53	64	38	19	0			
Hisar	111	102	87	84	76	0		
Rohatak	97	122	95	62	41	48	0	
Jind	80	81	56	45	37	41	31	0

III. ROAD CHARACTERSTICS AND DISCUSSION

A road characteristic is the road attributes which is used by system users to select the route. These characteristics of road are divided into two parts: primary characteristics and secondary characteristics. Influence of the primary characteristics in route selection is more than the secondary characteristics. The presented work uses both characteristics in route selection to achieve reliable results. The primary characteristics include travel distance, travel time, congestion, degree of difficulty, economy, facilities, user friendly and driver stress level. These eight characteristics are widely used in route selection while other are secondary characteristics. They may include- number of turns, traffic signals, road condition, unnecessary detours, traffic incidents, weather, catastrophe, bridge clearance, lane width, bridge load, speed limit etc. Table III shows the importance of road attributes for transportation mode. Most of the attributes are very important for one kind of transport mode while less important for others as explained in table given below.

First two PA like travel time and travel distance are the most important attributes from all other attributes in navigational route selection system. Experience based fuzzy scores are assigned to each road attribute. The fuzzy score 1 is assigned to denote the shortest travel time on a route among the set of feasible routes and a score of 0 is used to denote the longest travel time on a route. Travel distance attribute is similar to the travel time attribute. The fuzzy score 1 of this attribute denotes the shortest travel distance among the set of feasible routes and fuzzy score 0 denotes the longer travel distance. Several causes of road congestion are accidents, road construction, traffic light etc. More congestion on road is denoted by high fuzzy score and low score represents the less congestion on road. Degree of difficulty includes narrowness of road, traffic signals, number of turns, road construction etc. which is difficult for the drive.

A fuzzy score 1 indicates ideal road situation, means easy to drive and score 0 indicates difficult to drive. For example National Highways 1, State Highways 0.8, Urban Road 0.6, Rural Road 0.3. Economy includes toll plaza, vehicle maintenance cost, fuel consumption etc. A fuzzy score 1 denotes best economy (no toll, low maintenance cost, low fuel consumption) and 0 denotes the worst economy (maximum toll, high maintenance cost, high fuel consumption). Facilities include the available facilities (filling station, automobile station, restaurant, tea shop, ATM machines etc.) on a road. A fuzzy score 1 denotes that more facilities are available on a route and 0 denotes that there is no facility on route. A score 1 denotes that route is more user friendly and low stress level of the driver and 0 score represents that the route is less user friendly and high stress level of the driver.

TABLE III. IMPORTANCE WISE CLASSIFICATION OF ROAD ATTRIBUTES FOR TRANSPORTATION MODE

	Train	**Truck**	**Bus**	**Car**	**Motor Cycle**
Primary Attributes (PA)					
Travel Distance	VM	VM	VM	VM	VM
Travel Time	VM	VM	VM	VM	VM
Congestion	I	VM	VM	I	LI
Degree of Difficulty	LI	VM	VM	I	LI
Economy	LI	VM	VM	VM	LI
Facilities	LI	I	VM	VM	VM
User Friendly	NA	NA	NA	NA	NA
Driver Stress Level	LI	VM	VM	I	I
Secondary Attributes (SA)					
Road Condition	NR	VM	VM	VM	LI
Unnecessary Detours	LI	VM	VM	I	LI
Traffic Incidents	NR	VM	VM	VM	LI
Weather	VM	VM	VM	VM	I
Catastrophe	VM	VM	VM	VM	LI
Bridge Clearance	VM	VM	VM	I	LI
Lane Width	NR	VM	VM	I	LI
Bridge Load	VM	VM	VM	I	LI
Speed Limit	VM	VM	VM	VM	VM

Legend: VM- Very Important I – Important

LI- Less Important NR- Not Related NA- Not Analysis

TABLE IV. PRIMARY AND SECONDARY ATTRIBUTES VALUES

	Route1	**Route2**	**Route3**	**Route4**
Primary Attributes (PA)				
Travel Distance (km)	80	90	85	75
Travel Time (min)	90	110	140	160
Congestion	No	High	Low	No
Degree of Difficulty	No	Minor	Major	Minor
Economy	Very Low	High	Very High	Low
Facilities	More	Few	No	No
User Friendly	More	Less	Less	Medium
Driver Stress Level	Low	Medium	High	Very High
Secondary Attributes (SA)				
Road Condition	Best	Good	Average	Worst
Unnecessary Detours	Few	Medium	More	Few
Traffic Incidents	Medium	More	More	Few
Bad Weather	More Suitable	Suitable	Not Suitable	Not Suitable
Catastrophe	Low	Medium	High	Medium
Bridge Clearance (ft.)	≥ 15	≥ 15	≥ 10	≥ 9
Lane Width (ft.)	≥ 14	≥ 11	≥ 9	≥ 9
Bridge Load (pound)	9000	5000	5000	3000
Speed Limit (kmph)	100	80	60	50

The acceptable values for some other road attributes (bridge clearance, lane width, bridge load and speed limit) are ≥ 15 ft., ≥10 ft., 9000 pounds and ≥ 40 kmph respectively.

The routes which satisfy these conditions will be assigned high score, otherwise low score will be given to them.

IV. OPTIMAL ROUTE SELECTION CRITERIA

As discussed in previous section, a driver can select a route based on primary and secondary road attributes. Two types of the query systems (General and Special) are designed here. In general query system the results are returned on the basis of primary road attributes whereas in special query, the results are returned on the basis of both the road attributes.

The experience based weight is assigned to each road attributes as shown in Table V. These weights are assigned on the basis of road attribute values as given in Table IV. The fuzzy logic is used to select the optimum route for designed system as shown in Fig. 4. In this section a decision making route selection methodology for drivers is presented to select the optimum route for vehicle navigation system using fuzzy logic which is shown in Fig. 4. At first setting of the number of road attributes (primary and secondary) is done then the values of each attributes which are normalized between 0 and 1 are noted, after that classification of road attributes is made for transportation mode. Set the fuzzy rules to each road attribute and assign the experience based fuzzy score to each road attribute.

Rank is assigned to each feasible route on the basis of attribute score. The two input membership function f(x) and g(x) are represented in Fig. 5. Here f(x) is used to show “LOW”, ”SHORT”, ”MINIMUM”, ”FEW” and ”MINOR” and g(x) shows “HIGH”, ”LONG”, ”MAXIMUM”, ”MORE”, ”MAJOR”. Five trapezoidal membership functions- “VERY BAD”, “BAD”, “FAIR”, ”GOOD” and “VERY GOOD” as shown in Fig. 6. An example of primary and secondary attributes based fuzzy rules is as follows:

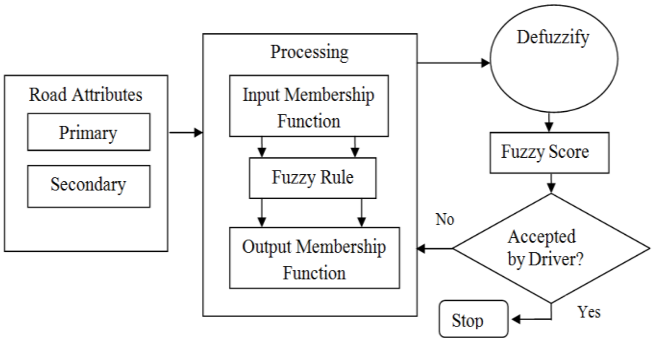


Fig. 4. Optimum route selection methodology

Primary Attribute based Fuzzy Rules as Follow:
Rule 1:

IF Travel distance is SHORT on Route1 AND
Travel distance is LONG on Route2
THEN Choose Route1 AND Route2 should not be taken

TABLE V. FUZZY SCORE OF ROAD ATTRIBUTES

	Route1	Route2	Route3	Route4
Primary Attributes (PA)				
Travel Distance			0.6	
Travel Time	0.8	0.5	0.6	0.9
Congestion	1.0	0.9	0.8	0.4
Degree of Difficulty	1.0	0.1	0.1	0.5
Economy	1.0	0.9	0.4	0.7
Facilities	0.9	0.2	0.0	0.7
User Friendly	1.0	0.4	0.2	0.0
Driver Stress Level	1.0	0.2	0.7	0.5
	1.0	0.8		0.2
PA Average Score			0.4	
	0.9	0.5		0.4
Secondary Attributes (SA)				
Road Condition	1.0	0.8	0.5	0.1
Unnecessary Detours	0.9	0.5	0.2	0.9
Traffic Incidents	0.5	0.2	0.2	0.8
Bad Weather	0.9	0.6	0.0	0.0
Catastrophe	0.8	0.6	0.4	0.6
Bridge Clearance	1.0	1.0	0.7	0.5
Lane Width	1.0	0.8	0.4	0.4
Bridge Load	1.0	0.6	0.6	0.3
Speed Limit	0.9	0.8	0.6	0.4
SA Average Score	0.8	0.6	0.4	0.4
Total Average Score (PA + SA)	0.8	0.5	0.4	0.4

Rule 2:

IF Travel time is LARGE on Route1 AND
Travel time is SHORT on Route2
THEN Choose Route2 AND Route1 should not be taken

Rule 3:

IF Distance is LONG and Congestion is LOW on Route1 AND
Distance is SHORT and Congestion is HIGH on Route2
THEN Choose Route1 AND Route2 should not be taken

Secondary Attribute based Fuzzy Rules as Follow:

Rule 1:

IF	Unnecessary Detours is FEW on Route1 AND Unnecessary Detours is MORE on Route2
THEN	Choose Route1 AND Route2 should not be taken

Rule 2:

IF	Catastrophe is HIGH on Route1 AND Catastrophe is LOW on Route2
THEN	Choose Route2 AND Route1 should not be taken

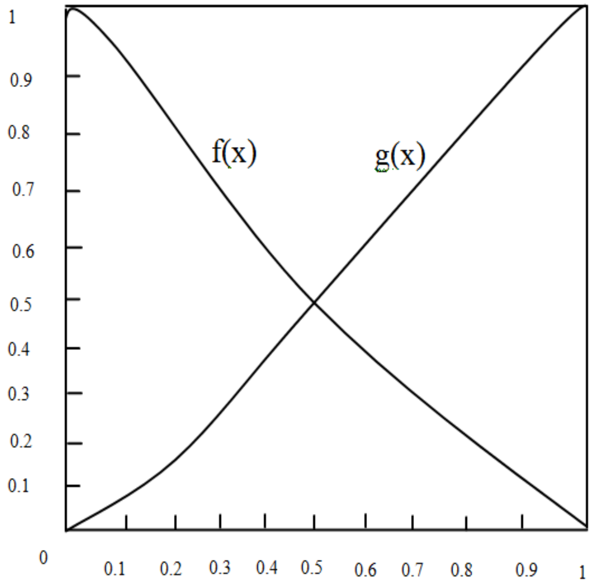


Fig. 5. Input membership functions f(x) and g(x) here f(x) for ("LOW", "SHORT", "MINIMUM", "FEW", "MINOR") and g(x) for ("HIGH", "LONG", "MAXIMUM", "MORE", "MAJOR")

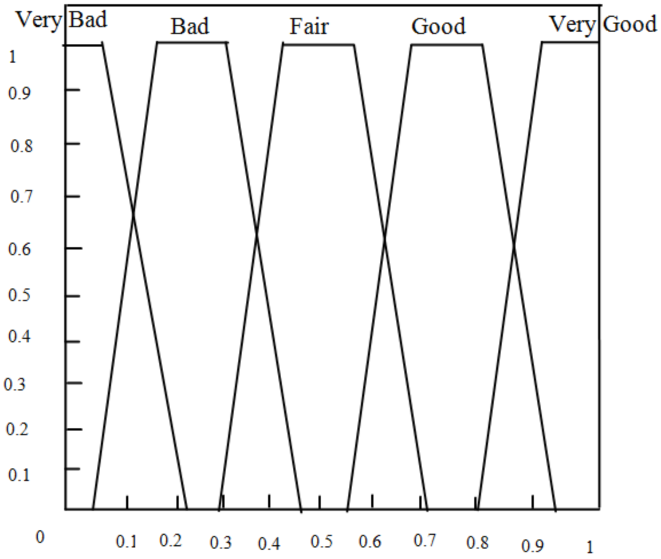


Fig. 6. Output Trapezoidal membership functions for ("VERY BAD","BAD","FAIR","GOOD","VERY GOOD")

V. Result and Discussion

A route selection criterion is based on fuzzy concept. The designed system is tested over Haryana route and alternate routes are considered from origin to destination. The program for developed system is written in programming language (PHP, MYSQL) and it is designed for two types of queries (general and special). In case of general query, the designed system returned the results on the basis of average score of PA and in case of special query; the average score of (PA + SA) is used to display results as shown in Table VI. The system users can view all alternate routes from origin to destination just by entering starting and ending point of journey as well as they can select optimal route for their journey. Four feasible routes and their descriptions are given here in Table I and Table V. Route1 and Route2 could be a route with heavy use of National Highway and the average score of primary and secondary attributes is 0.8 on Route1 as shown in Table VI which is the largest score among four feasible routes.

Fuzzy rules are used to specify the optimal route:

Rule 1:

IF	Travel distance is SHORT on Route1 AND Travel distance is LONG on Route2
THEN	Route1 is VERY GOOD and Route2 is very BAD

Rule 2:

IF	Distance is LONG and Congestion is LOW on Route1 AND Distance is SHORT and Congestion is VERY HIGH on Route2
THEN	Route1 is GOOD and Route2 is VERY BAD

TABLE VI. Average Score of PA, SA and (PA+SA) Corresponding Routes

	PA Score	SA Score	(PA+SA) Score
Route1	0.9	0.8	0.8
Route2	0.6	0.5	0.5
Route3	0.4	0.4	0.4
Route4	0.4	0.4	0.4

It's quite clear that Route1 will be optimal route for general and special types of driver's queries. The average score of PA, SA attribute is 0.9, 0.8 respectively of Route1 and average score of (PA+SA) is 0.8 which is highly fuzzy score among all alternate routes, so it is recommended route for the driver. Route2 has the next highest score it will be the next recommended route. The designed system provides reliable result as compared to other existing system because almost all road attributes are considered here to assign score to each route.

VI. Conclusion and Future Work

An intelligent route selection system is presented in this article. The system provides the optimum route for particular drivers to his/her preference. The concept of primary and secondary road attributes is discussed and a criterion for optimum route selection is set. The route selection criteria is based on fuzzy logic in which output Trapezoidal membership functions for Route1, Route2, Route3 and Route4 are " GOOD", "FAIR", "BAD" and "BAD" respectively. The first output membership function shows the optimal route among all alternate routes. This vehicle navigation system will be helpful for drivers in decision making to select more precise optimal route. In future ORSVNS will be a part of more intelligent route selection system for transportation.

References

[1] Grantham K.H. Pang, K. Takahashi,T. Yokota and H. Takenaga," Adaptive Route Selection for Dynamic Route Guidance System Based on Fuzzy-Neural Approaches", IEEE Transactions on Vehicular Technology, Vol. 48, no. 6, November 1999.

[2] A brief report on Auto and Auto Ancillaries in India, ASA & Associates LLP, July2015.

[3] Dunn Engineering Associates, P.E., Consulting Services," Alternate Route Handbook", U.S Department of Transportation, Federal Highway Administration, May 2006.

[4] Peter Bonsall," The influence of route guidance advice on route choice in urban networks", ©Kluwer Academic Publishers, Springer, Vol. 19(1), pp. 1-23, 1992.

[5] Mohamed Abdel-Aty and M. Fathy Abdalla," Modeling drivers' diversion from normal routes under ATIS using generalized estimating equations and binomial probit link function", Kluwer Academic Publishers. Printed in the Netherlands, Vol. 31(3), pp. 327–348, 2004.

[6] Alexander Aved, Tai Do, Georgiana Hamza-Lup, Ai Hua Ho, Lap Hoang, Liang Hsia, Kien A. Hua, Fuyu Liu, Rui Peng," A real-time route diversion management system ",Intelligent Transportation Systems Conference, IEEE, pp. 1131-1136, 2007.

[7] M. Shou-Ren HU, Chung-Yung WANG, Chih-Peng CHU and Ken-Chen LEE," Value of traffic information for route diversion control scheme under traffic incidents ",Journal of the Eastern Asia Society for Transportation Studies, Vol. 6, pp. 2487 - 2501, 2005.

[8] Kim, Dong Sun," A dynamic route diversion model on urban freeway O-D pattern predicton", Journal of the eastern Asia Society for Transportation studies, Vol.5, October 2003.

[9] Constantinos Antonioua, Haris N. Koutsopoulos, Moshe Ben-Akiva and Akhilendra S. Chauhan," Evaluation of diversion strategies using dynamic traffic assignment", Transportation Planning and Technology Vol. 34(3), pp. 199-216, April 2011.

[10] Hey Ran Kim and Kyung Soo Chon," Modeling En-route diversion behavior under on-site traffic information", Journal of the eastern Asia society for transportation studies, Vol. 6, pp. 1833-1843, 2005.

[11] Anastasia Spiliopoulou, Maria Kontorinaki, Ioannis Papamichail and Markos Papageorgiou," Real-Time Route Diversion Control at Congested Motorway off-Ramp Areas - Part I: User Optimum Route Guidance", Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, October 6-9, 2013.

[12] A. Spiliopoulon, M. Kontorinaki, I. Papamichail and M. Papaageorgious," Real-time route diversion control at congested off-Ramp areas partII: Route guidance versus off-ramp closure", Transportation: can we do more with less resources? transportation. Procedia - social and behavioral sciences. Elsevier Ltd., Vol. 111(5): pp. 1102–1111, 2013.

[13] Xingang Li, Yakang Cao, Xiaomei Zhao, and Dongfan Xie," Drivers' Diversion from Expressway under Real Traffic Condition Information Shown on Variable Message Signs ",KSCE Journal of Civil Engineering, Vol. 9(7), pp. 2262-2270, 2014.

[14] Mohd Rapik Saat and Jesus Aguilar Serrano," Multicriteria high-speed rail route selection: application to Malaysia's high-speed rail corridor prioritization", Transportation Planning and Technology, Taylor & Francis,Vol. 38(2), pp. 200–213, 2015.

[15] S.S. Keshkamat, J.M. Looijen and M.H.P. Zuidgeest," The formulation and evaluation of transport route planning alternatives: a spatial decision support system for the Via Baltica project, Poland", Journal of Transport Geography, Elsevier Ltd., Vol. 17, pp. 54–64, 2008.

[16] Grantham K.H. Pang, K. Takahashi,T. Yokota and H. Takenaga," Intelligent Route Selection for In-Vehicle Navigation Systems", Transportation Planning and Technol., Vol. 25 (3), pp. 175–213, 2002.

Skip List Data Structure Based New Searching Algorithm and Its Applications: Priority Search

Mustafa Aksu
Department of Computer Technologies
Vocational School of Technical Sciences,
Sutcu Imam University
Kahramanmaras, Turkey

Ali Karcı
Department of Computer Engineering
Faculty of Engineering, Inonu University
Malatya, Turkey

***Abstract*—Our new algorithm, priority search, was created with the help of skip list data structure and algorithms. Skip list data structure consists of linked lists formed in layers, which were linked in a pyramidal way. The time complexity of searching algorithm is equal to O(lgN) in an N-element skip list data structure. The new developed searching algorithm was based on the hit search number for each searched data. If a datum has greater hit search number, then it was upgraded in the skip list data structure to the upper level. That is, the mostly searched data were located in the upper levels of the skip list data structure and rarely searched data were located in the lower levels of the skip list data structure. The pyramidal structure of data was constructed by using the hit search numbers, in another word, frequency of each data. Thus, the time complexity of searching was almost Ɵ(1) for N records data set. In this paper, the applications of searching algorithms like linear search, binary search, and priority search were realized, and the obtained results were compared. The results demonstrated that priority search algorithm was better than the binary search algorithm.**

Keywords—Algorithms; Priority search; Algorithm analysis; Data structures; Performance analysis

I. Introduction

Various disciplines in computer sciences benefit from algorithms and data structures directly or indirectly, and different data structures were used as solutions to various problems. The limitations like processing, time complexity, required hardware or inefficiency of current algorithms conclude in defining new algorithms such as searching, sorting, and graph algorithms [1].

Sometimes, an algorithm was preferred on another one because of its processing, time complexity, etc. For example, binary search was preferred instead of sequential search to increase searching complexity. Skip list data structures based searching algorithm presented in this study is another option instead of binary search. Considering these factors, it is evident that new algorithms and data structures will continue to emerge as needed [2].

In computer science, the linked list is a data structure consisting of a group of nodes, which together represent a sequence (Fig. 1). The principal benefit of a linked list over a conventional array is that in the linked list elements can easily be inserted or removed without reallocation or reorganization of the entire structure, because the data items need not to be stored contiguously in memory or on disk. Linked lists allow insertion and removal of nodes at any point in the list, and can do so with a constant number of operations if the link previous to the link being added or removed was maintained during list traversal. Linked lists by themselves do not allow random access to the data, or any form of efficient indexing. Thus, many basic operations may require scanning most or all of the list elements [3], [20]. The time complexity of linked list is linear, so, the time complexity of searching in linked list of size N is O(N) [15], [19].

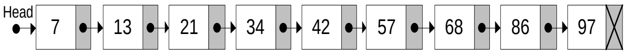


Fig. 1. Linked list

In this study, a new searching algorithm based on the skip list was developed and it was compared to other searching algorithms by doing some applications. The rest of paper was arranged as follows: Related works have been presented in Section II. The skip list data structure must be clarified for the sake of the understandability of developed searching algorithm. Due to this case, Section III explains the skip list data structure. The methodology of proposed algorithm has been explained in Section IV and Section V demonstrates the experimental results and significance of work. The conclusion has been given in Section VI.

II. Related Works

Skip list data structure, which was introduced by Pugh [8], is a data structure alternative to binary tree search structure. Search, insertion and deletion algorithms of nodes in skip list data structure is discussed in article written by Pugh [8]. The time complexity of searching in the skip list data structure is O(lgN). In addition, several studies have been conducted so far on the improvement and analysis of skip list data structure algorithms. In [2], how randomly creation of levels and different “P” thresholds (0.25, 0.5, 0.75) effect the performance was studied and solutions were proposed.

An optimized search algorithm for skip lists was analyzed in [6]. In [7], the probabilistic analysis of the search cost was considered in a slightly different way, namely, performing the asymptotic analysis of the total search cost or path length.

In [12], proposed exploring techniques based on the notion of a skip list to guarantee logarithmic search, insert and delete costs. The basic idea is to insist on that between any pair of elements above a given height are a small number of elements of precisely that height.

Other studies are about level optimization in skip list data structure [1], formal verification of a lazy concurrent list-based set [4], a simple optimistic skip list algorithm [5], average search and update costs in skip lists [9], skip lists and probabilistic analysis of algorithms [10], the binomial transform and the analysis of skip lists [11], deterministic skip lists [12], a skip lists cookbook [13], and concurrent maintenance of skip lists [14].

Various data structures and algorithms were also created apart from skip list data structure such as Tiara: A self-stabilizing deterministic skip list and skip graph [22],Corona: A Stabilizing Deterministic Message-Passing Skip List [23] and Skip lift: A probabilistic alternative to red–black trees [24].

III. Skip List Data Structure

Linked lists were used in skip list data structure and it aimed to facilitate searching, insertion and deletion through placing elements in a pyramid-like order at different levels. In this data structure, elements were placed at different levels randomly.

First, all nodes were placed at level 0 and, starting from left row and skipping each 2^ith node (i=0,..,MaxLevel (15 or 31)), pointers representing each level are created towards the top. The list at level 0 is the linked list at the bottom in skip list data structure and encompasses all nodes. Each list from bottom to the top were arranged as an index of the previous list [1], [19] (Fig. 2).

When levels in skip list data structure were created (level 0, level 1,.., level k), it was done randomly (Pugh's random Level algorithm [8]; for P=1/4). Let us say that the number of ordered nodes in skip list data structure is N. Level 0 consists of these entire N ordered nodes (Fig. 4- Level 0).

Level 1 is created if every other element of the list at Level 0 has also an extra link to the element four ahead of it (Fig. 4 – Level 1). Since the maximum number of elements at Level 1 level equals to $\left\lceil \frac{N}{4} \right\rceil + 1$, so on, the data structure will be constructed.

The height of skip list depends on the probability P threshold value given in Pugh's "random Level algorithm". The effects of P threshold values were studied in a previous study [2] and skip list is more efficient when P threshold value is equal to 1/4. While if P=1/2, the height of skip list approaches to height of balanced tree (lgN). If P=1/4, one out of every four nodes in Level 0 copied to Level 1 (an upper level), and this process was continued in the same way until all data structure were constructed. This process resulted in the height of skip list will be half of the height of the balanced tree. These cases are seen in Fig. 2 and Fig. 4 respectively.

A group of data consisting of the elements {zinc, bee, fox, hill, dive, lift, null, total, vary, other, see} on a skip list shown in Fig. 2. The true skip list structure, which was constituted from these elements, is shown in Fig.3.

Time complexity is O(N) for search, insertion and deletion processes when linked and ordered lists are used. On the other hand, the time complexity is O(lg N) in skip list data structure [8], [15] when the same process were performed.

In a search algorithm, a node was searched from upper levels to lower levels. During insertion, first, the node to be inserted was searched. If not found, new value is inserted to the matching location starting from a random level and pointers and lists are updated. The process was repeated for other levels where a node is to be inserted. Search was performed from the top level to lower levels for removal operations. The node was deleted when found and pointers and lists were updated. The process was repeated on other levels where the node is available [2].

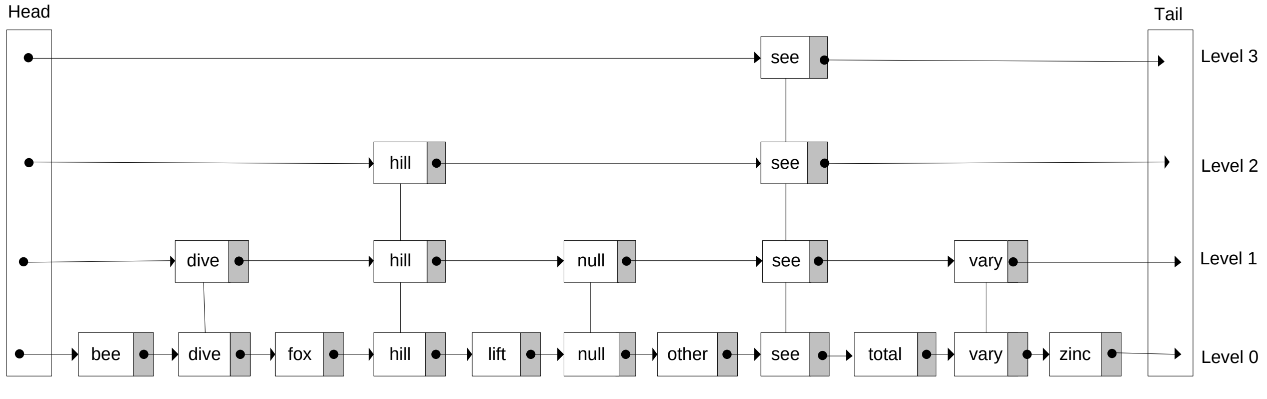


Fig. 2. Skip list (for P=1/2)

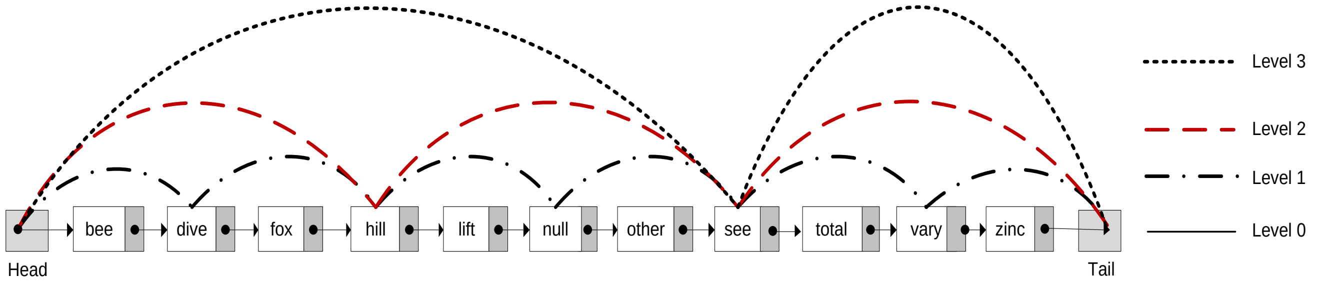


Fig. 3. Skip list (Real structure of skip list for Fig. 2)

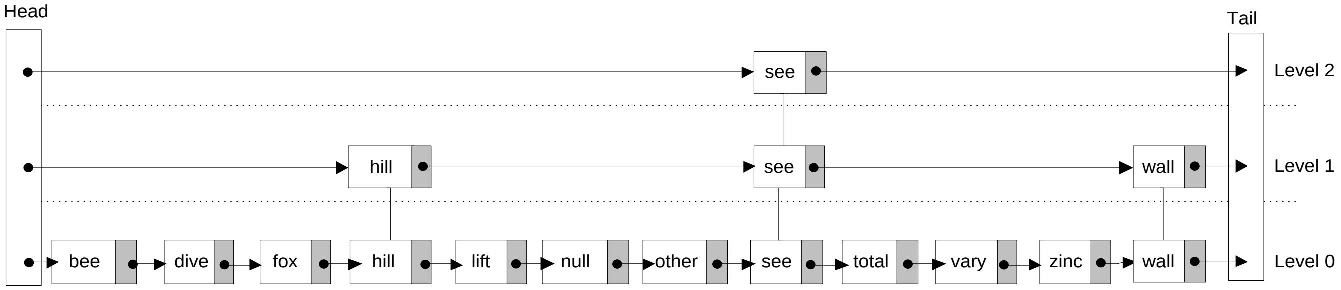


Fig. 4. Skip list (for P=1/4)

IV. Priority Search and Binary Search

The innovative search algorithm which was called priority search uses the skip list data structure. It was benefited from the pyramidal layered-structure of the skip list data structure. The standard searching algorithm (algorithm 1) in the skip list data structure starts at top-level to the lowest level until it finds the searching data or it ends up in the lowest level. The developed new searching algorithm (Algorithm 2) was based on the hit search number for each searched data. If a datum has greater hit search number, then it was upgraded in the skip list data structure to upper level. That is, the mostly searched data were located in the upper levels of the skip list data structure and rarely searched data were located in the lower levels of the skip list data structure. The time complexity of searching in the skip list data structure (Algorithm 1) is O(lgN), but the time complexity of searching algorithm in priority search (Algorithm 2) approximates to Θ(1). In another word, the mostly searched data were located in the top-level of the skip list data structure, thus, the searching for these data has time complexity as ~Θ(1). The rarely searched data were located in the lowest level and their searching time complexities approximate to O(lgN). The time complexity of searching by using priority search algorithm changes between Θ(1)-O(lgN).

When ‘dive’ two times, ‘null’ four times and ‘vary’ three times were searched as in Table I, the results in Table II will be obtained. The skip list data structure for data in Table II is seen in Fig. 5, in which priority search algorithm (Algorithm 2) was used. It was performed by using frequencies (hit search numbers). That is, the searched data is upgraded once for each search process. Therefore, the mostly searched data were located at the top of skip list data structure (pyramidal structure) and rarely searched data were located at the bottom of skip list data structure.

```
Algorithm 1 {Search in skip list }
SearchNode(slist, key)
 HEAD ←slist→head
 LEVEL ←slist→level
 if (HEAD→next[0] = NULL) or (LEVEL<0)
     return false
 for i ← LEVEL downto 0 do
  while(HEAD→next[i]≠NULL
         and HEAD→next[i]→value < key)
  HEAD ←HEAD→next[i]
  HEAD ←HEAD→next[0]
  if (HEAD ≠ NULL and HEAD→value = key)
      return true;
 return false;
```

TABLE I. Frequency-Wise Level Distribution Of Nodes On Fig. 4

nodes	bee	dive	fox	hill	lift	null
frequency	0	0	0	1	0	0
level	0	0	0	1	0	0
nodes	**other**	**see**	**total**	**vary**	**zinc**	**wall**
frequency	0	2	0	0	0	1
level	0	2	0	0	0	1

The priority search algorithm was used in the skip list data structure due to its pyramidal structure. Additionally, the standard search algorithm (Algorithm 1) for the skip list of size N has time complexity as O(lgN). Data were sorted in ascending order in the skip list data structure when skip list data structure were constructed (Fig. 5 Level0, Level1, Level2, Level 3, and Level 4). The most important property of skip list data structure is its pyramidal structure and ordered data in it.

The searching process started at the first element in the list and carried on till the end of list, when data were unordered.

So, the searching algorithm is a linear algorithm in term of the number of data in the list. The time complexity of linear search is O(N). The searching process considered the data as unordered whether data were ordered or not. But it is not a suitable search process for ordered data [16], [17], [21].

TABLE II. FREQUENCY-WISE LEVEL DISTRIBUTION OF NODES ON FIG. 5

nodes	bee	dive	fox	hill	lift	null	other	see	total	vary	zinc	wall
frequency	0	2	0	1	0	4	0	2	0	3	0	1
Level	0	2	0	1	0	4	0	2	0	3	0	1

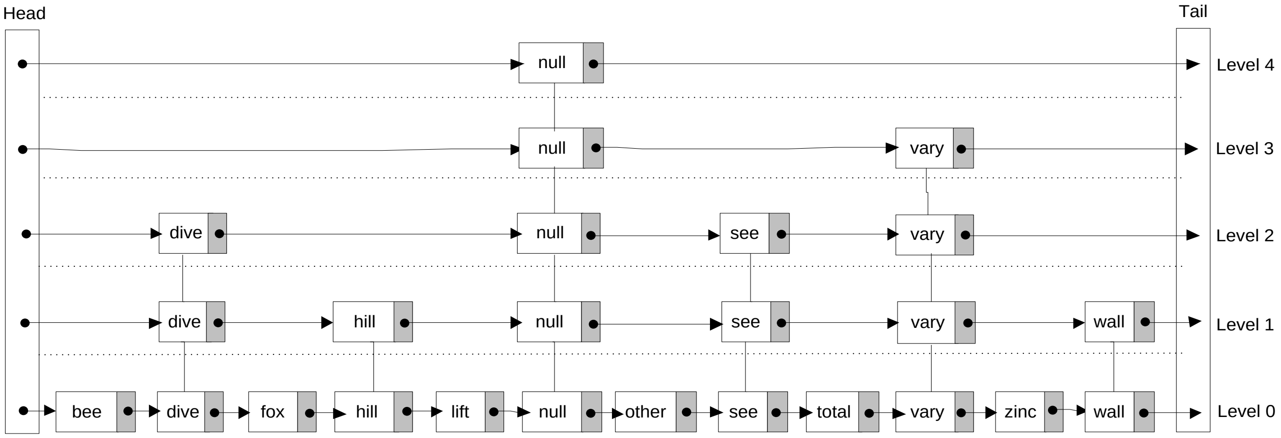


Fig. 5. Obtained Priority Search schemes for searching 'dive' two times, 'null' four times and 'vary' three times on Fig. 4)

```
Algorithm 2 {Priority search}
PrioritySearch(slist, search_value)
 HEAD←slist→head ,
 LEVEL←slist→level
 update[MaxLevel +1]
 while (LEVEL>=0)
 if(HEAD->next[LEVEL]->value=search_value)
  for i ← LEVEL downto 0 do
    while (HEAD→next[i] ≠ NULL and
      HEAD→next[i]→value <search_value)
    HEAD←HEAD→next[i]
    update[i] ← HEAD
   end for
  HEAD←HEAD→next[0]
  intlvl = LEVEL+1;
   if(lvl>slist->level)
     update[lvl] = slist->head
     slist->level = lvl
   end if
  HEAD->next[lvl] = update[lvl]->next[lvl]
  update[lvl]->next[lvl] = HEAD;
 return true
 end if
 if(HEAD->next[LEVEL]->value<search_value)
    HEAD= HEAD->next[LEVEL]
 if(HEAD->next[LEVEL]->value>search_value)
    LEVEL=LEVEL-1
end while
return false;
```

Another searching algorithm is binary search algorithm for ordered data. In order to use this algorithm, data have to be ordered on the list. If data were unordered, initially they must be ordered by using any sorting algorithm.

The mechanism of binary searching algorithm is as follows [15], [16], [18], [20]:

- If list or array is not sorted, it is firstly sorted.
- Sorted array is divided into two equal sub-arrays or approximately equal sub-arrays.
- The searched data is compared with the middle element of array. If it is equal then, it is found. If searched data is less than the middle element of array, then right sub-array is discarded and data will be searched in the left sub-array. If searched data are greater than the middle element of array, then searched data will be searched in the right sub-array.
- The searched data will be scanned on the left or right sub-array in the same manner.
- The process goes on in the same manner until searched data is found or search is terminated.

The time complexity of binary search algorithm is O(lgN) for an array of size N elements. Moreover, the time complexity of binary search for balanced binary trees is also O(lgN).

V. EXPERIMENTAL RESULTS: PRIORITY SEARCH AND OTHERS

The proposed algorithm was implemented by using C++ and tested successfully on distinct arrays. In order to compare Priority Search (PS), Linear Search (LS), and Binary Search (BS), random arrays and sorted arrays were used. The searching times of PS, LS and BS for sizes from 1000 to 100000 of arrays were illustrated in the Table III and Table IV. Moreover, each algorithm was applied to same size arrays 100 times and all times for all executions was added up and then their average was computed. This means that the effect of data permutation will be minimized and the comparison will be more equitable. If there is one search for algorithm, the comparisons may be non-equitable. For example, searched data for PS may be on the top level of skip list data structure, and then its time will be Θ(1). If the searched data for BS is not found in the binary search tree, then its time will be longer. This case may be available for each search algorithm. Due to this case, there were 100 executions for equitable comparisons of search algorithms.

All results were obtained on the same computer and the results in Table III and Table IV demonstrated that when size of array is small, BS shows normal performance; when the size of array increases, the performance of PS increases and PS is better than LS and BS. The results were illustrated in Fig. 6.

TABLE III. SEARCHING TIMESFOR LS, BS AND PS FOR SORTED ARRAYS (IF THE SEARCHED DATA ARE NEAR TO THE BEGINING OF ARRAY) (MS=MILLISECOND)

# of nodes	1000	5000	10000	30000	50000	100000
LS	0.0032 ms	0.0103 ms	0.0167 ms	0.0374 ms	0.0671 ms	0.1382 ms
BS	0.00015 ms	0.00018 ms	0.00020 ms	0.00022 ms	0.00023 ms	0.00025 ms
PS	0.00009 ms	0.00011 ms	0.00013 ms	0.00016 ms	0.00017 ms	0.00019 ms

Table III, Table IV and Fig. 6, Fig. 7 depict that PS is better than LS and BS with respect to searching time. The time complexities for searching PS, and BS on sorted arrays are O(lgN). The time complexities for searching LS on sorted array is O(N). While computing time complexity for any algorithm, the dominant (term with the greatest degree) term is regarded as time complexity. The asymptotic behaviors of PS and BS are similar; however, the constant coefficients are different and this case makes PS be the best algorithm.

It is noticeable in Table III and Table IV; PS algorithm has better performance than LS and BS. Moreover, PS algorithm is better than BS algorithm as seen in Fig. 7. Searched data in PS algorithm were located to the top of Skip List, hence time complexity will be Θ(1) for these data.

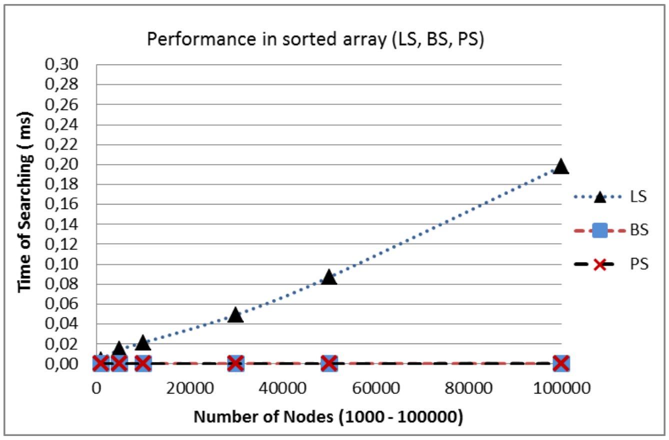


Fig. 6. Performance comparison for LS, BS, PS (If the searched data are in middle of array)

TABLE IV. SEARCHING TIMES FOR LS, BS AND PS FOR SORTED ARRAYS (IF THE SEARCHED DATA ARE NEAR TO THE END OF ARRAY) (MS=MILLISECOND)

# of nodes	1000	5000	10000	30000	50000	100000
LS	0.0047 ms	0.0171 ms	0.0327 ms	0.0858 ms	0.1471 ms	0.2876 ms
BS	0.00012 ms	0.00014 ms	0.00017 ms	0.00020 ms	0.00022 ms	0.00025 ms
PS	0.00008 ms	0.00010 ms	0.00012 ms	0.00015 ms	0.00017 ms	0.00020 ms

The results in Table III were obtained when the searched data were located near to the beginning of array. Whereas, Table IV shows the situation where the searched data were located near to the end of the array. Comparing the results of LS algorithm in both tables, it was seen that the search time increases if the data were located at the end of array. However, the results were the same for BS and PS algorithms no matter where the searched data was located.

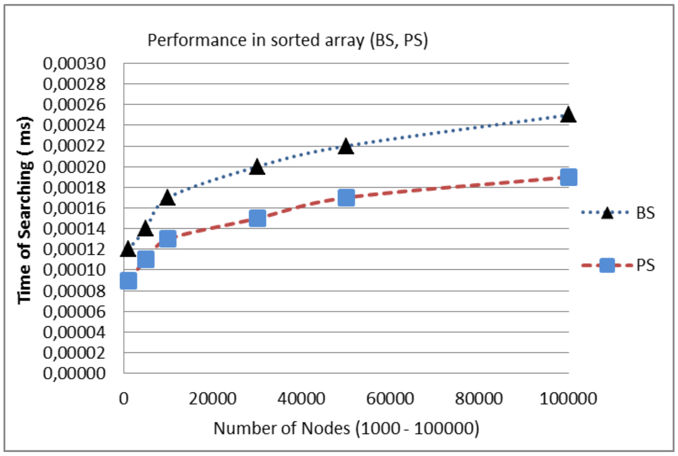


Fig. 7. Performance comparison for BS and PS (Sorted arrays)

When arrays are unsorted, the performance of linear search algorithm is better than the other algorithms, since remaining algorithms require the sorted arrays to show better performances.

A. Significance of work

Priority search algorithm locates the most searched data to the top of the pyramid-shaped skip list data structure. For these reason, enabling time complexity Θ(1) of frequent searched data were important.

The priority search algorithm may be used in the search engine like Google, Yandex, etc. The greater frequency (search hit number) the upper level for searched data; the smaller frequency the lower level for searched data. The mostly searched data were located in the top level of skip list data structure, so, searching this data will take less time. The rarely searched data were located in the lowest level of the skip list data structure, so, its searching time will take longer. If searching process was grouped with respect to frequencies of data, the searching would be easier. There many data (may be billion data, etc.) in the internet. If data were located in a large skip list data structure for search engine, it would be more advantageous.

This data structure is also advantageous for dictionary operations, since the most hit data will be on the top level of skip list data structure and its searching will take shorter time; the least hit data will be on the lowest level of the skip list data structure and its searching time will take longer time.

VI. Conclusion

Skip list data structure was created with the help of linked list data structures. Thanks to its layered structure, skip list data structure presented in this study reduces the time complexity of search, insertion and deletion processes in linked list data structure to O(lgN), which was O(N).

The applications of linear search, binary search and priority search were realized, and obtained results were compared. The obtained results verified that priority search was better than the linear search and binary search considering the applications. Priority search superior than binary searching and linear searching due to its application results. The time complexity of priority search algorithm was between Θ(1)-O(lgN); the most searched data has time complexity as Θ(1), the least searched data has time complexity as O(lgN).

To summary priority search algorithm could be used in searching processes more efficiently. It enables saving remarkable time when larger sets of data were handled.

References

[1] M. Aksu, A. Karcı, and Ş. Yılmaz, “Level optimization in Skip List data structure,” in Proc. 1ST International Symposium on Innovative Technologies in Engineering and Science (ISITIES2013), 2013, pp. 389-396.

[2] M. Aksu, A. Karcı, and Ş. Yılmaz, “Effects of P Threshold Values in Creation of Random Level and to the Performance of Skip List Data Structure,” BitlisEren University Journal of Science, Vol. 2, No. 2, 2013, pp. 148-153.

[3] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, Introduction to Algorithms, MIT Press, 2009.

[4] R. Colvin, L. Groves, V. Luchangco, and M. Moir, “Formal verification of a lazy concurrent list-based set,” in Proc. Computer Aided Verification, Lecture Notes in Computer Science, Vol. 4144, 2006, pp. 475-488.

[5] M. Herlihy, Y. Lev, V. Luchangco, and N. Shavit, “A Simple Optimistic Skiplist Algorithm,” in Proc. Structural Information and Communication Complexity, Lecture Notes in Computer Science, Vol. 4474, 2007, pp. 124-138.

[6] P. Kirschenhofer, C. Martinez, and H. Prodinger, “Analysis of an optimized search algorithm for skip lists,” Theoretical Computer Science, Vol. 144, 1995, pp. 199-220.

[7] P. Kirschenhofer, and H. Prodinger,”The path length of random skip lists,” ActaInformatica, Vol. 31, No. 8, 1994, pp. 775-792.

[8] W. Pugh, “Skip Lists: A Probabilistic Alternative to Balanced Trees,” Communications of the ACM, Vol. 33, No. 6, 1990, pp. 668-676.

[9] T. Papadakis, J. I. Munro, and P. V. Poblete, ”Average search and update costs in skip lists,” BIT, Vol. 32, 1992, pp. 316-332.

[10] T. Papadakis, “Skip lists and probabilistic analysis of algorithms,” PhD Thesis, University of Waterloo, Tech. Report CS-93-28, 1993.

[11] P. V. Poblete, J. I. Munro, and T. Papadakis,“The binomial transform and the analysis of skip lists,” Theoretical Computer Science, Vol. 352, 2006, pp. 136-158.

[12] J. I. Munro, T. Papadakis, and P. V. Poblete, “Deterministic Skip Lists,” in Proc. SODA '92 Proceedings of the third annual ACM-SIAM symposium on Discrete algorithms, 1992, pp.367-375.

[13] W. Pugh, “A Skip List Cookbook,” Dept. of Computer Science, University of Maryland, College Park, Technical report, CS–TR–2286.1, 1990.

[14] W. Pugh, “Concurrent Maintenance of Skip Lists,” Dept. of Computer Science, University of Maryland, College Park, Technical report,TR–2222.1, 1989.

[15] M. T. Goodrich, and R. Tamassia, Algorithm Design and Applications, Wiley, America, 2014.

[16] M. J. Dinnen, G. Gimel’farb, and M. C. Wilson, Introduction to Algorithms, Data Structures and Formal Languages, Pearson Education, Second edition, 2009.

[17] R. Sedgewick, and K. Wayne, Algorithms, Addison Wesley, Fourth Edition, America, 2011.

[18] K. Mehlhorn, and P. Sanders, Algorithms and Data Structures; The Basic Toolbox, Springer, 2007.

[19] M. McMillan, Data Structures and Algorithms Using C#, Cambridge University Press, 2007.

[20] C. A. Shaffer, Data Structures & Algorithm Analysis in C++, Dover Publications, 2011.

[21] D. E. Knuth, The Art of Computer Programming—Sorting and Searching, Volume 3, Addison Wesley, Second edition, 1998.

[22] T. Clouser, M. Nesterenko, and C. Scheideler, “Tiara: A self-stabilizing deterministic skip list and skip graph,” Theoretical Computer Science, Vol. 428, 2012, pp. 18-35

[23] R. M. Nor, M. Nesterenko, and C. Scheideler, “Corona: A Stabilizing Deterministic Message-Passing Skip List,” Theoretical Computer Science, Vol. 512, 2013, pp. 119-129.

[24] P. Bose, K. Douïeb, and P. Morin, “Skip lift: A probabilistic alternative to red–black trees,” Journal of Discrete Algorithms, Vol. 14, 2012, pp. 13–20.

Self-Organized Hash Based Secure Multicast Routing Over Ad Hoc Networks

Amit Chopra
PhD Research Scholar, CSE dept.
MMEC, M. M. University
Ambala, HARYANA, INDIA

Dr. Rajneesh Kumar
Professor, CSE dept.
MMEC, M. M. University
Ambala, HARYANA, INDIA

***Abstract*—Multicast group communication over mobile ad hoc networks has various challenges related to secure data transmission. In order to achieve this goal, there is a need to authenticate the group member as well as it is essential to protect the application data, routing information, and other network resources etc. Multicast-AODV (MAODV) is the extension of an AODV protocol, and there are several issues related to each multicast network operation. In the case of dynamic group behavior, it becomes more challenging to protect the resources of a particular group. Researchers have developed different solutions to secure multicast group communication, and their proposed solutions can be used for resource protection at different layers i.e. application layer, physical layer, network layer, etc. Each security solution can guard against a particular security threat. This research paper introduced a self-organized hash based secure routing scheme for multicast ad hoc networks. It uses group Diffie-Hellman method for key distribution. Route authentication and integrity, both are ensured by generating local flag codes and global hash values. In the case of any violation, route log is monitored to identify the malicious activities.**

Keywords—Security; Multicast; Group Communication; MAODV; Key management; HASH

I. Introduction

Multicast based communication is a vital network service, which sends the data from a source to multiple destinations simultaneously by creating copies only when the links to the destinations split. Multicast routing tree can be constructed to transmit the data from the sender to all the destinations with a minimum multicast tree cost that is used to evaluate the utilization of network resources [1]. Multicast packets are transmitted to all members of a group with the same reliability as regular unicast packets. Multicasting can reduce the cost of communication, consumption of bandwidth, sender and router processing and delivery delay [2].

A. Security issues and challenges

Multicast Ad hoc networks operate in open environment having no access constraints to the network resources. Following are security issues related to network operations:

- Secure routing issues: Use of Shared medium, open network environment, Lack of Centralized Monitoring, Limited Resources, Physical vulnerability [3].
- Network Security Issues: Confidentiality, Integrity, Availability, Non-Repudiation [3].

B. Security Attacks over Multicast Ad Hoc Networks

Multicast communication supports both unicast and multicast operations, so various security attacks can be categorized as per these activities which are given below [6][12][20]:

- Unicast Operation Attacks: Black Hole Attack, Worm Hole Attack, Sybil Attack, Flooding Attack, Routing Table Attack, DoS Attack [33]
- Multicast Operation Attacks: MACT-Attacks, Group Lead Selection Attack, Link breakage Attack, Routing Table attack [20][21].

C. Security Constraints of Multicast Group Communication

It is quite complicated to enforce security rules for group communication due to various facts i.e. Dynamic Group Behavior, Group Operations: Join/Leave, Open Communication: Inside/Outside the Group etc. Security requirements should consider Group Member Authentication, secure key distribution/management, detection/prevention of any threat to the group or entire network, mobility and scalability, etc. Following are the different categories of security provision for group communication:

- Key-based Secure Multicast Routing: With this kind of security provision, any key distribution method based on cryptographic algorithms can be used to secure the group communication. Each group member exchanges the keys for communication. Key distribution faces some issues like key pair production, distribution and management, etc. Key-based communication becomes more challenging for scalable and dynamic groups. Shared Keys are used for node Identification purpose, called authentication which may be quite time-consuming process and its performance depends upon the number of keys to be processed and the number of participants involved in group communication [29][30][31[32].
- Key Generation: It is essential to secure data transmission over open shared medium which may have several security threats. Cryptography is a process that is used to secure data. One can use any cryptography method (Symmetric/Asymmetric) to maintain the level of confidentiality and integrity [11].
- Key Distribution: Key distribution is a process which assigns the generated keys to each node. There are

different ways for Public/Private Key generation and distribution [32]. After key pair generation, it is also essential to distribute them in a secure way to legitimate group members only. Key distribution process can be performed using key agreement protocols. For secure communication, one can use Diffie–Hellman/Needham–Schroeder protocol for symmetric key distribution or any certification authority for asymmetric key distribution [29].

- Key Management: For key pair management, one can use centralized, decentralized or distributed approach. All key management schemes enforce the rules for secure group communication by utilizing the basic concept of cryptography [32].

D. Intrusion Detection and Prevention

In this case, any unauthorized access to network resources can be referred as intrusion or attack and the methods those can identify its symptoms, called intrusion detection tools [21]. Intrusion detection depends upon various factors i.e. routing information, packet drop, extra control overhead, unavailability of services, flooding, over consumption of network resources, signal jamming, etc. Following are the categories security threats:

- Active attack: An intruder can directly modify the resources, and it can be detected.
- Passive attack: The Intruder just analyzes the network information without any alteration. Captured data may be further utilized to trigger another type of attack [3], and it is quite difficult to observe the passive attacks.

This article contains different sections i.e. Section-I introduce the basic requirements of secure multicast communication, Section-II explores the related research work in relevant fields. It provides brief overview of the key based security solutions, and also investigates some intrusion detection/prevention schemes, Section-III explains proposed scheme, Section-IV & V describes simulation setup/results, Section VI shows the security analysis and Section-VI concludes the outcome of the proposed scheme and its future use.

II. Related Work

A. Key based security solutions

As per above discussion, Researchers have developed various solutions to secure multicast network operations by introducing the concept of group key generation, Key distribution, mutual authentication for dynamic groups, secure group key exchange, intrusion detection and prevention algorithms etc. Hui Xia et al. [4] proposed a multicast trusted routing algorithm with QoS multi-constraints which is based on a modified ant colony algorithm. This algorithm combines the security trusted model and the modified tree based ant colony algorithm with the QoS multi constraints and this combination is used to explore the trustworthy multicast forward paths that prevents the network from various security threats.

Dr. N. Sreenath et al. [5] proposed an enhancement for the Secure Enhanced-On Demand Multicast Routing Protocol (EODMRP) to prevent it from various security attacks such as flooding and black hole attacks. Simulation results show that there is some improvement in packet delivery ratio in presence of black hole attack, with marginal rise in average end-to-end delay and normalized routing overhead and in case of flooding attack, it uses simple statistical packet dropping method that prevents the attacks from malicious nodes effectively.

Ahmed. M. Abdel et al. [6] explored the new possible security threats which can degrade the performance of multicast ad hoc routing which includes the different network operations i.e. election of group lead, link errors, repair and route management etc. To interrupt group lead selection process, intruder tries to select a node as lead that does not belong to multicast tree and later on group can be split into multiple groups. Intruder can also send the route repair requests for the routes which actually do not exists, in order to initiate real route maintenance for entire network. To protect the network against these attacks, hop by hop authentication method can be used to validate each route request and certificate based approach can be used to identify the group members and it can also be used for leader election process as well as for route maintenance. Simulation results show its performance in terms of improved PDR under the constraint of compromised network resources.

Ratna Dutta et al. [7] proposed and analyzed a generalized self-healing key distribution using a vector space access structure in order to reach more flexible performance of the scheme. Proposed method reduces storage, communication and computation costs over previous approaches, and is scalable to very large groups in highly mobile, volatile, and hostile wireless networks.

Sencun Zhu et al. [8] presented an overview of the various approaches that have been recently proposed to address the group key management issues and finally discussed several new research directions. Authors focused on ad-hoc and sensor networks and explored most common issues related to detection of compromised nodes, key distribution, Group rekeying schemes etc.

Zahraa Sabra et al. [9] proposed end user solution which is capable to provide secure environment for VoIP communication with respect to QoS parameters using hybrid ad hoc networks. They used AES, ECC192 to implement security features and Voice codec G.729b for simulation purpose.

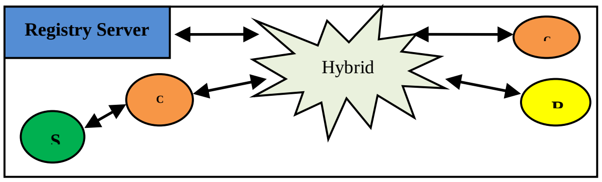


Fig. 1. Hybrid Network [9]

Fig. 1 above shows that as per the proposed scheme, Sender and receiver can communicate through cluster heads which are randomly selected and these are connected to hybrid networks. Shared keys are computed using SHA-512. Sender side cluster head sends query to Registry server which is forwarded to receiver side cluster head. If cluster heads share same keys, only then a link is established on the basis of authentication, after that sender and receiver can start VoIP communication. If a new cluster head is selected then registry server updates this information. Simulation results show that this scheme offers authentication against non-repudiation and traceability under QoS constraints.

Vennila Rajamanickam et al. [10] proposed an inter cluster communication and rekeying technique for multicast security in MANET using shared private keys generated by key manager. Low cost rekeying method is used when a node joins the cluster. Simulation results show the efficiency of proposed method in terms of low overhead and less computation cost.

E.A.Mary Anita et al. [11] designed a Worm Hole Secure ODMRP (WHS-ODMRP) that uses a certificate based authentication method in route discovery process. Authors also analyzed the performance of On Demand Multicast Routing Protocol (ODMRP) under the attack of worm hole using different scenarios. Simulation results show the comparison of WHS-ODMRP and ODMRP protocols and proved that proposed protocol can enhance the performance by reducing the packet loss caused by malicious nodes.

E.A.M. Anita [12] proposed a certificate based localized authentication scheme to prevent Sybil attack. Results show that the proposed method can sustain the Sybil attack and it is able to maintain the network performance in terms of throughput.

Jiwen Guo et al. [13] proposed a secure minimum-energy multicast (SMEM) algorithm to ensure multicast communication. In order to improve the stability of trust mechanism, the new trust values (calculated by the Bayesian theorem in CR networks) are modified by the iterative control criterion. Trust mechanism aims at guaranteeing the security of network environment, in which the trust information is encrypted to ensure the creditability of trust values. Results show that the time complexity of SMEM algorithm is polynomial.

Ding Wang et al. [14] investigated authentication related issues and presented a improved scheme to prevent the attacks over user credentials, called Kim-Kim scheme. Proposed Scheme offers three different phases i.e. first of all users are registered and key pairs (public & private keys) are produced as saved on server side. Whenever user wants to communicate with other, produced keys are saved on user’s device and to initiate communication, user side and server side credentials are used for mutual authentication purpose. Authors also investigated the potential threats for the proposed schemes i.e. compromise of node, keys, shared medium or server’s session etc. They performed cryptanalysis for each possible threat and they also raised some open issues like dependency of security goals over cryptography methods and offline security threats etc.

Babak Daghighi et al. [15] explored the various schemes which can be used for secure group communication and investigated the issues related to key exchange under the mobility constraints. They focused on the mobility of host as well as member in a group. It is quite complex to manage the validity of keys as any node can leave or join a group frequently. If a group member node leaves the group, that node should not be able to reuse the keys as well as the other resources of the group. If a node wants to join the group, authentication is required but if its keys have been expired, then there is need to reproduce the keys again but regenerated key should be unique. Frequently group updates may lead to extra overhead on group communication, storage key pool for group members and can affect the scalability, reliability and QoS etc. In order to develop a solution for secure group communication, all above factors are considered by researchers and they developed few solutions i.e. KMGM, GKMW, HKMS, TMKM, CDLM, HSK, FEDRP, GKMM, LKH, BALADE, KTMM, WSMM, M-IOLUS and SHKM etc. but each solution has its own limitations and there is requirement to explore Key management in highly mobile environment with the provision of QoS support which is essential for secure group communication.

Lin Yao et al. [16] developed a distributed key management scheme which can preserve the keys for nodes. Nodes can utilize the keys on the basis of their trust levels maintained by different nodes. Trust level of keys can protect from the various attacks and node can easily select the keys on the basis of their trust levels and it can eliminate the requirement of certification authority. Analytical and simulation models developed for this scheme show its performance in terms of less control overhead and efficient key management.

Lein Harn et al. [17] proposed an enhanced key management scheme using group Diffie–Hellman (GDH) key agreement protocol by introducing secret sharing method. Analytical model shows its performance in terms of its resistance against the well-known attacks over shared keys.

B. Security solutions forIintrusion Ddetection and Prevention

V.Srihari et al. [18] did a survey of the security threats and their remedies for VoIP/SIP protocols. As per their studies, security attacks can be introduced on session management, signaling, call control and credits etc. For detection and prevention, service providers can use intrusion detection system, fake call monitoring system, call analysis and pattern recognition etc.

Jiazi Yi et al. [19] did vulnerability analysis of Relay Set Selection (RSS) algorithms for the Simplified Multicast Forwarding (SMF) Protocol which is used in mobile ad hoc networks. Study shows that network topology can be compromised by misconfigured routers or malicious nodes by using spoofing. Attackers can also inject information conflicts for RSS decision making process. To enable the security provision, authors explained the various attack vectors for different RSS algorithms.

A.m. Pushpa et al. [20] propose multicast activity-based overhearing technique to identify this attacker node in the multicast group. They analyzed the multicast announcement packet fabrication to keep the track of group behavior and on the basis of threshold value, nodes can be isolated from network. Each node collects the feedback for a specific node and also considers the feedback status about that node to make the final decision. If all feedback collected about a particular node below then the threshold, then it is finally isolated from entire network. Simulation results indicate the impact of attack on the performance metrics such as Packet Delivery Ratio (PDR) and delay of PUMA and MAODV multicast routing protocols. Results show the efficiency of proposed scheme in terms of detection of malicious nodes w.r.t. less control overhead and false negative alarm rate. Proposed scheme can be extended to identify the attacks over parent selection method, which is used by tree based multicast routing protocols.

J.K.Harika et al. [21] proposed a secure multicast protocol for intrusion detection systems that uses hybrid cryptography to isolate the unwanted network overhead. In hybrid scheme, nodes can negotiate the session key for secure communication that fulfils the requirement of Authentication. Results show that the proposed scheme can defend the network from various attacks such as reply attack, rushing attack, IP spoofing and man in the middle attack etc.

A. Fidal castro et al. [22] proposed an artificial intelligence based solution which utilizes the analytical equations for network intrusion detection and prevention and can guard against several attacks i.e. black hole, Neighbor, route disruption etc. It builds the new rule as per the identified attack and this information is shared with each node. Simulation results show its performance in terms of rate of intrusion detection and response under the constraints of different mobility/traffic patterns.

Hui Xia et al. [23] proposed a scheme which estimates the route as per the assigned trust value and this value can be used to identify the attacker nodes. Trust level can be calculated for nodes as well as for routes also. For genuine and intruder nodes, different threshold values can be set and if any node's trust value is less than the recommended threshold, this can be identified as intruder. Node's trust level is used to build route's trust levels with route's state. Threshold value is updated, if there are variations in trust values and idle routes are identified and ignored by routing table. Performance of proposed scheme is evaluated by varying speed and density of intruders. This work can be extended further by considering various factors such as delay, threshold variations etc.

T. Stephen John et al. [24] developed an agent based method to identify the intruder nodes in network. Mobile agents are introduced by the sender node and they can adopt any forwarding route to find out any intruder. Broken routes are managed by intermediate nodes. ODMRP and D-ODMRP routing protocols were used for simulation purpose and results show ODMRP's performance is better than D-ODMRP in terms of delay, control overhead, PDR, energy utilization w.r.t. network size and node density etc. Proposed scheme can be extended ty introducing the concept of inter communication process for mobile agents.

Wei Yuan et al. [25] developed a routing scheme, called topology hidden multicast routing (THMR) which can isolate the routing information to prevent the network from well-known attacks over routing. Receiver insures the identification of sender and shared keys. Route information is isolated for intermediate nodes, in order to avoid the attack on routing table. To make a route request, first of all node produces shared keys for session using RSA algorithm and then generates broadcast messages which are validated by intermediate nodes. At the destination end, received packets are verified on the basis of the relevant keys and all are discarded, if their keys are already compromised. If packets are accepted then a replay is prepared using shared keys and this is again validated by intermediate nodes and finally it is accepted by source node. In case of route errors, short lived public keys are used to propagate broken link information and finally a new route is built, if it can't be repaired. It shows its resistance against various attacks i.e. impersonation, DoS, packet analysis, fabrication and routing attacks etc. Simulation results show its performance in terms of key computation time, delay and latency as compared to MAODV.

A. Menaka Push et al. [26] explored the packet drop/fabrication attack and introduced a watchdog algorithm based scheme to analyze it. For authentication purpose, it calculates the node's distance from core and each node keeps the track of its neighbor and in case of excessive packet drop, identified node is eliminated from group communication. If value of core's distance is altered by malicious node, then it can be identify by the neighbors by verifying the actual distance of core and its surrounding neighbors. If there is any difference between hop count and distance value, then current parent node can be identified as malicious node and finally it is neglected by group. Simulation results show it is able to maintain PDR and control overhead under the compromised situations but improvement in network performance and the impact of security threat, both depend upon the actual location of malicious nodes from the core and these two factors are inversely proportional to each other. Proposed scheme can be further enhanced for another multicast routing protocol.

P. Anitha et al. [27] developed a dynamic pre-keys distribution scheme to protect the network from Sybil attack and they Integrate the proposed scheme with On-Demand multicast routing, called S-ODMRP. Key distribution utilizes the relevant information of each node and common keys are used to establish a secure session and key can be easily validated, if it is common between two nodes. Simulation results show its performance in terms of improved PDR under the constraints of security threat.

N.M. Saravana kumar et al. [28] proposed a key management solution for multicast group operations. It can adopt dynamic behavior of group members as they can join/leave the group at any time using member authentication based on their signatures. Key pairs contain different subset of keys for inside or outside group communication. Key pairs enforce the rules for various operations i.e. group join/leave and data exchange etc. If any other node forcefully joins the group, that cannot access the information due to the absences of previously generated session/group keys. If ex-group member wants to re-join the group, as per the record of session keys, authentication can be done before group joining. Proposed method performs well under Security QoS constraints i.e. integrity, confidentiality, key calculation time, data processing time etc.

Xiao Wang et al. [29] explored the possible threats over multicast group communication and developed the solution by considering various factors i.e. mobility, scalability and key management etc. All nodes are arranged into self-organized form having one hop distance to each other and a Group manager (GM) is defined for multiple members. Diffie–Hellman key agreement protocol is used for key management and it uses different keys i.e. session key which is common for GM and group members, mobility key and a field keys are used when nodes move to another groups. Keys are generated for a particular group only. Keys are updated as per different network operations i.e. node movement, group join/leave, link breakage etc. Analytical and simulation models show the performance of proposed method in terms of consumed energy for key calculations, control overhead and efficient key management for groups etc.

III. Proposed Scheme

This paper presents a Secure MAODV Routing Solution based on Group Diffie-Hellman (GDH) Key distribution algorithm. Following are the basic steps of GDH algorithm (including Phase-I, II & III). Phase IV is used for node authentication before group joining.

```
Phase I: Proc init (p,g) {
//initialize all nodes and assign p,g values
                Initilize Node(s):=n;
                Calculateg();
               For each node $N_i$ {
              Assign ($N_i \rightarrow p, N_i \rightarrow g$) }                    (1)
Proc Calculateg () {        p= getPrime()
//All nodes calculate the value of universal g on the basis
of initial value of G using private key i, for key exchange.
initial G=getPrimitiveRoot(p)
$Univeral\ g = \{N_i \rightarrow G^{a,b,c,d,e,f,\ldots n}\}$     }                     (2)
Phase II: Proc KeyExchange $(N_i, N_j)$ {
           $A_i = N_i \rightarrow g^a \ mod \ N_i \rightarrow p$                      (3)
           $B_j = N_j \rightarrow g^b \ mod \ N_j \rightarrow p$                      (4)
           $N_i \rightarrow key_i$=$B_j$                                   (5)
           $N_j \rightarrow key_i$=$A_i$            }                      (6)
Phase III: Proc KeyAgreement $(N_i, N_j)$ {
            $s_i$=$N_i \rightarrow key_{i\square}^{a} \ mod \ p$                      (7)
            $s_j$=$N_j \rightarrow key_{i\square}^{b} \ mod \ p$                      (8)
            $N_i \rightarrow skey_{i\square}^{\square} = s_j$                         (9)
           $N_j \rightarrow skey_{j\square}^{\square} = s_i$    }                     (10)
Phase IV: Proc Join_Group $(N_i)$ {
            If (IsAuthentic($N_i \rightarrow p, N_i \rightarrow g, N_i \rightarrow skey_{i\square}^{\square}$ )
            {          Set $N_i \rightarrow Join_{\square}$=True;
                       Update(Multicast Table, $N_i$)
            } else { Set $N_i \rightarrow Join_{\square}$=False; } }
```

A. Key Assignment and Node Authentication

To establish shared keys, all nodes participate in GDH algorithm key generation process to produce a universal value of g which cannot be reproduced by individual node. All candidates submit their private keys i.e. $\{N_1 \rightarrow G^a\}$, $\{N_2 \rightarrow G^b\}$, $\{N_3 \rightarrow G^c\}$,..$\{N_i \rightarrow G^n\}$ to produce g. After generating g, first node becomes group lead and starts key negotiation with upcoming members to generate key pairs for distribution purpose.

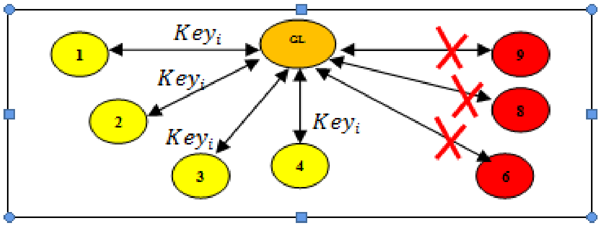


Fig. 2. Key assignment and node authentication

Once keys are assigned to each candidate node, group join process is initiated. During key negotiation and distribution process, if nonmembers try to join a group, they are authenticated on the basis of their $\{N_i \rightarrow G^i\}$ values and finally their group join requests are discarded. After successful key distribution process, local flag codes and global HASH values are generated and each member node is aware of these codes. If any member node does not verify the incoming and outgoing requests, activity logger generates alerts for authentication violations.

B. Control Message Authentication

A node can generate four different types of Route Requests: RREQ is used for route discovery and maintenance, RREQ-J for group joining, RREQ-R and RREQ-JR for tree merge. Only authorized nodes can generate RREQ messages with unique local flag codes. On the basis of these codes, incoming RREQ messages can be verified along with the encrypted flag to ensure route integrity and it is discarded, if local flag code of received RREQ does not match with the calculated local flag code.

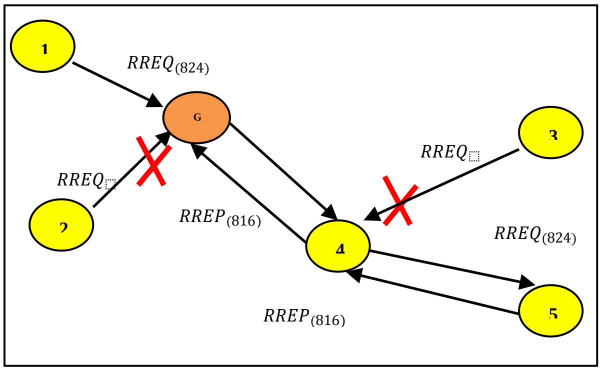


Fig. 3. Local Flag Codes

Table I. shows the list of local flag codes generated by one way HASH function during network operations. Following are the procedures used to verify the authenticity and integrity of messages. In case of any violation, Log alert messages are generated.

```
Proc setFlags (Msg, $S_i$)
{
Get_Local_Flag_Code($Msg \rightarrow Type, 1w\_Hash()$);      (11)
Get_Global_Hash_Value($Msg \rightarrow Type, SHA512()$);     (12)
        $encrypt(Msg \rightarrow flag, S_i \rightarrow key_i)$              (13)

LogInfo('Outgoing_Msg_From Node= $S_i$ at $Time = T_i$');
LogInfo('Outgoing Local Flag Code= $Msg \rightarrow Lfc$ at $Time = T_i$');
LogInfo('Outgoing Global Hash Value= $Msg \rightarrow Gh$' at $Time = T_i$);
}

Proc CheckFlags (Msg, $R_i$)
{
$Lf$ = Chk_Local_Flag_Code($Msg \rightarrow Type, 1w\_Hash()$);

$Gh$ = ChkGlobal_Hash_Value($Msg \rightarrow Type, SHA512()$);
                                                            (14)
      $d = decrypt(Msg \rightarrow flag, R_i \rightarrow key_j)$          (15)
       If ($Lf$ = $Gh$ = $d$ =True)
       {
         Accept=1;
       }
       else {
              Accept=0;
LogAlert ('Invalid_Msg_sent by Node= $S_i$ at Time=T' );
LogAlert ('Incomming Local Flag Code= $Msg \rightarrow Lfc$ at Time=T');
LogAlert ('Incomming Global Hash Value= $Msg \rightarrow Gh$' at $Time = T_i$); }
         return Accept;
       }
//Send a Message
        $S_i \rightarrow Send(setFlags(Msg, S_i), R_i,)$            (16)

//Receive a Message
       If ($R_i \rightarrow Recv$(CheckFlags($Msg$), $S_i$)==1)    (17)
        {
          $R_i \rightarrow accept(Msg)$
        }
        else {
        $R_i \rightarrow discard(Msg)$
        }
```

Where $Sender = S_i$, $Receiver = R_i$, in a particular group G_i, $at\ Time = T_i$, Msg indicates data to be sent, $Msg \rightarrow Type$ contains Request/Response control messages, $1w\_Hash()$ calls one way Hash function to calculate local flag code, $SHA512()$ calls SHA512 function to calculate global HASH values

Flag code based verification can filter the fake route requests and their replies. Unauthorized nodes cannot calculate local flag codes because these codes are available for group members only. SHA512 algorithm generates global HASH values which are used to ensure integrity of control messages.

TABLE I. LOCAL HASH CODES

MAODV Control Messages	Local Flag Codes
RREQ	824
RREQ-R	3208
RREQ-J	3976
RREQ-JR	6920
RREP	816
RREP-R	3200
RREP-J	3968
RREP-JR	6912
MACT-J	3998
MACT-P	3294
MACT-GL	7486
MTF-UP	3096
MTF-DOWN	15400
GRPH	932
GRPH-U	3316
GRPH_M	4084
HELLO-MSG	28938
Multicast Table Entry	22780518

If any node wants to a join multicast group, its keys are verified and a global HASH value is used for message authentication further. In case of invalid keys, all RREQ and RREP are discarded and if keys are valid but global HASH value based authentication fails, even then node cannot join the group. After successful joining of the group, activities of each node are logged and warning messages are generated for all unmatched flag codes and HASH values.

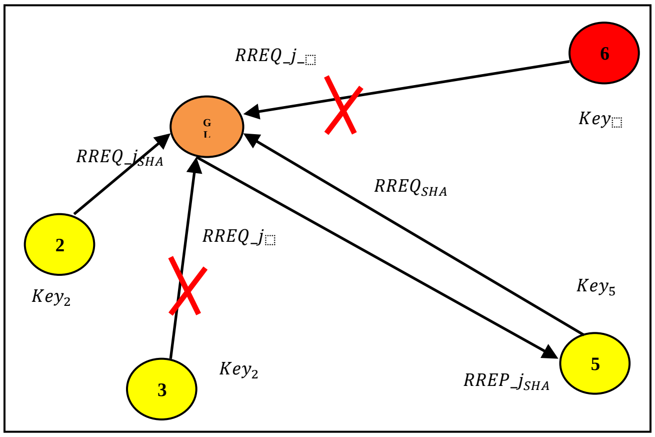


Fig. 4. Global HASH Values

Fig.4 above sows that RREQ_J of node 6 is discarded due to invalid keys, RREQ_J of node 3 is also rejected due to invalid global HASH value. REQ_J of node 2 and 5 is accepted due to successful message authentication.

C. Key Revocation

If any member node leaves a particular group, all assigned keys are revoked and that node cannot rejoin the group till further key negotiation.

TABLE II. GLOBAL HASH CODES

MAODV Control Messages	Global HASH Values
RREQ	8f9e705ccf9bbd05349fd0940428822385ddfd73e9321f9c28f008db7527bd6e881e22a418ca1562cdaf16df33ad332d13bf0744737b7f406b7c5b893d31fcc9
RREQ-R	63b2d0c9b925dd6c8820798099ae6c099245d9c4eaaeb6a3b604c09bebe30b39a62b9c99162922ebc05fed156a45ebd849ed088cef6abe6f87f010bbded5e37a
RREQ-J	f56c4416632652f2b3469ffdada9c9f245a8c025e128e471d05f75e625320024ee61bd29bbbd7dc1454102107ffe7b6fb288c1f0e0a56132f5b7c8cc3eb2be18
RREQ-JR	151e550443631e80078cca115ee978e0498dd5942661f5389e39694e5f8f619bc6a100cc9ebf60fafdad83c44b49e64e86595a530a06ac1ed1d537273968eee9
RREP	fd92856a81d58d4abd15a032b3a47d5df24d178b6142dcd86cd888a7902206089e052c082c0bbe0d7fd20731e773e2f0a57cb009ec68dd97c8d6dfc3483cbac2
RREP-R	168c5626507d1e5e892a5b0dad06c3da6fa82ee52cc80be7f6ec0f39320ffdc6e57a8f4dbbfa9b26f2f17d573e0a931c06a5ec28b8110b84e53dab4138199bec
RREP-J	d45e6c5407eb39095a27092c79cfb140d1e4f17950d9c377122f2fb03a14a21d529edb2f1b066795e8efb2007cdfe806390b329c340a0b50a3b351892bbefa8f
RREP-JR	ac9586c62d99bbcbef4e5eddaa30253d51ecaec779236ed8fce7f310c6759789a5e37b7e54ec16972331120227bdc560a4abe86257bc48ad3e3e23ba73e772de
MACT-J	60019dad8270475155c29098b47f3c075636440fe1debc428ca0d681763e0eb7c432267888bf05a3655564045049f8a747cf3e8bbbe896625f59855e6f4d1065
MACT-P	d35108b3c2ccacf75650125f208c955c7819a3ca454162a99997585e6e3b4154e661a74a0fb5d2fb544432a8cd0836e14d2551069489fa5583fdc71e937c0279
MACT-GL	7c60039bbc556772b9ab8728228291a9e7b6c333624bd94aaa38b6b8b327feda74e1934e75d9a48f5ff98917906720b503f94e2f261e264495a422d5ce17076b
MTF-UP	8701d222c072f43a4b70b049bc6a5db86e8fd014db4f218206d63496e7e8c25e9f496a9da7682959fb20cc7ddb7538cbed60197be68769ea8b01190dd1c03d43
MTF-DOWN	de1a48482d525753dac00107508d155b2d03a7610a96d740a6e81ffa89df92dfb1de6e428ed573ca61e63dfaf1557a850946444ad1e24e788454fb006802d8e7
GRPH	d13248a6fc8be1710e1c4a01b574c1f57cb2ffed8715f14cdc3917ba0d99982894dc5c5b8960ec71efc64c8455323d9e89c7c91956f43d98a1c97d6d6ac12cec
GRPH-U	c16b48e4cdde2b1030d578a3a36bcd766ce9cac04e8303d30a0252c707fbdfaf30864859f945246cb9e8267e37a2098390eede801b7790975ff44df5952899a2
GRPH_M	4db9c2544085faef3ba1bfba49044a254ae1e8662cef8f2ae8e6f4e1e3df53974f94ef106b821f874b6b8afdca9004009977726854f8d5f52a686f3c2234b5e6
HELLO-MSG	ec3256d70176dbd67b5370135ac3432a8654436b984ecd61de1fce8faf258326402a0d2bba95f82375c1c10a2e1895fb754e4160bbdb9583719e3a0ecbf39b13
Multicast Table Entry	ae2e360c2c7c3342a03aed80de66f3ac2afd89dca20d04824beb9a44a3124a667b91f2bd6178b52cd69b4a5371e6ee3d4f9e8c03f200b33fb9c42ea2447d7b55

Local flag codes and global HASH values both are verified for various network operations i.e. Tee construction/Tree Merge/Tree Pruning (MACT-J/MACT-P/MACT-GL), Group Leader selection etc. Without authentications of these codes, no multicast operation can be performed.

IV. SIMULATION SETUP

TABLE III. SIMULATION SCENARIO

Simulation Scenario	
Total Nodes	30
Sender Node(s) Density variation	1,5,10,15
MAC Protocol Standard	802.11
Key Distribution Algorithm	Group Diffie-Hellman (GDH)
Wierless Terrain	1200x1200
Multicast Ad Hoc Routing Protocol	MAODV
Simulation Time	10 Minutes
Group Size	1
Propagation Model	TwoRayGround
Traffic Type	CBR
Packet Size	512 Bytes
Sampling Interval	0.1 Second
Network Simulator	NS-2.35
One way HASH Funtion	For Local Flag Codes
SHA512 HASH Funtion	For Global HASH Values
Mobility Model	Random WayPoint

Table III. above shows the simulation scenario used for analysis purpose.

V. SIMULATION RESULTS

Following graphs show the performance of entire network using different parameters, i.e. Throughput, Packet Delivery ratio and Routing Load etc.

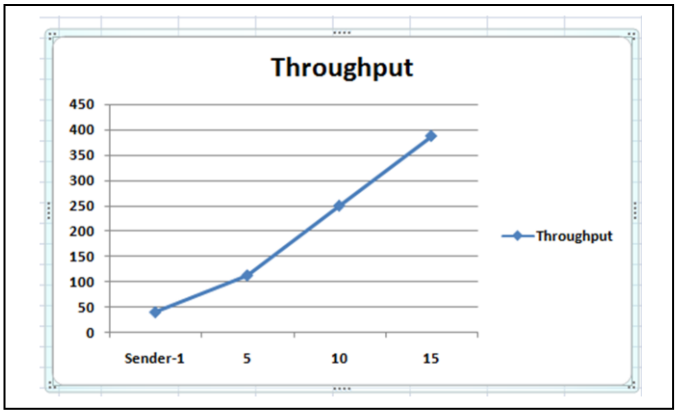


Fig. 5. Throughput

Fig.5 Above shows Throughput of the network with the sender density 1,5,10 and 15. It shows the improvement in Throughput, which is increasing w.r.t. sender's density.

Fig.6 Above shows Routing Load of the network with the sender density 1,5,10 and 15. It shows that Routing Load is decreasing w.r.t. sender's density. Fig.7 shows the significant improvement in Packet delivery Ratio of the network with the sender density 1,5,10 and 15.

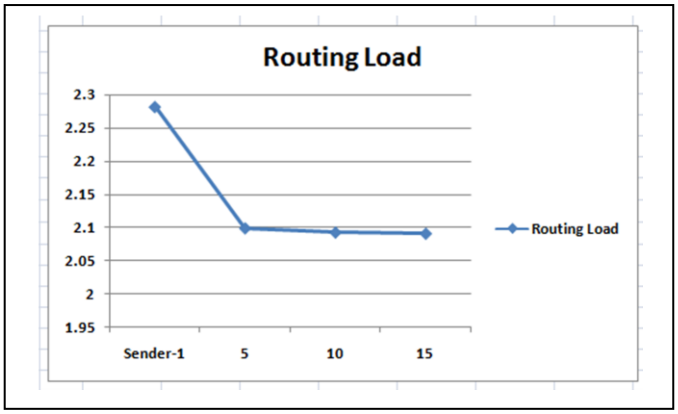


Fig. 6. Routing Load

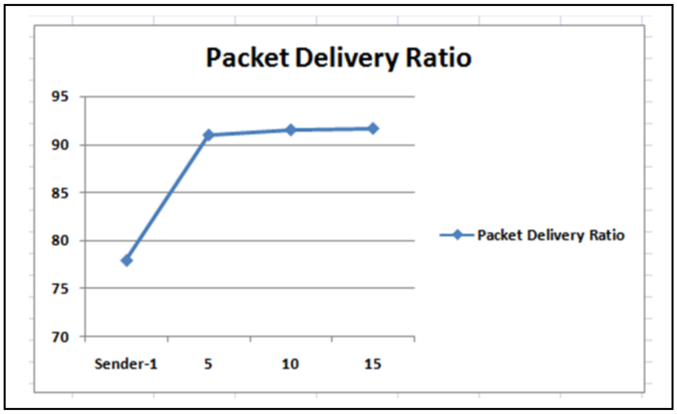


Fig. 7. Packet Delivery Ratio

VI. Security Analysis

A. *Secure Key management and node authentication*

GDH algorithm supports secure communication based on group authentication. At the initial stage, it requires a large prime number P and its primitive root G. All candidate nodes participate to generate a universal g on the basis of their private keys using P and G.

$$Univeral\ g = \{N_i \rightarrow G^{a,b,c,d,e,f,\ldots n}\} \tag{18}$$

If an intruder generates:

$$g^j = \{N_i \rightarrow G^{a,b,c,d,e,f,\ldots n+x}\} \text{ or} \tag{19}$$

$$g^j = \{N_i \rightarrow G^{a,b,c,d,e,f,\ldots n-x}\} \tag{20}$$

Then $g^j \neq g$, that means to generate exact value of universal g, intruder must use the key combination equivalent to original participant keys and it depends upon the number of candidate those want to form a group. Shared group key s_i can be generated, only if $(a, b) \in N_i \rightarrow g^{a,b}$

$$s_i = N_i \rightarrow {key_i}^a \ mod\ p \tag{21}$$

$$s_j = N_j \rightarrow {key_i}^b \ mod\ p \tag{22}$$

$$s_i = s_j \text{ , only if } (a, b) \in N_i \rightarrow g^{a,b} \tag{23}$$

If node leaves the group, then s_i is revoked and it cannot rejoin the group without key negotiation.

B. *Secure Multicast Tree construction and maintenance*

Local flag codes and global HASH values cannot be intercepted because all these are produced at the time of group formation and regeneration of exact codes by malicious nodes is not feasible. RREQ and RREP are used for route discovery. Intermediate nodes verify them on the basis of local flag codes and forward those by embedding an encrypted flag in their header with global HASH values. All messages without valid flag codes, global HASH values and encrypted flags are declared as unauthorized messages are declared as unauthorized messages and finally discarded. During RREQ propagation phase, intermediate authorized nodes set their status ON_TREE, if they are not on the tree and update multicast routing table and multicast packets are further propagated. RREQ/RREP messages from unauthorized nodes are not entertained.

An encrypted flag is merged with Multicast Route Activation (MACT) header. MACT-J is used for tree construction or when a node wants to join group. After receiving MACT-J, its flag is decrypted and local flag code and global HASH values are verified to update multicast routing table. Unauthorized MACT-J is rejected and routing table is not updated. MACT-GL is used for new group lead selection and MACT-P for Tree pruning. For new group leader selection, all shared keys of eligible candidate are verified and an encrypted flag is merged with MACT-GL header for authentication purpose. If the selected group leader cannot decrypt the flag, next candidate is selected for leadership and so on. Tree pruning is invoked when a node leaves the multicast tree and after that upstream node becomes a leaf node. Tree pruning is controlled by verifying the shared keys and HASH values of upstream/downstream nodes. After successful verification, MACT-P is processed otherwise it is filtered out. Finally, it can prevent group leader selection attack/MACT fabrication attack.

The only authorized group leader is allowed to generate periodic Group Hello messages (GRPH). After receiving a GRPH message, intermediate nodes update their multicast table after message authentication. Upstream and downstream node authentication is performed on the basis of shared key and HASH value for processing of GRPH-U, GRPH-M, RREQ-JR and RREP-JR etc.

VII. Conclusion

In this paper, GDH algorithm for key generation and distribution was used along with MAODV routing protocol. Key negotiation starts at the time of the group joining. All candidates use their private keys to calculate the universal value of g. After that shared key pair is generated and distributed to each participant node only. Fist node becomes the group leader and generates local flag codes and global HASH values, and embeds them with each message. These codes are used to verify the authenticity and integrity of all messages. Local flag codes/global HASH values are used to verify critical multicast operations i.e. Group Leader Selection, Tree Construction and Maintenance etc. All invalid requests and responses are rejected.

As per security analysis, it can be observed that reproduction of shared key is not feasible due to the absence of private keys of each node. In the case of compromised keys, the intruder cannot intercept the local flag codes and global HASH values and without using codes, all routing messages are discarded. Simulation results show that in the presence of multiple senders, Throughput and PDR of the network increase w.r.t. Sender's density, whereas routing load is reduced. Finally, simulation and analysis results conclude that proposed scheme can protect the routing information without generating extra control overhead, and it can be further extended to adopt the compromised network environment using different multicast routing protocols.

REFERENCES

[1] Hui Cheng et al., "Hyper-mutation based Genetic Algorithms for Dynamic Multicast Routing Problem in Mobile Ad Hoc Networks", 11th International Conference on Trust, Security and Privacy in Computing and Communications-2012 IEEE, pp. 1586-1592.

[2] N. Bhalaji et al., "Performance Comparison of Multicast Routing Protocols under Variable Bit Rate Scenario for Mobile Ad hoc Networks", Recent Trends in 19 Network Security and Applications Communications in Computer and Information Science vol. 89, 2010, Springer-2010, pp. 114-122.

[3] C. Siva Ram Murthy, B.S. Manoj, "Ad Hoc Wireless Networks", 14 impression- Pearson-2012, Chapter (5-11), pp. 191-641.

[4] Hui Xia et al., "Multicast Trusted Routing with QoS Multi-Constraints in Wireless Ad Hoc Networks", International Joint Conference of IEEE, TrustCom-IEEE-2011, pp. 1277-1282.

[5] Dr. N. Sreenath et al. "Countermeasures against Multicast Attacks on Enhanced-On Demand Multicast Routing Protocol in MANETs", ICCCI-IEEE -2012, pp. 1-7.

[6] Ahmed. M. Abdel Mo'men, Haitham. S. Hamza, IEEE Member, and Iman. A. Saroit, "New Attacks and Efficient Countermeasures for Multicast AODV", IEEE-2010, pp.51-57

[7] Ratna Dutta et al. "Computationally secure self-healing key distribution with revocation in wireless ad hoc networks", Ad Hoc Networks, vol. 8 (6), August 2010, Elsevier 2010, pp.597-613

[8] Sencun Zhu et al., "Scalable Group Key Management for Secure Multicast: A Taxonomy and New Directions", Network Security, 2010, Springer-2010, pp. 57-75.

[9] Zahraa Sabra and Hassan Artail,"Preserving Anonymity and Quality of Service for VoIP Applications over Hybrid Networks", IEEE Mediterranean Electro-technical Conference, 2014, pp.421-425

[10] Vennila Rajamanickam, Duraisamy Veerappan, "Inter cluster communication and rekeying technique for multicast security in mobile ad hoc networks", IET Information Security, IEEE, 2013, pp.234-239

[11] Anita, E.A.M. Bai, V.T., Raj, E.L.K., Prabhu, B, "Defending against worm hole attacks in multicast routing protocols for mobile ad hoc networks", Advances in Computing and Communications in Computer and Information Science vol. 190, 2011, Springer-2011, pp. 1-5.

[12] E.A.M. Anita, "Sybil Secure Architecture for Multicast Routing Protocols for MANETs", Advances in Computing and Communications in Computer and Information Science vol. 190, 2011, Springer-2011, pp. 111-118.

[13] Jiwen Guo et al., "Secure Minimum-Energy Multicast Tree Based on Trust Mechanism for Cognitive Radio Networks", Wireless Personal Communications November 2012, vol. 67 (2), Springer-2012, pp. 415-433.

[14] Ding Wanga, Nan Wang, Ping Wang, Sihan Qing, "Preserving privacy for free: Efficient and provably secure two-factor authentication scheme with user anonymity", Information Sciences, Elsevier-2015

[15] Babak Daghighi, LaihaMatKiah, Shahaboddin Shamshirsband, Muhammad abib Ur Rehman, "Toward secure group communication in wireless mobile environments: Issues, solutions and challenges", Journal of Network and Computer Applications, Vol. 50, Elsevier -2015, pp. 1-14

[16] Lin Yao, Jing Deng, JieWang, Guowei Wu, "A-CACHE: An anchor-based public key caching scheme in large wireless networks", Computer Networks, Elsevier-2015, pp.78-88

[17] Lein Harn, Changlu Lin, "Efficient group Diffie–Hellman key agreement protocols", Computers and Electrical Engineering, Elsevier-2014, pp. 1972–1980

[18] V. Srihari, P. Kalpana, "Security Aspects of SIP based VoIP Networks: A Survey", ICCTET, IEEE-2014, pp-143-150

[19] Jiazi Yi et al., "Vulnerability Analysis of Relay Set Selection Algorithms for the Simplified Multicast Forwarding (SMF) Protocol for Mobile Ad Hoc Networks", 15th International Conference on Network-Based Information-IEEE-2012, pp. 255-260

[20] A. Menaka Pushpa, Dr. K. Kathiravan, "Secure Multicast Routing Protocol against Internal Attacks in Mobile Ad Hoc Networks", IEEE GCC Conference and exhibition, 2013, pp-245-250

[21] J.K.Harika, Dr.C.Jayakumar, "An Acknowledgement Based Secure Data Transmission in MANETS", ICICES,IEEE-2014, pp-1-5

[22] A. Fidalcastro, E. Baburaj, "An Advanced Grammatical Evolution Approach for Intrusion Detection on multicast routing in MANET", ICICES, IEEE-2014, pp.1-4

[23] Hui Xia, Jia Yu, Zhi-yong Zhang, Xiang-guo Cheng, Zhen-kuan Pan, "Trust-enhanced multicast routing protocol based on node's behavior assessment for MANETs", International Conference on Trust, Security and Privacy in Computing and Communications, IEEE-2014, pp.473-480

[24] T. Stephen John and A. Aranganathan, "Performance analysis of proposed mobile autonomous agent for detection of malicious node and protecting against attacks in MANET", International Conference on Communication and Signal Processing, IEEE-2014, pp.1937-1941

[25] Wei Yuan, Liang Hu,Kun Yang, "A Topology Hidden Anonymous Multicast Routing for Ad Hoc Networks", GlobeCom, IEEE-2013, pp.599-604

[26] A. Menaka Pushpa, K. Kathiravan, "Resilient PUMA (Protocol for Unified Multicasting through Announcement) against Internal Attacks in Mobile Ad Hoc Networks", ICACCI, IEEE-2013, pp.1906-1912

[27] P. Anitha, G. N. Pavithra, P. S. Periasamy, "An Improved Security Mechanism for High-Throughput Multicast Routing in Wireless Mesh Networks Against Sybil Attack", PRIME, IEEE-2012, pp.125-130

[28] N.M. Saravanakumar, R. Keerthana and G.M. Mythili, "Dynamic Architecture and Performance Analysis of Secure and Efficient Key Management Scheme in Multicast Network", Artificial Intelligence and Evolutionary Algorithms in Engineering Systems Advances in Intelligent Systems and Computing Vol. 324, 2015, pp.775-784

[29] Xiao Wang, Jing Yang, Zetao Li, Handong Li, "The energy-efficient group key management protocol for strategic mobile scenario of MANETs", EURASIP Journal on Wireless Communications and Networking, Springer-2014, pp.1-22

[30] Mahalingam Ramkumar and Nasir Memon, "An Efficient Key Pre-distribution Scheme for Ad Hoc Network Security", IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, Vol.23 (3),IEEE-2005, pp.611-621

[31] Sourabh Chandra, Smita Paira, S k Safikul Alam, Goutam Sanyal, "A comparative survey of symmetric and asymmetric key cryptography", ICECCE, IEEE, 2014, pp.83-93

[32] Youssef BADDI, Mohamed Dafir ECH-CHERIF El KETTANI, "Key Management for Secure Multicast Communication: A Survey", IEEE-2013, pp.1-6

[33] R. Di Pietro, S. Guarino, N.V. Verde, J. Domingo-Ferrer, "Security in wireless ad-hoc networks – A survey", Computer Communications, Vol. 51,Elsevier-2014, pp.1-2

Ranking Documents Based on the Semantic Relations Using Analytical Hierarchy Process

Ali I. El-Dsouky
Computers and Systems Department,
Faculty of Engineering,
Mansoura University,
Egypt

Hesham A. Ali
Computers and Systems Department,
Faculty of Engineering,
Mansoura University,
Egypt

Rabab S. Rashed
Electrical Engineering Department,
Faculty of Engineering,
Kafr elsheikh University,
Egypt

***Abstract*—With the rapid growth of the World Wide Web comes the need for a fast and accurate way to reach the information required. Search engines play an important role in retrieving the required information for users. Ranking algorithms are an important step in search engines so that the user could retrieve the pages most relevant to his query.**

In this work, we present a method for utilizing genealogical information from ontology to find the suitable hierarchical concepts for query extension, and ranking web pages based on semantic relations of the hierarchical concepts related to query terms, taking into consideration the hierarchical relations of domain searched (sibling, synonyms and hyponyms) by different weighting based on AHP method. So, it provides an accurate solution for ranking documents when compared to the three common methods.

Keywords—Semantic rank; ranking web; ontology; search engine; information retrieval

I. Introduction

Web based information retrieval systems; especially search engines are the basic tools to assist users to find information on the World Wide Web. Despite the vital role in reaching information, many of the returned results are irrelevant to the user’s needs as they are ranked based on the string matching of the user's query. This has created a semantic gap between the meanings of the keywords in the retrieved documents and the meanings of the terms used in users' queries.

Search is the most popular applications on the Web. The bulk of traditional retrieval systems usually make use of metadata keywords matching with the query. However, these systems don’t take into account the semantic relationships between query terms and other concepts that might be significant to users. Thus, the addition of explicit semantics can improve the search process. Semantic search is an application of the Semantic Web to search. It tries to improve traditional search results (based on Information Retrieval technology) using data from the Semantic Web [1]. This approach offers an enhancement to traditional search as it allows retrieval to incorporate the underlying terms semantics [2]. It improves the traditional search that focuses on word frequency by trying to understand hidden meanings in the retrieved documents and users’ queries [3, 4]. The problem of poor retrieval information system exists when users cannot clearly express their information needs or poor ranking methods to evaluate pages if they are related to query or not.

In order to overcome the irrelevant documents that result from search process, there are many solutions such as: using query expansion (QE), taking into account the semantic meaning; or by improving the ranking of documents, taking into account not only the occurrence of query terms, but also the semantic relation between the user search and the document context.

QE is considered a viable solution, expanding process by expanding query keywords with related terms. With an expanded query, the retrieved documents are not only based on the query terms, but also on the related terms to that query which can improve the search process. This is suitably broadened and more accurate results may be obtained by retrieving more relevant documents .Web search ranking algorithms play an important role in ranking web pages so that the user could get good results more relevant to the user’s query.

This paper presents two methods to solve these problems. The first is an expansion query method taking into consideration the relations between expanded query terms in the ranking process of documents, by organizing all terms of an expanded query as a tree model of multi-levels, regarding their hierarchical relationships defined in a specific ontology. The second method is a ranking process for documents based on the semantic relation between document contents and the query terms.

II. Related Work

Search engines accuracy is improved based on how they will search for the meaning of query terms, and how they will present the results to users by evaluating the documents containing the query terms. There are many solutions for improving the search engine: by expanding query taking into account the semantic meaning related to user's query terms; or by improving the evaluation of documents not only by the occurrence of terms, but also by how it semantically relates to the topic search.

Query expansion (QE) is a technique used to aid users to express their requirements. There are many works in QE techniques, such as the mechanisms of relevance feedback [5] and statistical term co-occurrence [6]. The drawback of relevance feedback and statistical term co-occurrence methods is the analysis of pervious results documents which may provide a relationship between extracted terms and the original query. But this cannot be ensured if there are no sufficient documents used for analysis before a search process.

The semantic meaning is a method based on ontology to disambiguate the query meaning [7]. This method is used to expand query terms by their synonyms using WordNet ontology, or by adding synonyms and terms related to them based on ontology domain. But adding these terms to query without taking into consideration their hierarchical relationships may affect the relevance of documents to the main query terms [8].

Ranking methods are applied to arrange the documents in order of their relevance, importance and content score using web mining techniques to do this [9]. Web mining techniques are applied in order to extract only relevant documents from the database and provide the intended information to users. They classify the web pages and internet users by taking into consideration the contents of the page (WCM), behavior of internet user in the past (WUM), and web structure mining based on links in pages (WSM) [9-13].

There are many ranking algorithms that can be classified based on the parameters used to describe them and the parameters used to calculate the ranking score. We will discuss this in the following section.

Page rank algorithm is an algorithm used by Google to rank pages. It is based on a web graph, where web pages are represented as nodes; and links as edges between pages. The page rank depends on the number of links it has. The page linked to many pages with high PageRank receives a high rank itself [14-16].

Weighted links rank (WLRank) is the modification of the standard page rank algorithm [17]. This algorithm provides weight value to the link based on three parameters; the length of the anchor text, tag in which the link is contained, and relative position.

Time Rank Algorithm is based on the visit time of a webpage [18] to overcome the keywords query match without taking into account the context of user meaning. User's preferences in content and in a link are used to rank pages [19-20]. Also, user behavior can be used to indicate the importance of webpages and websites, by analyzing the individual user sessions to rank the web pages [21].

Semantic ranking is based on the domain ontology by similarity between ontology concepts and document page terms using the term frequency of terms [22]. Semantic ranking is based on the user logs and IS A and Part of hierarchy relations, the extension similarity is based on the user browsing patterns and their hyperlinks, the content similarity between two nouns are constructed based on the IS A and Part of hierarchy using user's web log to find the semantic ranking web page [23]. Modifying graph base sentence ranking by summarizing a text to nodes and edges as relations, to hypergraph to overcome the group of relationships between sentence, where a sentence represents as nodes and edges may be group relationship or pairwise relationship Text hypergraph for summarization and hypergraph based semi-supervised learning algorithm for sentence ranking [24].

III. ANALYTIC HIERARCHY PROCESS (AHP)

The Analytic Hierarchy Process (AHP) is an effective tool for dealing with complex decision making; it aids the decision maker to determine the priorities of used criteria. It based on a series of pairwise comparisons and then synthesizing the results, it also incorporates a useful technique for checking the consistency of the decision maker's evaluations.

The AHP generates a weight for each evaluation criterion according to the decision maker's pairwise comparisons of the criteria. The higher the weight, the more important the corresponding criterion is. Next, for a fixed criterion, the AHP assigns a score to each option according to the decision maker's pairwise comparisons of the options based on that criterion. The higher the score is, the better the performance of the option with respect to the considered criterion. Finally, the AHP combines the criteria weights and the options scores, thus determining a global score for each option and a consequent ranking. The global score for a given option is a weighted sum of the scores it obtained with respect to all the criteria [25].

In AHP (Analytic Hierarchy Process Matrix) a matrix is constructed where the Rows and Columns have the same parameters. The first row and the first column have the same parameter and the so on for other rows and columns. once the matrix is arranged ,the comparison between each row with all columns are done to determine the score, where a maximum score implies that the row is more important than the column. The diagonal of the matrix is allocated a score of 1. The score value of cell below the main diagonal is just inverse of the scores in the corresponding row. Likewise calculate all the columns. Add the columns. Calculate the new table to normalizing the scores; divide each value of a cell of a column by the column total. Likewise do for all columns. Add the rows of this new table. This will be the Normalized score for each parameter. Convert into percentage by dividing the normalized score for a parameter with the column total of the Normalized Score Column and multiplying by 100. This will be the Percent Ratio Scale Of Priority (PRSP) for each parameter and will also be the priority of your customer.

TABLE I. THE SCORE MATRIX

	X	Y
X	1	3
Y	1/3	1
Sum	1.3	4

TABLE II. NORMALIZED AND PRIORITY TABLE

	X	Y	Sum	Priority
X	1/1.3	3/4	1/1.3+3/4	(1/1.3+3/4)/S1*100
Y	0.33/1.3	1/4	0.33/1.3+1/4	(.33/1.3+1/4)/S1*100
Sum	1/1.3+.33/1.3	3/4+1/4	S1	

IV. Semantic Similarity

The semantic similarity techniques are used to determine how two concepts or terms are similar, they are used in many applications such as intelligent information retrieval, knowledge integration systems, sense disambiguation, classification and ranking, detection of redundancy, and detection and correction of malapropisms [26,27]. Semantic similarity between words is measured by using semantic web (ontology) which define words with their define meaning, and describes the relationships between terms or concepts and their properties.

There are many techniques used to semantic similarity using domain ontology, wordnet, and corpus. Also semantic similarity can be measured based on the information content based approaches that use ontology structure and corpus-based features such as Resnik [28], Jiang & Conrath [29], Lin [30], and structure based approaches such as path length [31], Leacock.& Chodorow [32], Wu & Palmer [33].

Semantic similarity is important approach in information retrieval, semantic similarity can evaluated using page count, and text snippets retrieved from search engine for two terms. Using page count to count the result of searching of each term alone, and pages contain two terms to evaluate how they depend or independent terms [34]. Google used to evaluate semantic relatedness to calculates the similarity between two words, and distance between them [35].

In text snippets retrieved , searching about two terms and extract a snippet from results such as Wikipedia pages for two terms and processing the result to extract only the main terms in original form, then using a five similarity measure of association that is simple similarity. Jaccard similarity comparing the similarity and diversity of given sample set. Dice similarity also related to the jaccard measure. Over Lap method is used to find the overlapping between the two sets. The cosine similarity is a measure of similarity between two vectors of n dimensions by finding the angle between them [34].

V. The Proposed Search Engine Technique

The proposed engine enhances a search engine through two methods. The first is the disambiguation of query terms by expansion process using general purpose ontology and domain ontologies selected by searching in the domain it is dealing with. The domain ontology is selected by searching in the domain dealing with and taking into consideration the relation of expanded terms through ontology domain description. The second method improves the ranking process taking into account the semantic relation between terms found on the page. This engine retrieves a high amount of the available semantic documents and enhances current search technology on the web. It performs the basic functionalities of the traditional search engine including: crawling web documents, indexing, ontology selection, query manipulation and expansion, and thus ranking documents.

As Fig.1 depicts, the architecture of the proposed engine indicates the two suggested methods, each of them composed of some modules. The Search engine has a main module that is a user interface module, and an additional module that is semantic search for ontology domain search.

- User Interface Module: is an easy interface for user to enter their queries and show required results.
- Semantic Search Module: In this module, the process of searching for the semantic documents is related to the domain search using the user queries to provide a suitable ontology.

A. *The Query Expansion Method*

This method is an expansion query process to disambiguate the query terms and to explain the meaning of query terms using their synonyms from WordNet ontology (general purpose ontology), and their related terms from domain ontology taking into account their relationships. It consists of three modules fig.2: query manipulation (expansion), semantic query and weighting module (building tree model, using AHP algorithm).

- Query manipulation: In this module, query is interpreted by performing preprocessing, stemming and disambiguating the query. Disambiguating the query is done by adding semantic meaning to terms with their synonyms using general purpose ontology (WordNet).
- Semantic query Module: In this module after connect to WordNet to extract the synonyms for each query terms and based on domain ontology extract hyponyms for query terms and their sibling, we construct all semantic meaning to query terms as a vector of terms.
- Weighing Modules: consist of two parts.

The first step is building a hierarchy tree based on domain ontology and the synonym terms in two-level trees. A tree model is a technique used to build a tree with multi-levels. All terms of an expanded query are organized as a tree with multiple levels regarding their hierarchical relationships defined in a selected ontology. In this model, the synonyms are located at the same level as the query terms and the hyponyms are distributed at a lower level. The relevance scores generated by those expansion terms and documents are evaluated upon the degree of relation between terms, original query and documents.

The second step is to evaluate the weight values based on AHP algorithms. AHP is a multi-criteria decision support methodology used in management science. We estimate the mutual importance values between relevance generated by original query terms and synonyms and hyponyms estimated based on the AHP score [36]. Where the original query terms and their synonyms are in the same degree of importance, but their hyponyms terms have different degree.

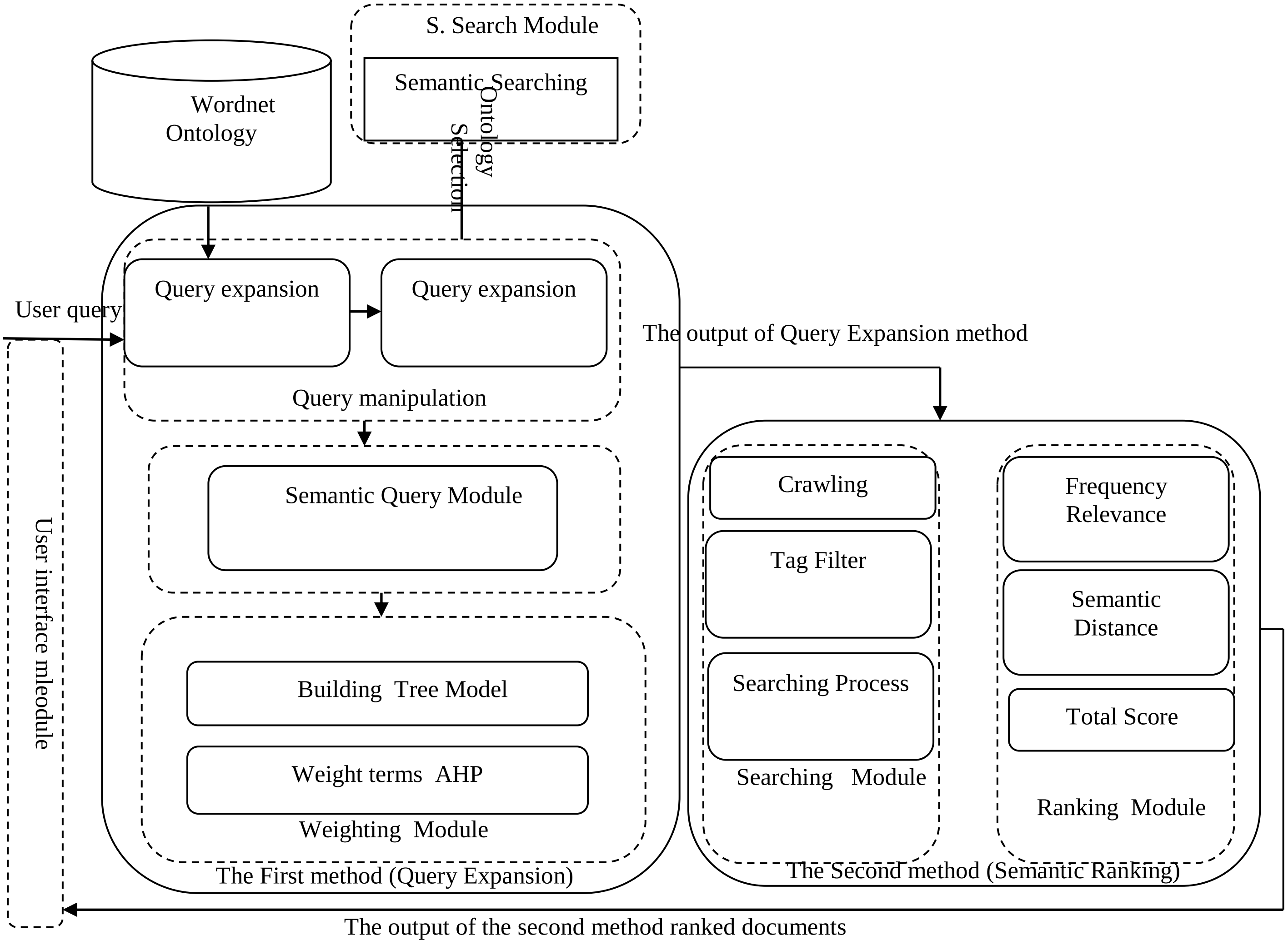


Fig. 1. Proposed System Architecture

B. The Semantic Ranking Method

Ranking process is considered an important step in any search engine. A good search engine is evaluated by whether the user's requirement exists in relevant documents which are returned, and evaluated by ranking techniques. This method consists of two modules a Searching Module, and Ranking Module (Semantic distance in content, term frequency) as shown in fig.3.

- Crawling: crawling the documents and indexing them [37,38].In crawling we based on crawler built using java code enter a start url and extract a list of urls from pages ,indexing process by parsing url document using jsoup java tools that deal with html pages ,it parsing html based on tags ,which allow us extract each text tag separately, split them based on (. dot) for each statement or (" " space) for terms , removing stopwords and stemming them ,calculate the frequency of each term and storing them in database.
- Tag Filter: Most information are represented in internet pages in HTML documents, which it contains a set of markup tags that represent the content. These tags have different priorities in documents. Many retrieval information works deal with tf (term frequency),VSM(vector space model) and many other techniques deal with all document as a whole.

But HTML have many parts (tags) which mean different priorities, such as the document that have query term in title tag mean related to the query more than the document have a query term in other tag, the query term in <a href> is related to another page that explain it in detail ,and so on, then it becomes difficult to weight all document as the same in final ranking [39]. Due to the above mention some works deal with document as classify document based on tags, but it deal with single query term [40], also dealing with document tag by adding extra weight to term found in special tag [41].In this paper we deal with main document tags (title, head, body) construct a weight to each tag based on AHP dealing with semantic distance between query terms in each tag.

We implement our system using jsoup as a tool in java working with real world HTML(jsoup: Java HTML Parser),it provides a way to extract html tags ,extract the text for each tag select("title"), select("body"), select("head")), then split text by " ,_-" any special characters, remove stopwords ,and stemming each term to restore in original form, connecting with WordNet ontology to return the synonyms for each term. All these data are stored in database relate the terms to original text contains and in which tag, to measure the semantic distance.

- Searching Process: Searching for documents that have query terms and their expansion and taking into account their frequency of each term found.

- Ranking Module: Ranking plays an important role in searching. In this paper the documents are evaluated based on the semantic relation between terms in statements (semantic distance) and term frequency. The related terms found are weighted based on the result of the tree module .These two values are calculated according to the following subsections.

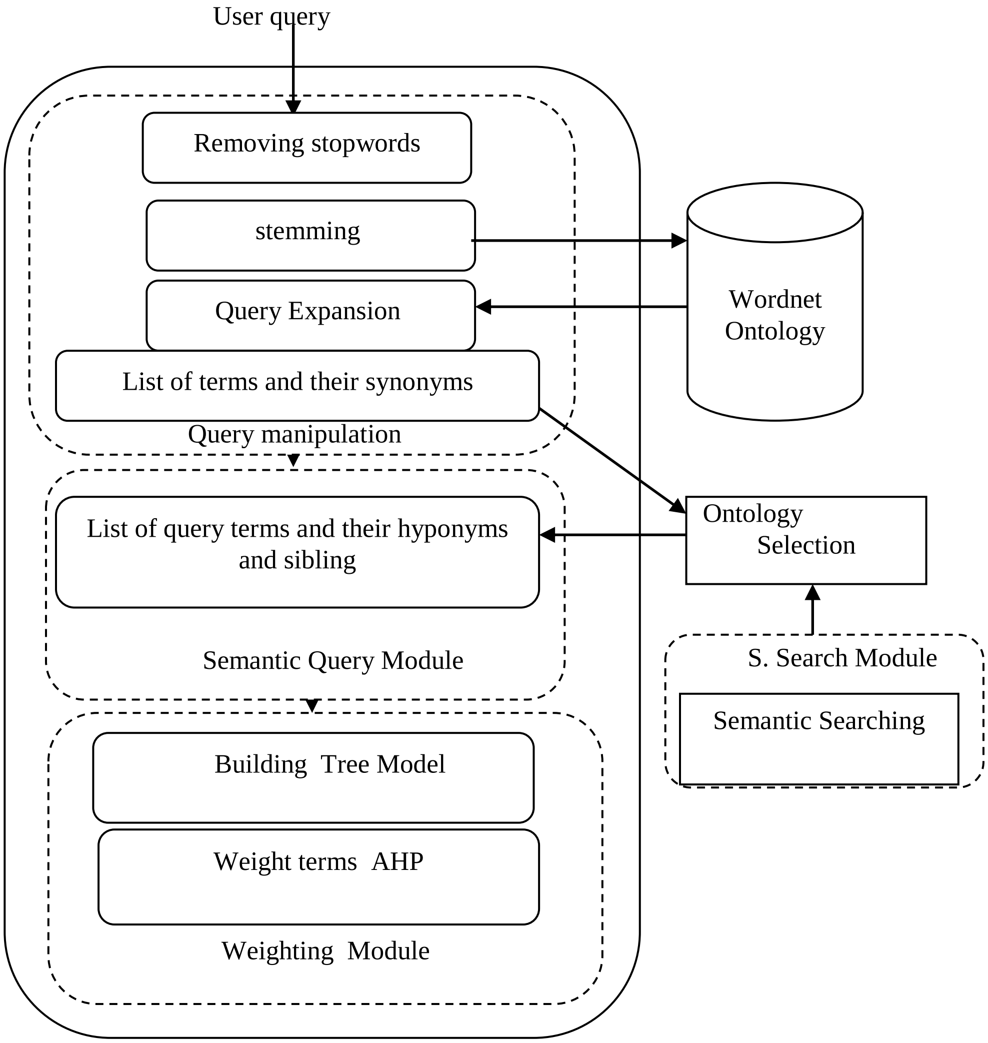


Fig. 2. The first Part of query expansion and weighting

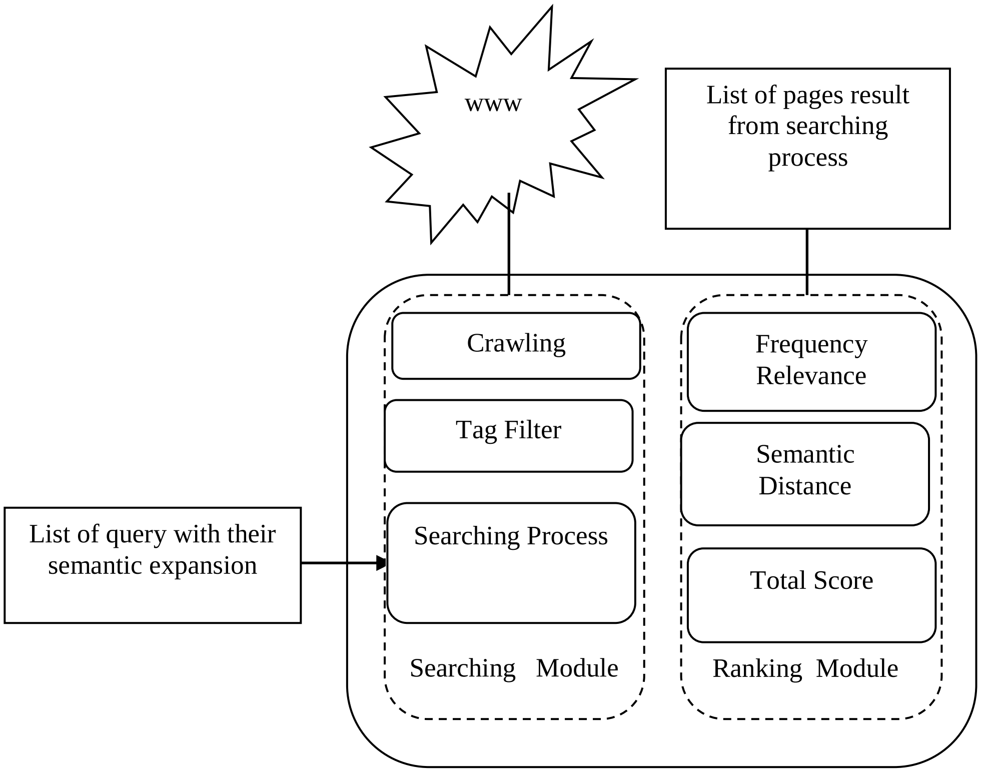


Fig. 3. The Second part Ranking Process

1) Frequency Relevance

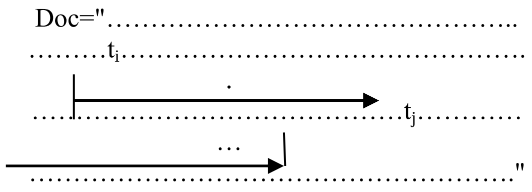
Frequency is used to evaluate how documents are related to the user query, by searching in the document for the number of occurrences of the required terms. In the previous works, they took into account the summation of the frequency of all terms found. But in our proposed method, due to the expansion of query; we take into account the semantic relationships (synonyms, hyponyms) to ensure that the page is related to the domain selected. We weight each frequency term to indicate their priority on the page. fi is the frequency of term i, so the total frequency relevance is the summation of all query terms.

$$F(D)=\frac{\sum_{i=1}^{k} w_i\ f_i}{|D|} \quad (1)$$

w_i : weight value estimated using AHP algorithm, f_i : frequency of term i, K: number of query terms, |D| :is the length of document to make normalized for total frequencies of terms.

2) Semantic Distance Function

Using the frequency relevance for query terms may introduce multiple topics and irrelevant information within relevant documents. In order to provide the semantic distance between two terms, the weights of their hierarchical structure in documents are taken into account. We measure the distance between query terms and their hyponyms found in documents; the terms that have higher distance between them become less related terms. The distance function is a weighting function to measure the semantic distance between the terms of queries and their hyponym found in the document. Where the frequency based of terms dealing with terms in any position within the documents, whatever these terms are related to each other or not. So, it is important for assessing if a term is close enough to query terms and their expansions, which indicates if the document related to specific topic or not, the position distance is adapted from one proposed[42]. Based on the relevance model, the main idea of the positional relevance model (PRM) is to further distinguish different positions of a term and discount the occurrences of a term at positions that are far away from a query term in a document. We modify this work to be suitable on document ranking and semantic distance as follows in (2).



We calculate the position between two terms, based on the semantic distance between two concepts in ontology which is calculated by measuring the distance (length of path between two concepts). We estimate the distance between two terms by the length between terms in statement.

$$P_{pos}(t_1|D) = \sum_{j=1}^{m} SD(t_1, t_j) = \\ = \sum_{j=1}^{m} (\frac{1}{len\left(t_1 \xrightarrow[j=1:m]{} t_j\right)})/|ST| \quad (2)$$

Where, SD(t_1, t_j) is the semantic distance between two terms t_1,t_j, where j terms from 1 to m (m: number of all query terms and their expansions); |ST| length of statement.

For each term t_1, we measure the distance between it and all the other terms, their synonymous, other query terms and their expansions (t_1 and t_j). For each sentence or statement or paragraph separated by (.,\n) are splitted remove the stopwords and return each term to their original form and measure the length between terms. $\mathrm{len}(\mathrm{t1} \xrightarrow[\mathrm{j=1:m}]{} \mathrm{tj})$ is the length between ti and tj. Because we deal with different length statements, normalize this result by divide by the length of statement |ST|.

In our proposed method, we calculate the semantic relation between query terms in each part in web page (title, head, and body). The semantic relation of each part is calculated as shown in (3):

$$\mathrm{SR(D)} = \sum_{\mathrm{h=1}}^{3} \mathrm{w_h} * \sum_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{w_i} * \sum\nolimits_{\mathrm{j=i}}^{\mathrm{k}} \mathrm{SD}(\mathrm{t_i}, \mathrm{t_j}) \qquad (3)$$

Where, SR(D): the semantic relations, k: number of query terms with expansion, SD(t_i,t_j): semantic distance between terms shown in (2), w_i : the weight of t_i, w_h is the weight for each tag in html document contain query terms.

For each term, calculate the semantic distance between this term and all other searched terms and their synonyms and hyponyms for each part in documents. If the term occurred many times in the paragraph or sentence, we deal with each statement separately, if no terms we deal with the paragraph as a whole.

The total semantic relation for each page is calculated as the summation of semantic relation for three weighted parts using AHP algorithms that indicate title tag with higher priority than body tag, and body tag with higher priority than the head tag (0.607002, 0.303344, and 0.089654).

3) *Total Score*

Based on the pervious notice ranking document based on the term frequency or cosine similarity between query terms and document contents, they does not take into account the semantic relation between terms found in documents, so to aggregate the advantage of the occurrence of query terms and how they are close related to each other, we add two values of frequency and semantic distance between query terms as shown in(4)

$$\mathrm{R(D)} = \mathrm{w1} * \textstyle\sum_{\mathrm{i=1}}^{3} \mathrm{SR(D)} + \ w2 * F(D) \qquad (4)$$

Where, $\sum_{\mathrm{i=1}}^{3} \mathrm{SR(D)}$: Is the Semantic relation calculated for document D for each part (title, head, body) in HTML documents, F(D) is the total frequency of terms found in documents.

VI. Implementation

Implementation is the stage of the project when the theoretical design is turned into a system. It can be considered the most critical stage in achieving a successful new system to give the user confidence that the new system will work and will be effective. The proposed system was implemented using Java and JENA software as a simulator.

As previously explained, the implemented system consists of some steps; it starts with the crawling process. Multi-threaded, multi crawlers were implemented to crawl both traditional and semantic web. The implemented crawlers make use of seed URLs to extract semantic web (pages with extended owl type). Another option for crawler is to enter the domain or keywords to search in swoogle API or Google API to search for all files by (file type: owl). After collecting web documents, their content is then parsed using Jena software to extract all semantic web details (concepts/classes, relations, instances /individuals). These data are stored in the index.

When the user enters his query through an interface, the implemented system comes with capabilities that enable the user to identify his intent by disambiguating his query using the WordNet database to extract the synonyms of his query. The important part of the interface is a list of semantic web document results with a summary such as their content description class number, property number and instances number. As well, it allows an immediate preview for related data found in the ontology, if the user selects one. The semantic web document with a high ranking is selected and is used as a domain ontology description to expand a user's query. After expansion, a tree of expanded terms is built and the weights are evaluated for expanded terms using AHP process.

A. *Data Sets*

Implementing our proposed system with the three real-world data sets below:

- Academic Staff & University (Staff): academic staff's full names (from 20 different universities) and their universities have been collected. (D1)
- Drug & Disease (Drug): This data set contains 200 drug names and the names of 183 different diseases they can cure. It was extracted from a drug list. (D2)
- Invention & Inventor (Invention): This data set contains 512 inventions' names with their chief inventors' full names (311 different people) from an inventive list on Wikipedia. (D3)

For D1, database is collected by crawling documents based on Google search engine for university (seed url).D2 is collected by crawling documents form Wikipedia site for pharmaceutical products, and Google search for drugs and diseases. The ontology which is used to expand query and build a tree to determine their priority is selected by user, when he/she enters his/her query search engine for semantic documents are working to present a ranked semantic documents and allowing immediate preview for ontology descriptions related to this query.

B. *Results*

For the experiment, two parameters are used to evaluate the information retrieval system. These two values are precision P and recall R, where Nc is the number of correct web pages returned, Nr is the number of related returned web pages, but they were not necessarily the correct web page, and Nt is the number of total returned web pages.

The documents are crawled and are stored in database. These documents are classified as 'relevant' and 'non-relevant'. The judgment for relevant and ranking the documents collect for D1 based on ranking result of (http://www.arwu.org/), for D2 the relevant and non-relevant ordering of documents are based on Google search.

The two values precision and recall are calculated using the following equations:

$$P = \frac{Nc}{Nt}$$

$$R = \frac{Nc}{Nc + Nr}$$

$$f - mesure = \frac{2*p*R}{P+R}.$$

To measure the performance of the suggested ranking method, there are four different documents search engines, named SR1, SR2, SR3 and SR4; respectively. They are implemented using Java, where SR1 represents a traditional keyword-matching search engine based on the user query terms only, which does not employ any QE techniques. SR2 is a search engine based on expanding user query taking into account their synonyms, SR3 is a search engine that does the search process by expanding queries based on the pervious retrieval pages for that domain based on the relative terms and their frequency[43]. SR4 uses the proposed ranking method based on expanding queries by disambiguating their meaning with synonyms using WordNet and their subclasses from a domain ontology taking into account their relative weights to that expanded term.

In our system, we evaluate the ranking by using the relative weights for expanded terms to measure how documents are related to query terms based on the priority of terms founds evaluated using AHP algorithm.

In D1 for example we search about academy staff by query terms "academy staff & university" are expanded based on the selected ontology to (staff, university, academia, faculty, research, clerical staff ,system staff, professor, research assistant, administrative staff, chair, dean, teacher ,organization, affiliated organization, course, lecture) with a relative weight to indicate the importance and their priority in documents(0.094811868,0.090237076,0.073245559,0.084874439,0.023292748, 0.008797915,0.099963995,0.05668714,0.053551544,0.028051472,0.037997043, 0.062339604, 0.067050112,0.021785616,0.02354974,0.037918655, 0.07472).

SR3 method based on the related terms from the previous query result, for the same search query "academy staff & university" which is expanded to (university, professor, school, faculty, technology, department, Dr., institute, lecture, PhD, edu). In the first method, our search is based on only the query terms; in SR2, the ranking method is based on the query terms and their synonyms.

In our method, the searching and ranking process does not only depend on the terms found or on their frequency, but it also takes into account the importance or priority of the expanded terms through domain ontology with the relationships, synonyms and hyponyms of a query term. This process is done by weighting values to indicate the important terms. SR3 is based on the related terms from the previous query result. For the same search query "academy staff & university" which is expanded to (university, professor, school, faculty, technology, department, Dr., institute, lecture, PhD, edu). In the first method, our search is based on only the query terms; while in SR2, it is based on the query terms and their synonyms.

In our method, the searching and ranking process not only based on the terms found or on their frequency, but it also takes into account the importance or priority of the expanded terms through domain ontology with the relationships synonyms and hyponyms of a query term, by weighting values to indicate the important terms.

TABLE III. THE PRECISION OF COMPARISON METHODS

	SR1	SR2	SR3	SR4
D1	0.53	0.67	0.73	0.9
D2	0.62	0.6	0.66	0.9
D3	0.55	0.53	0.35	0.9

TABLE IV. THE RECALL OF COMPARISON METHODS

	SR1	SR2	SR3	SR4
D1	0.588889	0.744444	1	1
D2	0.688889	0.666667	1	1
D3	0.611111	0.588889	1	1

TABLE V. THE F-MEASURE OF COMPARISON METHODS

	SR1	SR2	SR3	SR4
D1	0.557895	0.705263	0.843931	0.947368
D2	0.652632	0.631579	0.795181	0.947368
D3	0.578947	0.557895	0.518519	0.947368

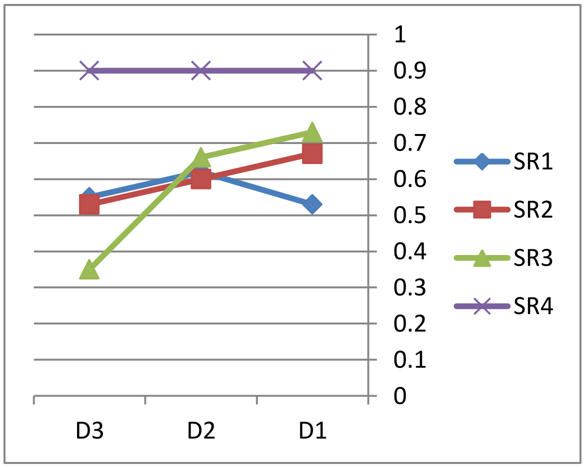


Fig. 4. The Precision of comparison between methods

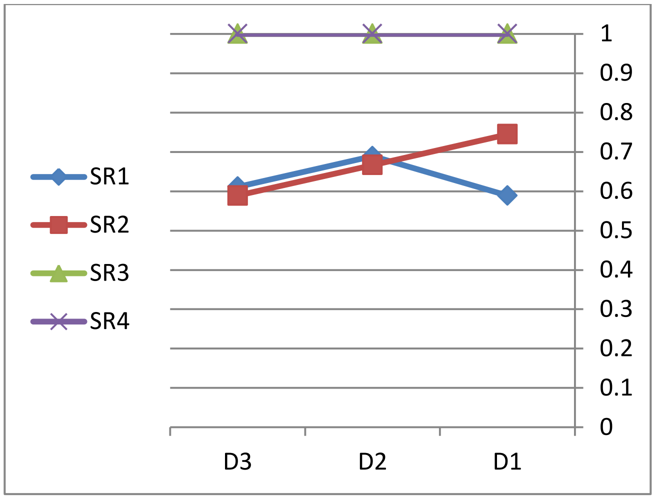


Fig. 5. The Recall of comparison between methods

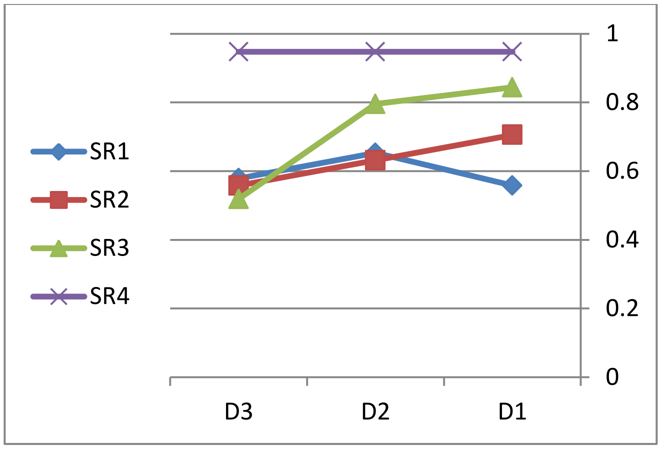


Fig. 6. The F-measure of comparison between methods

In SR1, the search results depend only on the query terms so it holds only the document; contain query terms unless it doesn't relate to the domain searched. In the SR2 and SR3, the relevant documents are increased based on the expansion, query terms by synonyms and related terms from pervious query results respectively. While in SR3, it depends on the good pervious results.

We notice that the expanded query in the SR3 method has the same recall as the proposed method, but it is still controlled using the related terms that expand from the previous query results.

SR4 depends on the expanded terms using domain ontologies that are searched for by our system, controlled by multiple parameters such as: properties of the concepts, properties and instances searched in ontologies (details of domain description).

For measuring the semantic similarity based on html tag, we take only three main tags (head, title, body)tags ,with the pervious weights .we measure the recall and precision for html documents based on tags, We notice the precision and recall of semantic similarity is increased based on numbers of query terms found in parts in html pages as shown.

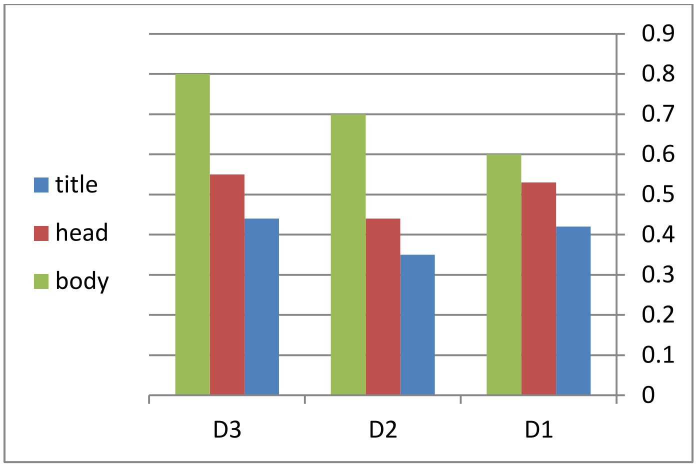


Fig. 7. The precision based on semantic distance using html tags.

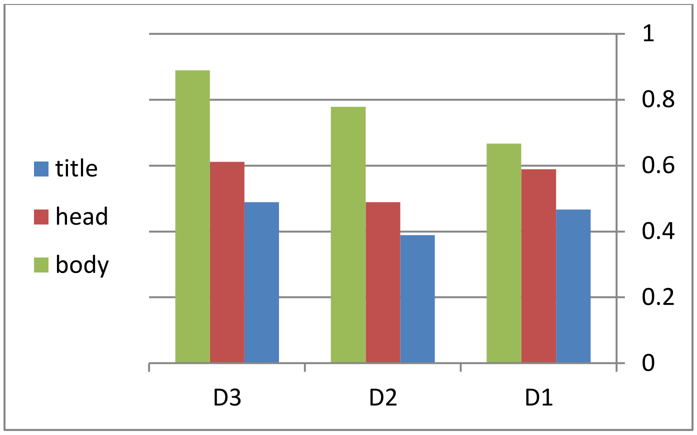


Fig. 8. The recall based on semantic distance using html tags

VII. Conclusion

In this paper, a system is proposed to improve the search process to overcome the traditional search problems by some methods, such as enhancing the expression of what the users actually mean and enhancing the evaluation process of the documents returned to users. The process of query expansion can be done using relevance feedback-based, statistical co-occurrence-based and domain ontology. But in the case of using domain ontology while dealing with all expanded terms from ontology that has different relationships with the same weighting which will affect in the evaluation to documents that contain them. The new proposed method used to search based on ontology and expanded query with domain ontology and ranking document taking into account the related weights in expanded terms as in the ontology domain in the hierarchical structure. These weights will affect the document accuracy related to the user main query terms.

VIII. Future Work

In this work we focus on single ontology for single domain. In future, we will focus on multiple ontologies which allow us to give an opportunity for employing the knowledge from different ontologies of single or different domains. Also, we will take into account the important html tags such as link (<a href> <.a>), bold tag <B>.

Reference

[1] Preethi , Ms.N., and Devi , Dr.T., Case and Relation (CARE) based Page Rank Algorithm for Semantic Web Search Engines. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

[2] Lee, T. B., Hendler, J., and Lassila ,O.,"The semantic web". Scientific American, vol. 284(5), May 2001.

[3] Ch.-Qin Huan,Ru-Lin Duan, Y. Tang, Zhi-Ting Zhu, Y.-Jian Yan, and Yu-Qing Guo, ."EIIS: an educational information intelligent search engine supported by semantic services".International Journal of Distance Education Technologies ,January 1, 2011.

[4] Robin Sharma , Ankita Kandpa,and Priyanka Bhakuni, Rashmi Chauhan, R.H. Goudar and Asit Tyagi." Web Page Indexing through Page Ranking for Effective Semantic Search". Proceedings of7'h International Conference on Intelligent Systems and Control (ISCO 2013).

[5] Yuan LIN,Hongfei LIN, and Li HE." A Cluster-based Resource Correlative Query Expansion in Distributed Information Retrieval ".Journal of Computational Information Systems 8: 1 ,2012, 31–38.

[6] W. W. Chu, Z. Liu and W. Mao."Textual document indexing and retrieval via knowledge sources and data mining". Commun. Inst. Inf. Comput.Mach. (CIICM), Taiwan, 2002, 5, (2), pp. 135–160

[7] A. Vizcaíno, F. García, I. Caballero, J.C. Villar, M. Piattini."Towards an ontology for global software development". IET Softw., 2012, 6, (3), pp. 214–225

[8] N. Tyagi and S. Sharma."Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page". In International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[9] N. Duhan, A. K. Sharma and K. K. Bhatia."Page Ranking Algorithms: A Survey". In proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

[10] Vishal Jain, Dr. Mayank Singh."Ontology Based Information Retrieval in Semantic Web: A Survey ", I.J. Information Technology and Computer Science, 2013, 10, 62-69

[11] M.Yadav,and Mr. P. Mittal." Web Mining: An Introduction ". International Journal of Advanced Research in Computer Science and Software Engineering, 2013,Volume 3, Issue 3, March , ISSN: 2277 128X

[12] Md. Z. Hasan, Kh. J. A. Chisty and Nur-E-Z. Ayshik . "Research Challenges in Web Data Mining". International Journal of Computer Science and Telecommunications , Volume 3, Issue 7, July 2012.

[13] S. Pal, V. Talwar, and P. Mitra ."Web Mining in Soft Computing Framework : Relevance, State of the Art and Future Directions". In IEEE Trans. Neural Networks, 13(5), PP.1163–1177,2002.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web" . Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[15] P. Devi, A. Gupta and A. Dixit. "Comparative Study of HITS and PageRank Link based Ranking Algorithms". International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014.

[16] Ch. D. Manning, P. Raghavan and H. Schütze ."Introduction to Information Retrieval". Book Introduction to information retrieval Cambridge university press New York, NY ,USA , 2008, ISBN:05218657199780521865715.

[17] R. Baeza-Yates and E. Davis ."Web page ranking using link attributes" . In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters , 2004, PP.328-329.

[18] H Jiang et al."TIMERANK: A Method of Improving Ranking Scores by Visited Time". In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[19] J.Jayanthi, Dr.K.S.Jayakumar. "An Integrated Page Ranking Algorithm for Personalized Web Search". International Journal of Computer Applications (0975 – 8887), Volume 12– No.11, January 2011.

[20] K.-J. Kim and S.-B. Cho . "Personalized mining of web documents using link structures and fuzzy concept networks". Applied Soft Computing, Volume 7, Issue 1, January 2007, Pages 398–410.

[21] Guangyu Zhu and Gilad Mishne." Clickrank: Learning session-context models to enrich web search ranking". TWEB,6(1):1, 2012.

[22] Ahmad Kayed, Eyas El-Qawasmeh, and Zakaryia Qawaqneh."Ranking web sites using domain ontology concepts".Information & Management, 47(7-8):350–355, 2010.

[23] Yajun Du and Yufeng Hai. "Semantic ranking of web pages based on formal concept analysis". Journal of Systems and Software, 86(1):187–197, 2013.

[24] Wei Wang, Sujian Li, Jiwei Li, Wenjie Li, and Furu Wei." Exploring hypergraph-based semi-supervised ranking for query-oriented summarization". Inf. Sci., 237:271–286, 2013.

[25] N. Bhushan, K. Rai." Strategic Decision Making Applying the Analytic Hierarchy Process". http://www.springer.com/978-1-85233-756-8, 2004.

[26] A. Hliaoutakis, "Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline", Master’s thesis, Technical University of Crete,Greek, 2005.

[27] A. Budanitsky and G. Hirst," Evaluating WordNet-based measures of semantic distance", Computational Linguistics, vol.32,1, March 2006.

[28] P. Resnik., "Using information content to evaluate semantic similarity". In Proceedings of the 14th international Joint Conference on Artificial Intelligence, 448–453. Montreal, Canada, 1995.

[29] J. J. Jiang and D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,", Proc. ROCLING X, 1997.

[30] D. Lin, "An Information-Theoretic Definition of Similarity,". Proc.Int’l Conf. Machine Learning, July 1998.

[31] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets,", IEEE Trans. Systems,Man, and Cybernetics, vol. 9, no. 1, pp. 17-30, Jan. 1989.

[32] C. Leacock., M. Chodorow, "Combining local context and WordNet similarity for word sense identification," In Fellbaum, C., ed., WordNet: An electronic lexical database, pp. 265-283. MIT press. 1998.

[33] Z. Wu . , M. Palmer, "Verb semantics and lexical selection," In 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133–138,1994.

[34] Strube , S.P. Ponzetto, "Wikirelate! Computing Semantic Relatedness Using Wikipedia," Proc. Nat’l Conf. Artificial Intelligence (AAAI ’06), pp. 1419-1424.2006.

[35] E. Iosif ,A. Potamianos, "Unsupervised Semantic Similarity Computation between Terms Using Web Documents". IEEE transactions on knowledge and data engineering, vol. 22, no.11, November 2010.

[36] A. Awasthi ,S.S. Chauhan . "Using AHP and Dempster–Shafer theory for evaluating sustainable transport solutions". Environ. Model. Softw.,2011, 26, (6), pp. 787–796

[37] H. Hama ,Thi Thi Zin ,P. Tin "Optimal Crawling Strategies for Multimedia Search Engines". Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, September 12-September 14,2009.

[38] C. Castillo ."effective web crawling". SIGIR Forum, ACM Press, Volume 39, Number 1, New York, NY, USA, p.55-56 (2005)"

[39] S. Pathak, S. Mitra." A New Web Document Retrieval Method Using Extended-IOWA (Extended-Induced Ordered Weighted Averaging) Operator on HTML Tags". IOSR Journal of Computer Engineering (IOSR-JCE),e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, PP 65-74,www.iosrjournals.org,2014.

[40] J. Deng, L. Chen." Web Documents Categorization Using Fuzzy Representation and HAC". In: Proceedings of the IEEE First International Conference, vol. 2, 2000, pp. 24-28.

[41] Y. Bassil , P. Semaan ." Semantic-Sensitive Web Information Retrieval Model for HTML Document". European Journal of Scientific Research, ISSN 1450-216X, vol. 69(4), 2012.

[42] Y.Lv and C. Zahi. "Positional relevance model for pseudo-relevance feedback." In SIGIR, pages 579-586,2010.

[43] Z. Li, M. A. Sharaf, L. Sitbon, X. Du and X. Zhou." CoRE: A Context-Aware Relation Extraction Method for Relation Completion", IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING,2013.

The Effect of Parallel Programming Languages on the Performance and Energy Consumption of HPC Applications

Muhammad Aqib
Department of Computer Science
FCIT, King AbdulAziz University
Jeddah, Saudi Arabia

Fadi Fouad Fouz
Department of Computer Science
FCIT, King AbdulAziz University
Jeddah, Saudi Arabia

***Abstract*—Big and complex applications need many resources and long computation time to execute sequentially. In this scenario, all application's processes are handled in sequential fashion even if they are independent of each other. In high-performance computing environment, multiple processors are available to running applications in parallel. So mutually independent blocks of codes could run in parallel. This approach not only increases the efficiency of the system without affecting the results but also saves a significant amount of energy. Many parallel programming models or APIs like Open MPI, Open MP, CUDA, etc. are available to running multiple instructions in parallel. In this paper, the efficiency and energy consumption of two known tasks i.e. matrix multiplication and quicksort are analyzed using different parallel programming models and a multiprocessor machine. The obtained results, which can be generalized, outline the effect of choosing a programming model on the efficiency and energy consumption when running different codes on different machines.**

Keywords—power consumption; quicksort; high- performance computing; performance; Open MP; Open MPI; CUDA

I. Introduction

With the addition of multiple cores, the capability of chips to process multiple instructions simultaneously has increased the performance. High-performance computing provides boost in performance but at some stages, it requires more resources to increase the performance. To provide an optimal solution which could be running efficiently and consumes fewer resources, like energy etc. the performance of the computing system must be analyzed.

Multiple performance analysis tools could be used to test the performance of different software applications [1]. This kind of performance analysis studies help to improve the performance of the software application and to provide an optimal solution. Tools that are utilized for the performance analysis of HPC applications use different approaches for the analysis purposes [2].

In earlier work, performance analysis criteria was based upon the computation of speed, the number of threads generated to perform a task and how the memory was utilized to perform those tasks [3]. When considering HPC architecture, it is supposed that there are a large number of processors that are dedicated to performing the computation tasks. So there is an obvious increase in the consumption of the energy resources as well. So, in addition to the optimization techniques to improve performance, it is also necessary to use energy-aware techniques.

Many optimization techniques could be applied to the code to be running in parallel. For example, loop optimization techniques could be implemented to improve the performance of loops in a code. The use of different programming APIs or architectures like Open MP [4], Open MPI [5], CUDA [6] etc. provides the programmers and application developers with the ability to running different blocks of codes in parallel on CPUs and GPUs. These APIs also provide a mechanism to running the code in parallel using multiple cores in HPC environment.

In this paper Open MP, Open MPI, CUDA were used to perform simple computation tasks i.e. matrix multiplication and to sort using quicksort. Matrix multiplication task is considered as one of the expensive tasks as it involves nested loops and performs multiplication and addition of numbers. In both cases, the code was implemented in C++ to measure the computational time and energy consumption in sequential manner. Then parallel programming API's have been used to get the results while performing the same operations in parallel. Comparing the results obtained implementing different models used in HPC with the results in the sequential mode made it possible to analyze the effect of parallel programming languages on the performance and energy consumption in HPC environment.

The rest of the paper is organized as follows. Section 2 describes the work done by other researchers to analyze the performance of parallel programming models and techniques. Section 3 discusses different tools/models/APIs available for parallel programming. In Section 4, the performance analysis model adopted in this study is presented. Section 5 contains the results obtained using different APIs. Section 6 discusses the results presented in section 5. Finally, conclusion and future work is presented in Section 7.

II. Related Work

Many researchers have provided energy consumption analysis of machines having HPC capabilities. Rejitha *et al.* have analyzed the effect of loop optimization techniques on the use of energy consumed by different techniques [7]. Although they have compared different such techniques but they did not implement them using all available models in HPC environment. In [8], authors have implemented MPI based solutions using different loop optimization techniques. But, their results are also limited to the use of MPI model.

Freeh *et al.* have directly measured the time and energy with the help of power meters consumed by the AMD-64 nodes [9]. The effects of bottlenecks in the memory and communication in these nodes have been measured. According to them, there is a trade-off between time and energy consumed by HPC applications. i.e. If bottleneck problem arises in any node, it will increase the amount of energy consumed for that application. But this could be reduced by increasing the execution time for that application.

Feng *et al.* have emphasized on the need to characterize the power characteristics of high performance applications to control the energy consumption of future HPC applications [10]. According to them, the operational costs to run an application depends on the characteristics of that application. Even if two applications are running on a system for the same amount of time, the energy consumed by them may differ depending upon their characteristics.

In [12, 13], different techniques to estimate energy consumption in embedded systems have been discussed. Although embedded systems in general are different from high performance systems they have a common case i.e. in both systems energy consumption is a critical issue. So the techniques used for comparison in embedded systems may give an idea on how to estimate the energy consumption in HPC systems.

Enos *et al.* in [14] have provided a mechanism to monitor the energy consumed by CPUs and GPUs installed in a HPC machine. This approach is capable of calculating the power consumption by individual CPUs and GPUs. For this purpose they have used hardware devices and other equipment to monitor the power consumed by the system components. In this paper, a software mechanism has been provided to measure the energy consumption.

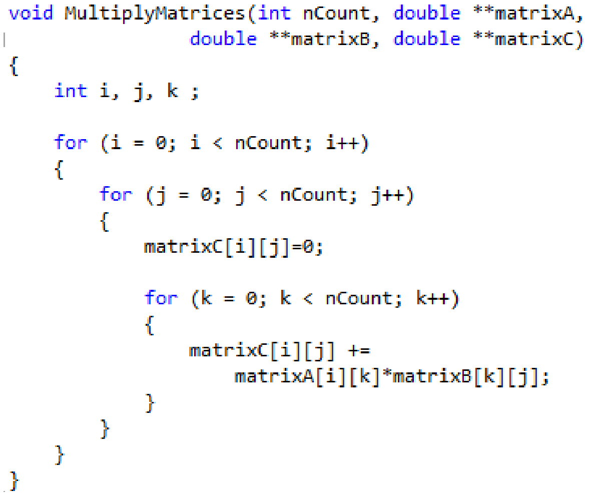
A very recent work done by Rashid *et al.* [15] provides an analysis of different sorting algorithms. They have implemented these algorithms on ARM based devices. So this work is basically related to mobile devices. But they have identified some factors which affect the energy consumption in those devices. According to them, algorithm implemented to perform a task and the language used affect the energy consumed by that application.

A model to calculate the energy complexity of different algorithms has been proposed in [16]. Although this is not directly related to high performance computing, it provides a model which deals with the energy consumption and the memory layout which is divided into two layers in this model.

III. Parallel Programming Models and Energy Consumption Analysis Tools

In this section, the parallel programming models and the software tool that were used to get data related to energy consumption analysis have been described. For parallel execution of the code blocks used in the experiments different programming models have been used.

Fig. 1 shows a simple code in C++ language to perform matrix multiplication without any optimization. All the loop instructions in this code run sequentially. Even if this code is run as it is on a multiprocessor machine, it will take the same time to execute.

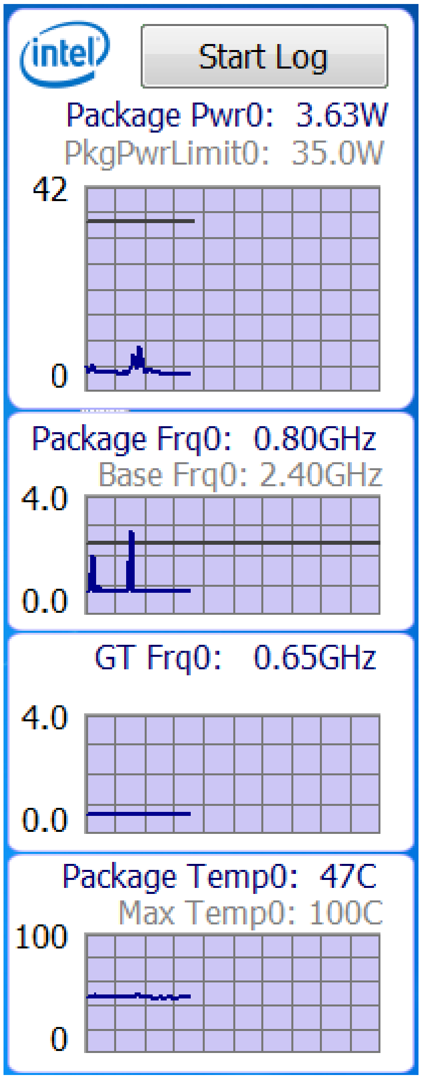


```
void MultiplyMatrices(int nCount, double **matrixA,
                      double **matrixB, double **matrixC)
{
    int i, j, k ;

    for (i = 0; i < nCount; i++)
    {
        for (j = 0; j < nCount; j++)
        {
            matrixC[i][j]=0;

            for (k = 0; k < nCount; k++)
            {
                matrixC[i][j] +=
                    matrixA[i][k]*matrixB[k][j];
            }
        }
    }
}
```

Fig. 1. Matrix multiplication code in C++ without optimization

Different energy consumption analysis tools have been used by researchers to measure the energy consumed during the execution of the code. In this study, Intel Power Gadget 3.0 [11] have been used for this purpose. It is a power monitoring tool developed by Intel. It supports second generation Intel Core processors to monitor the power consumption in that system. Desktop view of this gadget is show in Fig. 2.

Intel Power Gadget GUI have four different sections that shows different readings. “Package Pwr” section shows the overall power consumption and the average power limit. Current CPU frequency is shown as the “Package Frq”. If GPU is attached with the system, its frequency is shown under the label “GT frq”. The overall system temperature is shown in the section named “Package Temp”. It shows both, the current temperature and the max. temperature limit.



Package Pwr0: 3.63W
PkgPwrLimit0: 35.0W

Package Frq0: 0.80GHz
Base Frq0: 2.40GHz

GT Frq0: 0.65GHz

Package Temp0: 47C
Max Temp0: 100C

Fig. 2. Intel Power Gadget 3.0

This gadget generates the energy consumption log that provides the power consumption statistics. Log file includes the elapsed timed, package power limit, processor frequency, GT frequency, processor temperature, average and cumulative power of the processor [11]. For the purpose of this study, the “Processor Energy” have been used. This gives the total energy consumed by the processor including the energy consumed by processor cores, GPU, and by other devices.

To run the above code in parallel mode, different parallel programming models have been used. The same code has been implemented using C++ compatible APIs for each parallel programming model. The code has been implemented using Open MP, Open MPI and CUDA. In the following subsections, a brief introduction to these parallel programming models is given.

A. OpenMP

Open MP provides a set of compiler directives. It also includes a set of runtime library routines that are implemented using Fortran, and C/C++. These routines provide support for the parallelism using shared memory model [4].

B. Open MPI

The Message Passing Interface has been implemented in the form of Open MPI [5]. It fully supports the multithreading approach and could be used to develop applications that support concurrent access to memory. It also supports the old versions of MPI like LAM/MPI, LA-MPI and FT-MPI. It also provides options to check the data integrity for processes running in parallel.

C. CUDA

Compute Unified Device Architecture (CUDA) is also a parallel programming model and it is developed by NVIDIA. It runs on a graphical processing unit that supports CUDA. For parallel processing, it provides direct access to the virtual instruction set of GPU [6].

IV. Proposed Model for Energy Consumption Analysis

In this section, the model which was used to perform the analysis and the computing system specifications are presented. To run the programs a multicore hyper threaded machine has been used. The System specifications for that machine are given in the following table.

TABLE I. SYSTEM SPECIFICATIONS

Component	**Name / Capacity**
Operating System	Microsoft Windows 7
CPU	Inel® Xeon® CPU E5-2640 @ 2.50 GHz (12 CPUs)
GPU	Nvidia® Tesla K-40
Ram	8 GB
Analysis tool	Intel Power Gadget 3.0

A power consumption analysis model has been proposed. This model describes the process flow and all the steps performed during the analysis process. At the initial stage, before starting the program execution, the energy consumption analysis log needs to be started, and the destination folder for this log file to be selected. After starting the log, the program execution will start. But before starting the multiplication function, the execution time start will be recorded then the multiplication process will be started. After the completion of multiplication process, the time again will be calculated, and both starting and ending times will be written to a separate time log file. Now program will be terminated and the energy consumption analysis log will be stopped. After that, starting and ending time will be available in the time log file and from that time, the energy consumed during that period can be found. A flow chart describing this model is shown in Fig. 3.

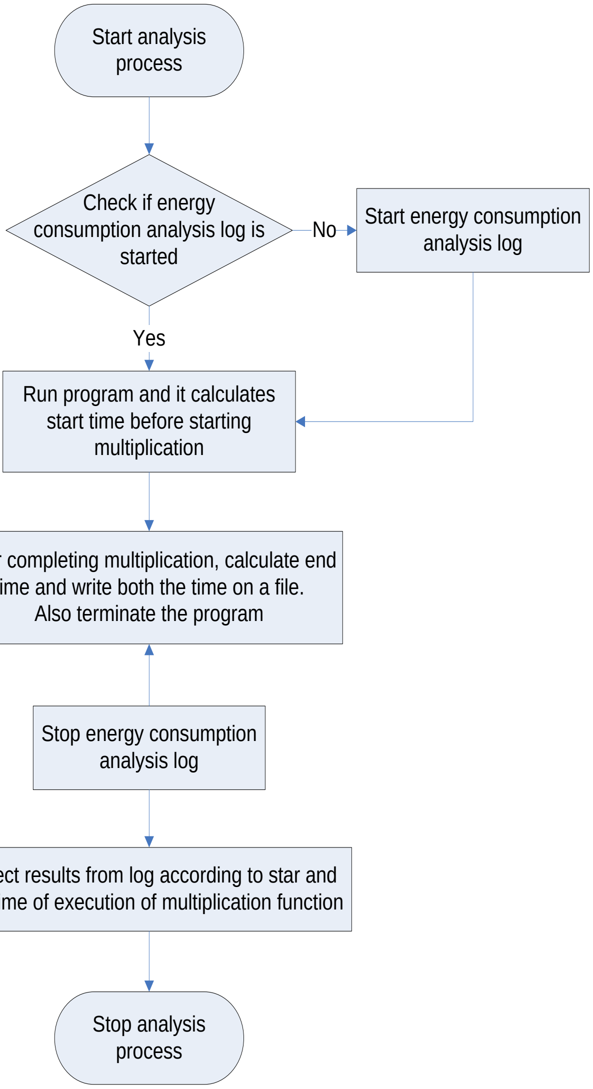

Start analysis process
Check if energy consumption analysis log is started
No
Start energy consumption analysis log
Yes
Run program and it calculates start time before starting multiplication
After completing multiplication, calculate end time and write both the time on a file. Also terminate the program
Stop energy consumption analysis log
Collect results from log according to star and end time of execution of multiplication function
Stop analysis process

Fig. 3. Proposed model for energy consumption analysis

V. Results

The following section presents the results obtained by running the matrix multiplication program for different matrix sizes and using different programming models. Also the results for running the quick sort algorithm for different array sizes and different programming models are given.

For comparison purposes, different matrix sizes that range from 500 × 500 to 5500 × 5500 have been used. Execution time has been recorded in seconds and the energy log sampling resolution was set to 500ms. This enables the monitoring of energy consumption and other related statistics twice a second. Table 2, and 3, show the results obtained by running each code to multiply square matrixes of five different sizes for each programming model (i.e. C++, Open MPI, Open MP, and CUDA). In table 2, the execution time consumed during the multiplication process is given.

TABLE II. Time Consumed by Different Programming Model to Multiply Matrices of Five Different Sizes

Matrix Size	**Time Consumption (sec)**			
	C++	**OpenMP**	**Open MPI**	**CUDA**
640 × 640	3.042	2.074	1.03	4.055
1280 × 1280	29.062	18.257	17.318	29.408
2560 × 2560	284.131	164.094	181.252	225.279
3840 × 3840	1236.349	571.047	650.066	755.43
5120 × 5120	3101.816	1922.444	1617.374	1789.212

Table 3 shows the results for the energy consumption analysis for the same set of data using the same models for matrix multiplication. Note that, For the purpose of energy consumption analysis, we have measured the overall energy consumed by the system.

TABLE III. Energy Consumed by Different Programming Model to Multiply Matrices of Five Different Sizes

Matrix Size	**Energy Consumption (mWh)**			
	C++	**OpenMP**	**Open MPI**	**CUDA**
640 × 640	24.051	16.408	8.184	37.497
1280 × 1280	280.361	186.676	158.254	302.709
2560 × 2560	2709.928	1723.636	1689.389	2347.044
3840 × 3840	11677.053	6028.354	6100.8	8019.528
5120 × 5120	29988.794	20762.018	16686.165	19439.64

For quick sort, array sizes have been considered between 128,00,000 to 1,024,00,000. Here it is worth mentioning that for the sorting comparison, array size for CUDA ranges from 12,80,000 to 102,40,000. Similar to matrix multiplication, execution time has been recorded in seconds and energy consumption resolution was also set to 500ms. These results will be discussed in detail in the following section.

TABLE IV. Time Consumed by Different Programming Model to Sort Arrays of Five Different Sizes

Array Size	**Time Consumption (sec)**			
	C++	**OpenMP**	**Open MPI**	**CUDA**
	× 10^5	**× 10^5**	**× 10^5**	**× 10^4**
128	60.312	112.142	12.012	84.087
256	229.315	431.019	41.058	608.026
512	901.225	1702.979	155.044	649.202
768	2016.177	3956.418	340.095	2133.056
1024	3564.942	6849.089	596.007	5869.163

In table 5, the results obtained by measuring the energy consumed by different programming models to sort the arrays of different sizes have been presented. Same like matrix multiplication, the sampling window was set to 500ms to collect the data for energy consumed by different programming models using the quick sort.

TABLE V. Energy Consumed by Different Programming Model to Sort Arrays of Five Different Sizes

Array Size	**Energy Consumption (mWh)**			
	C++	**OpenMP**	**Open MPI**	**CUDA**
	× 10^5	**× 10^5**	**× 10^5**	**× 10^4**
128	551.37	1067.009	94.185	10258.337
256	2740.969	5517.866	1575.13	16745.233
512	11410.318	22433.522	3516.288	24443.189
768	30765.32	60925.113	7831.1	45670.962
1024	64959.75	129887.901	14293.997	103397.188

Figures 4 and 5 show the results obtained by running the matrix multiplication code using the four programming models. Time comparison has been given in Fig. 4, whereas the energy consumption analysis is shown in Fig. 5.

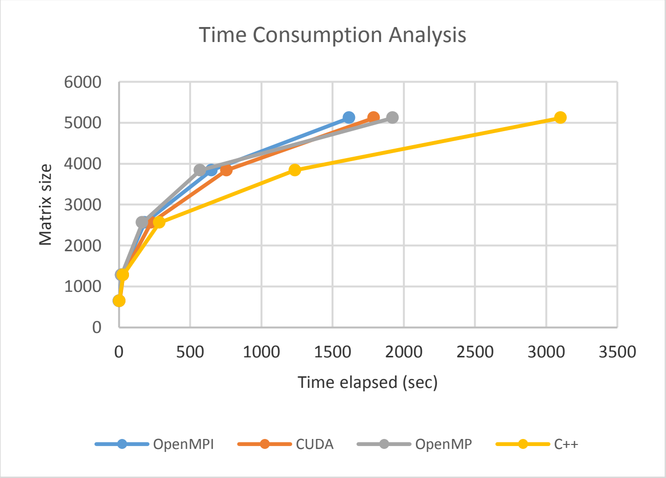


Figure 4. Time efficiency comparison of all four types for matrix multiplication

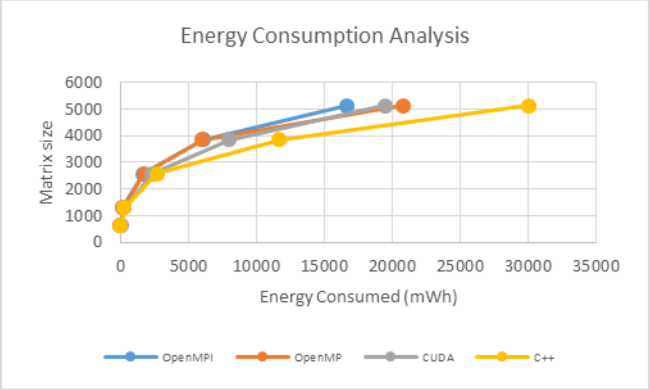


Figure 5. Energy consumed by four models for matrix mulitplication

Figures 6 and 7 show the results obtained by running the quick sort algorithm that is implemented using the four programming models. Time comparison has been given in Fig. 6, whereas the energy consumption analysis is shown in Fig. 7.

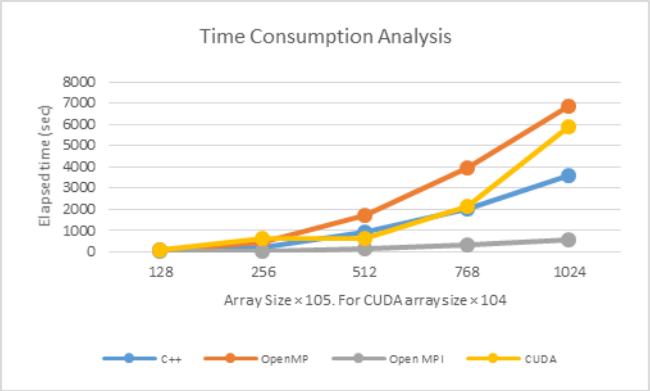


Figure 6. Time efficiency comparison of all four types for quick sort

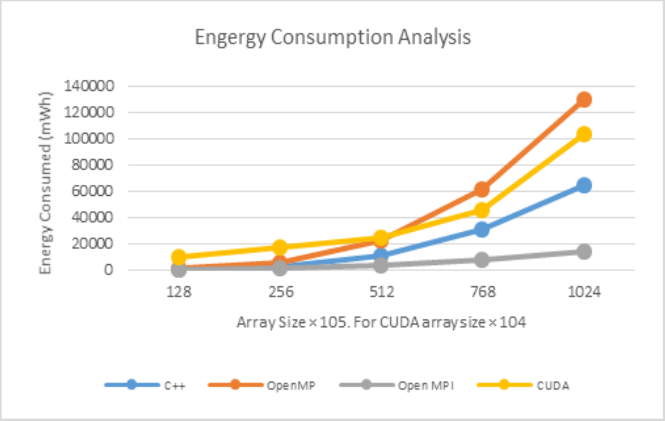


Figure 7. Time efficiency comparison of all four types for quick sort

VI. Discussion

The main purpose of this work is to analyze the performance and energy consumption analysis of different parallel programming models using the computing system and the model described in the previous sections. For this purpose, matrix multiplication and quick sort algorithm have been used. It is obvious that the parallel programming models improve the efficiency and reduce the energy consumption only if there are some blocks of codes that could be parallelized. For example, in matrix multiplication, it is not possible to run all the instructions in parallel, but as the multiplication takes place in the form of rows * columns, so this task could be assigned to multiple threads to run in parallel. Results shown in Fig. 4 and Fig. 5, show that models that support parallel execution of multiple threads produce good results when matrix size is large. For small matrix size, the time and energy consumption is same for all models. And even in some cases, sequential execution is better than the parallel. But when the size increases, the parallel execution produce good results both in terms of time and energy. The results in section 4 show that for large data manipulation, Open MPI performs much better than the other parallel models. On the other hand, results shown in Fig. 6 and Fig. 7 for quick sort show that in most of the cases, sequential execution (C++) produces good results as compared to parallel architectures. Although Open MPI is much more faster than sequential and consumes less energy as compared to sequential execution. But the other two approaches, Open MP and CUDA takes much longer than sequential and in result consumes more energy.

VII. Conclusion and Future Work

Results obtained by running test codes using four models C++ (sequential), Open MPI, Open MP, and CUDA have been discussed in the previous section. The results show that for small calculations, all the models produce the same results in terms of time and energy consumption. Even in some cases as in sorting, the parallel programming models need more resources and time to perform the task. Also, the results obtained by sequential execution are same for small matrix and array sizes. Parallel computation increases performance when running large and complex computations where it is possible to parallelize the code blocks. Though, every language provides different mechanisms to increase efficiency the default mechanism provided by those models was used. As was mentioned earlier in this paper, the computational tasks of matrix multiplication and sorting were performed on a certain machine. Although the results may differ when performing a different task and utilizing different machine the simple technique used in this work provide a quick and simple way to get a general idea about the performance and energy consumption of a particular programming model on similar machines for different tasks.

In future, this work will be extended by executing some other codes and using different machines or running real applications to get a better estimate of the performance and energy consumption.

REFERENCES

[1] Benedict, Shajulin, et al. "Automatic performance analysis of large scale simulations." Euro-Par 2009–Parallel Processing Workshops. Springer Berlin Heidelberg, 2010.

[2] Wang, Zhiming, et al. "Energy-aware and revenue-enhancing Combinatorial Scheduling in Virtualized of Cloud Datacenter." JCIT 7.1 (2012): 62-70.

[3] Benedict, Shajulin. "Energy-aware performance analysis methodologies for HPC architectures—An exploratory study." Journal of Network and Computer Applications 35.6 (2012): 1709-1719.

[4] Dagum, Leonardo, and Rameshm Enon. "OpenMP: an industry standard API for shared-memory programming." Computational Science & Engineering, IEEE 5.1 (1998): 46-55.

[5] Gabriel, Edgar, et al. "Open MPI: Goals, concept, and design of a next generation MPI implementation." Recent Advances in Parallel Virtual Machine and Message Passing Interface. Springer Berlin Heidelberg, 2004. 97-104.

[6] Kirk, David. "NVIDIA CUDA software and GPU parallel computing architecture." ISMM. Vol. 7. 2007.

[7] Rejitha, R. S., C. Bency Bright, and Shajulin Benedict. "Energy consumption analysis and energy optimization techniques of HPC applications." Energy Efficient Technologies for Sustainability (ICEETS), 2013 International Conference on. IEEE, 2013.

[8] Chowdhuri, Arghyadip, and M. Rajashekhara Babu. "Analysis of Loop Optimization Techniques in Multi-Core Environment using MPI-C." Analysis 2.4 (2011).

[9] Freeh, Vincent W., et al. "Analyzing the energy-time trade-off in high-performance computing applications." Parallel and Distributed Systems, IEEE Transactions on 18.6 (2007): 835-848.

[10] Feng, Xizhou, Rong Ge, and Kirk W. Cameron. "Power and energy profiling of scientific applications on distributed systems." Parallel and Distributed Processing Symposium, 2005. Proceedings. 19th IEEE International. IEEE, 2005.

[11] https://software.intel.com/en-us/articles/intel-power-gadget-20 last accessed on 05-12-2015.

[12] Zotos, Kostas, et al. "Energy complexity of software in embedded systems." arXiv preprint nlin/0505007 (2005).

[13] Castillo, Juan, et al. "Energy Consumption Estimation Technique in Embedded Processors with Stable Power Consumption based on Source-Code Operator Energy Figures." XXII Conference on Design of Circuits and Integrated Systems. 2007.

[14] Enos, Jeremy, et al. "Quantifying the impact of GPUs on performance and energy efficiency in HPC clusters." Green Computing Conference, 2010 International. IEEE, 2010.

[15] Rashid, Mohammad, Luca Ardito, and Marco Torchiano. "Energy Consumption Analysis of Algorithms Implementations." Empirical Software Engineering and Measurement (ESEM), 2015 ACM/IEEE International Symposium on. IEEE, 2015.

[16] Roy, Swapnoneel, Atri Rudra, and Akshat Verma. "An energy complexity model for algorithms." Proceedings of the 4th conference on Innovations in Theoretical Computer Science. ACM, 2013.

Image Transmission Model with Quality of Service and Energy Economy in Wireless Multimedia Sensor Network

Benlabbes Haouari
Laboratory of Energetic in Arid Zone
Tahri Mohammed University-Bechar
Street independence bp 417
Bechar, Algeria

Benahmed Khelifa
Department of Science, Faculty of
Exact Sciences
Tahri Mohammed University-Bechar
Street independence bp 417
Bechar Algeria

Beladgham Mohammed
Department of Electrical Engineering
Faculty of Technology, Tahri
Mohammed University-Bechar
Street independence bp 417
Bechar Algeria

***Abstract*—The objective of this article is to present the efficiency of image compression in the Wireless Multimedia Sensor Network (WMSN), the method used in this work is based on the lifting scheme coupled with the SPIHT coding to wavelet biorthogonal CDF 9/7. The effectiveness of this technique results in combining two advantages. First it allows saving energy to prolong the life of the network; second, it improves the Quality of Service (QoS) in terms of average throughput, the number of dropped packets, and End-to-End average delay. The authors examined two scenarios of the same network model in NS2 simulator, according to the above critical points of the energy economy and quality of service, the first scenario transmits an original image, and the second sends the compressed image with the used method. The simulation results presented show that the proposed system allows to extend the life of the network and minimizes the consumption of energy; and it can transmit the image in comfortable conditions for QoS of network, reduce the End-to-End average delay, no dropped packet, best average throughput and satisfied for bandwidth.**

Keywords*—*Wireless Multimedia Sensor Network; Multimedia; Compression; Routing; Energy Consumption; QoS; Energy Economy

I. Introduction

The emergence of Multimedia Sensor Networks Wireless and embedded computing systems opens the way for the deployment of new applications for surveillance, monitoring, and control of large systems, including those that extend over vast areas that require geographical and instrumentation scale [1].

WMSN are useful for many applications such as surveillance and monitoring, military, storage of potentially relevant activities, sport, medical and other fields. These applications bring new scientific and technological challenges that caught the attention of a large number of researchers in recent years.

The character and the specification of the multimedia data such as image cause problems in the transmission of data over the network and in the node itself. Among these constraints, the energy consumption and the quality of service represented average throughput, the number of dropped packets, and End-to-End average delay, knowing that the electronic transmission module take the largest share of energy consumption [2].

Image compression is the optimal solution to solve these problems simultaneously [3-6] while it saving a lot of energy, by sending a small amount of information (compressed image) in the circumstances of the comfortable Qos of the network in terms of flow of information, and no dropped packets in optimal time. The previous scenario is better than the second, which involves sending a very large amount of information (original image), which consumes more energy in unsatisfactory network conditions due to large information flows occupying bandwidth and dropping a high number of packets in a long time.

The paper comprises five sections followed by a conclusion. In the second section titled “energy consumption in wireless multimedia sensor network”, show the importance of image compression, and its importance in energy consumption in WMSN, as review some related work in this area. In the section entitled "QoS evaluation WMSN", show the importance of QoS in WMSNs, e.g. to help make decisions promptly, and explain some of the QoS criteria in WMSNs. In the next section titled "Radio Model (energy consumption)", mathematically illustrate power consumption in the radio equipment for the transmitter and receiver, and how to dissipate energy. The other sections include positioning of our work and our contributions in detail.

II. Energy Consumption in Wireless Multimedia Sensor Network

The energy consumption is very important criterion of performance in WMSNs. The large volume of media (eg image) requires bandwidth that consumes more energy. [5] The economy of energy for the transmission is necessary for nodes which allow extending the life of the network.

Data compression or source coding can be lossy or lossless. In WMSN, the lossy compression is often preferred for low-speed transmissions; it loses details in the image but within acceptable limits. Using this type of compression to data transmission in WMSN [10-12] allow energy saving visible, which leads to prolong the lifetime of the network, in contrast, the transmission the natural image that consumes seven times the energy, energy consumption are the additive value [23]. As prove it in this article.

Several research studies related to image compression in wireless multimedia sensor networks in the literatures [5] and [24-26].

In [5], ZainEldin H et al. discussed various compression techniques to WMSN a comparison between them and the factors that affect compression performance. Image quality, compression ratio, speed compression and energy consumption are the most important indicators discussed for compression performance.

In [24], Kumar v et al analyzes of all the obtained experimental results demonstrates that the incorporation of SVD and BTC in image compression along with OCT in an adaptive manner enhances the compression performance significantly. They proposed technique performs the best technique in terms of PSNR and MSE. But it requires slightly longer time that makes it suitable for large bandwidth. This compression technique depends on the parameter (x) that based on the observation of the standard deviation (σ) to decide what compression technique can be used as following: if (σ < x) use DCT, else if (σ > x) use SVD, else if ($35 \leq \sigma \leq 45$) use BTC

In [25], Ghorbel et al compare two image compression methods, Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) and test the ability of each method on wireless sensor networks (WSN) to. And they implement these methods on a real platform of sensor networks with TelosB sensor type. And they are running the performance evaluation and compare the two methods in terms of image quality, execution and transmission time, and lost packet and memory usage and energy consumption

Ma et al. [26] present the multimedia compression techniques and multimedia transmission techniques and provided an analysis of energy efficiency when applied to limited resources platform. For the image compression they discussed three important technical JPEG2000, JPEG (DCT) (EBCOT) and SPIHT. They analyzed in their working efficiency in compression terms, the memory requirement and computational complexity. They concluded that SPIHT is the best choice for compression methods lower power consumption due to its ability to provide higher compression ratio with low complexity. JPEG2000 (EBCOT) hit a high compression ratio, which means better quality than SPHIT, but the burden of higher computation and power consumption for resource-limited systems, due to the complexity of operations EBCOT Tier-1 and Tier-2 caused intensive complex coding.

III. THE QOS EVALUATION IN WMSN

Because of the critical and urgent nature of certain information from multimedia sensors such as in football, detection systems errors arbitration as technology on the goal line (goal-line technology or goal decision system) is an electronic device and / or video assistance to arbitration, used to determine for certain whether a ball has crossed fully or not the goal line in a very short time, it is necessary to provide QoS support for these applications.

These applications require that the data received must reach the base station during a limited time to the data is useful and acceptable [13]. WMSN with QoS reflects the delay requirements, bandwidth and residual energy. It also optimizes energy management.

In general, a criterion is a system or standard of measurement defined as a parameter for quantitatively assessing a method, event, and entity by using special procedures to perform measurements [14].

Among the QoS evaluation objective criteria in WSN:

Average energy consumption: [15] the average energy consumed by all nodes in the network, and lifetime: [16, 17] Time until the first sensor node or group of sensor nodes in the network runs out of energy.

End-to-end delay: the average time taken by a data packet to arrive in the destination. It also includes the delay caused by route discovery process and the queue in data packet transmission. Only the data packets that successfully delivered to destinations that counted. The delay of end to end and is the additive value [15, 23].

End-to-end Average delay = ∑ (arrive time – send time) / ∑ Number of connections.

Throughput: [15] the average time from the first bit until the last bit of data packets has left the transmitting node to destination a certain node, excluding protocol overhead, and excluding retransmitted data packets, which generally is lower than network access connection speed (the channel capacity or bandwidth).

Throughput = ∑ number of packet send / ∑ send time.

Packet Delivery Ratio: [15, 18] is measured as a percentage of receive packets with respect to Generated Packets (packets sent), Packet Delivery Ratio is the multiplicative value [23]. The main parameters that monitored are usually bandwidth, latency, dropped packet and request response time.

Packet Delivery Ratio = ∑ number of packet receive / ∑ Number of packet send

Dropped Packets: the total number of dropped packets during the transmission

Dropped Packet = number of packet send – number of packet received.

IV. RADIO MODEL (ENERGY CONSUMPTION)

Model of the energy consumption of the radio equipment for transmitter and receiver dissipate energy, to operate the electronic radio, as shown in Figure 1. The formula for the transmission of energy consumption and reception data of K-bit from two sensors at a distance *d* is as follows [20, 21]:

$$E_{Tx}(k,d) = E_{Tx-elec}(k) + E_{Tx-amp}(k,d) \qquad d > 1$$

$$E_{Tx}(k,d) = E_{elec} * (k) + \in_{amp} * k * d^{2} \tag{1}$$

And receive data, radio spending:

$$E_{Rx}(k) = E_{Rx-elec}(k)$$

$$E_{Rx}(k) = E_{elec} * k \quad (2)$$

Total consumed energy of each node

$$= \Sigma E_{Rx} + \Sigma E_{Tx} \quad (3)$$

= Total consumed energy of data receiving + total consumed energy of data transmitting [22].

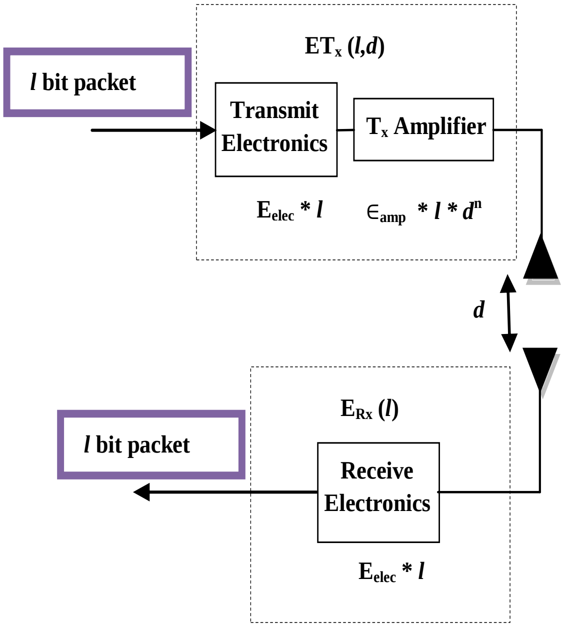


Fig. 1. Radio Model (energy consumption)

V. Network Model and Simulation

The proposed work is implemented using NS2 version 2.35 and Matlab R2014a. The simulation is carried out on personal computer with Intel(R), Pentium (R) CPU P6200, processors rated at 2.13 GHz, main memory of 4 GB and 64 bit Microsoft Windows 7 operating system and in Ubuntu 14.04.

A. *The compression method*

The Biorthogonal CDF 9/7 Wavelet Based on Lifting Scheme and SPIHT Coding is a Discrete Wavelet Transform (DWT). The wavelet transform that uses functions located both in real space [7,8]. This method is detailed in [9] Beladgam.M et al have used a grayscale image to prove the effectiveness of this method in terms of image evaluation criteria objectives such as the compression ratio , PSNR and MSSIM, where they are compared against other methods. For these reasons, take decision to use this method to transmit a color image in WMSN. In the MATLAB simulation program, the color image compression flowchart 512x512 Lena to be transmitted is described in Figure 2. In a first step, each pixel of the original image color is multiplied by the bitrate 0.75 pixels per second (bpp). Then, convert the color space of the image from RGB to YCbCr. Third, apply Wavelet Decomposition (CDF9 / 7 + lifting) Each layer is Independently and apply SPIHT Encoding Each layer is Independently For Each matrix Y, Cb and Cr. Eventually image size is obtained compressed to be transmitted (Bit Streams) to generate packets that are saved in a CSV file; the algorithm to generate these packets is as follows:

Algorithm 1

```
image_compressed_size = bit Streams/8      /* image compressed size by bytes */
Initialize the packet size (packet_seize)      /* in bytes */
Packet_number1 = image_compressed_size/packet_seize
Packet_number2= rounding the (Packet_number1)
if  (Packet_number1 > Packet_number2)
{
   packet_number = Packet_number2 +1
else
    packet_number = Packet_number1
}
/* creating a matrix contains the number of packets and the size of each packet * /
s=0
for (i=1 ;i<= packet_number; i++)
 {
if  (packet_seize < (image_compressed_size - s))
{
T[i]= packet_seize;
s=s+ packet_seize;
 else
  T[i]= image_compressed_size -s;
}
}
    /* Then, Write the content of the matrix T [packet_number] that contains the size of each packet in a CSV type file, then uses this file in the image transmission model in NS2 simulator  * /
```

B. *Wireless Multimedia Sensor Network*

In this experiment, using the NS-2 simulator, is particularly well suited to packet switched networks [19], using the reference image Lena color. Two scenarios are realized, the first is to transmit the original image and calculate the energy consumed for each node of the network and overall energy, and also calculates the average of the network and the rate of dropped packets, and calculated the same for the second scenario, which will send a compressed image.

In this article, adopt a WMSN formed by fifteen nodes multimedia randomly denoted by N= {n1, n2, ..., n15}, deployed sensors in area of 600*600 M^2, of the different distances *d* between tow nodes and only one sink node. All sensing nodes are used for data collection in the surveillance zone and do not move after deployment, in these 15 nodes three nodes multimedia (n2, n9 and n13) transmit data at the same time to the sink node, as shown in Figure 3. The main distinguishing features of the system are as the followings:

- The sink node has highest ability of communication and computation, and has not the energy problem.
- All the multimedia sensor nodes have the same initial energy and the capacity of communication and computing.

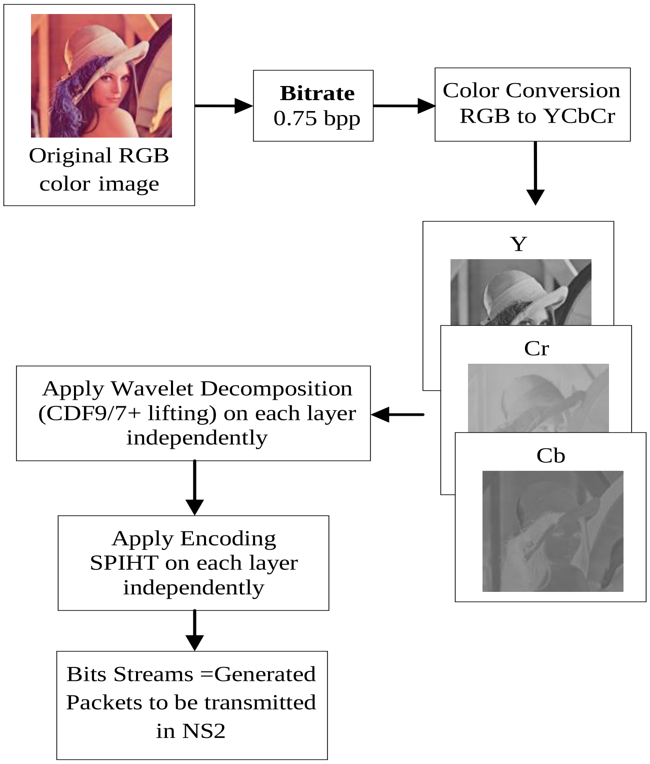

Original RGB color image
Bitrate 0.75 bpp
Color Conversion RGB to YCbCr
Y
Cr
Cb
Apply Wavelet Decomposition (CDF9/7+ lifting) on each layer independently
Apply Encoding SPIHT on each layer independently
Bits Streams =Generated Packets to be transmitted in NS2

Fig. 2. The image compression flowchart

Only focus on overall energy consumption by the entire network as a single unit, the energy consumed by each sensor, Packet Delivery Ratio, Average End-to-End Delay and Average Throughput, while communications between node and sink devices outside the network are outside the scope of this article.

The simulation parameters are shown in Table I.

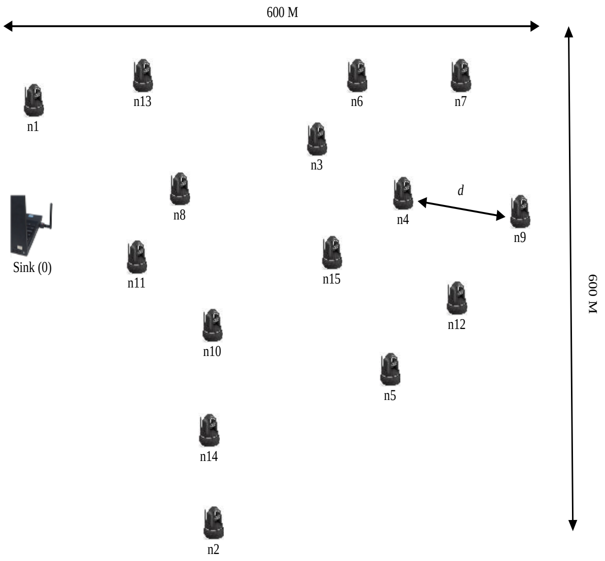


Fig. 3. Network model

TABLE I. NS2 SIMULATION PARAMETERS

Parameter	**Value**
Network	Static
Simulation area	600 m* 600 m
Network sizes (node number)	15
Routing protocols	AODV
Traffic source	Constant Bit Rate (CBR)
Initial energy (J)	100
E_{Tx}	0.6 nJ/bit
E_{Rx}	0.3 nJ/bit
Antenna model	Antenna/omniantenna
Interface queue type	Queue/Drop tail/priqueue
Link layer type	LL
Min packet in ifq	50

VI. RESULTS AND DISCUSSION

The results of the simulation are shown in Table II, which shows:

All evaluation criteria: energy consumption average, the number of packets to be transmitted, Packet Delivery Ratio, Average End-to-End Delay and Average Throughput show that the results of scenario 2 (Compressed Image Transmission) are much better than the results of the first scenario.

Network lifetime: the simulation time of the second scenario is about 14.55% of the first scenario, which indicates that the length of life of network transmitting a compressed image with the method studied can extend this period almost seven times that the transmission of original image.

Figure 4 shows the simulation graph for the rate of energy consumption for each sensor node in both scenarios, and as see that the energy consumption in the second scenario is approximately equal to 14, 52% of Energy consumed in the first scenario.

TABLE II. SIMULATION RESULTS

Criteria	**Scenario 1**	**Scenario 2**
Simulation time (s)	31.108277	4.526776125
Average Energy Consumption (J)	9,561964533	1,388441
Generated Packets	2364	225
Received Packets	1702	225
Packet Delivery Ratio (%)	71.9966	100
Number of dropped data (packets)	660	00
Number of dropped data (bytes)	673200	00
Average End-to-End Delay (ms)	3353.02	970.678
Average Throughput [kbps]	436.74	390.83

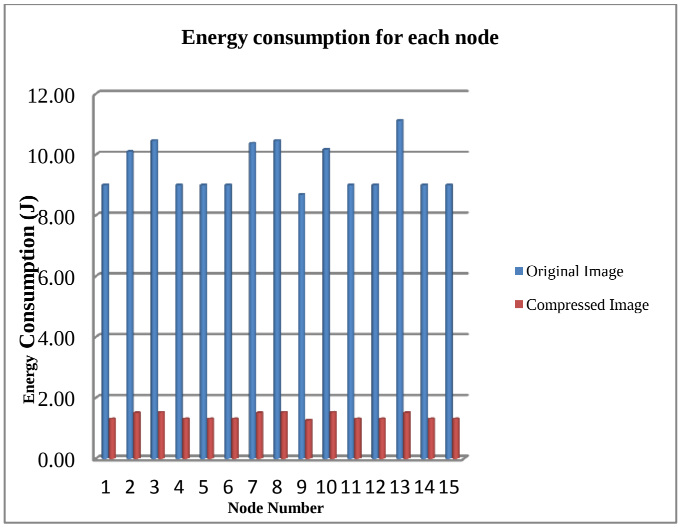


Fig. 4. Energy consumption for each node

VII. Conclusion

In this article, the authors use an algorithm of Biorthogonal CDF 9/7 Wavelet Based on Lifting Scheme and SPIHT Coding in Wireless Multimedia Sensor Network.

This proposed scheme is shown to provide energy savings of about 14, 52% (that is to say seven times that the transmission of original image) at each network node, and can improve the overall energy economy, and also offer the best terms of Quality of Service, during the data transmission with a throughput without affecting bandwidth, without dropped packets (Packet Delivery Ratio = 0), and with an acceptable delay (Average End-to-End Delay).

From the foregoing, which confirms the effectiveness of the use of this method in all types of networks, whether large or small sizes, especially in networks that can monitor environments with unfavorable terrain such as volcanoes, mountains or forests.

Acknowledgment

The authors would like to be grateful to the Editor and anonymous reviewers for their invaluable comments and suggestions, which have improved the quality of the paper.

References

[1] C. Duran-Faundez, "Transmission d'images sur les réseaux de capteurs sans fil sous la contrainte de l'´energie," Docteur Doctorat, Automatique, Université Henri Poincaré, Nancy 1, 23 juin 2009.

[2] I. F. Akyildiz and M. C. Vuran, "Factors Influencing Sensor Network Design," in *Wireless Sensor Networks* wiley, Ed., wiley ed, 2010.

[3] M. Nasri, A. Helali, H. Sghaier, and H. Maaref, "Adaptive image compression technique for wireless sensor networks," *Computers & Electrical Engineering,* vol. 37, pp. 798-810, 2011.

[4] T. Ma, M. Hempel, D. Peng, and H. Sharif, "A Survey of Energy-Efficient Compression and Communication Techniques for Multimedia in Resource Constrained Systems," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS,* vol. 15, pp. 963-972, 2013.

[5] H. ZainEldin, M. A. Elhosseini, and H. A. Ali, "Image compression algorithms in wireless multimedia sensor networks: A survey," *Ain Shams Engineering Journal,* vol. 6, pp. 481-490, 2015.

[6] S. Aswale and V. R. Ghorpade, "Survey of QoS Routing Protocols in Wireless Multimedia Sensor Networks," *Journal of Computer Networks and Communications,* vol. 2015, pp. 1-29, 2015.

[7] S. G. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 37, pp. 2091-2110, 1989.

[8] B. Huang, S. López, Z. Wu, S. Al-Mansoori, and A. Kunhu, "Hybrid DWT-DCT-Hash function based digital image watermarking for copyright protection and content authentication of DubaiSat-2 images," *Proc. of SPIE,* vol. 9247, p. 924707, 2014.

[9] M. Beladgham, A. Bessaid, A. M. Lakhdar, and A. Taleb-Ahmed, "Improving Quality of Medical Image Compression Using Biorthogonal CDF Wavelet Based on Lifting Scheme and SPIHT Coding," *SERBIAN JOURNAL OF ELECTRICATL ENGINEERING,* vol. 8, p. 17, 2011.

[10] D. Zordan, B. Martinez, I. Vilajosana, and M. Rossi, "On the Performance of Lossy Compression Schemes for Energy Constrained Sensor Networking," *ACM Transactions on Sensor Networks,* vol. 11, pp. 1-34, 2014.

[11] H. Benlabbes, K. Benahmed, and M. Beladgham, "The Image Compression Methods In Wireless Multimedia Sensor Networks," *Progress in Machines and Systems,* vol. 4, p. 12, 2015.

[12] M. Hooshmand, M. Rossi, D. Zordan, and M. Zorzi, "Covariogram-based Compressive Sensing for Environmental Wireless Sensor Networks," *IEEE Sensors Journal,* pp. 1-1, 2015.

[13] A. Alanazi and K. Elleithy, "Real-Time QoS Routing Protocols in Wireless Multimedia Sensor Networks: Study and Analysis," *Sensors (Basel),* vol. 15, pp. 22209-33, 2015.

[14] S. Aswale and V. R. Ghorpade, "Survey of QoS Routing Protocols in Wireless Multimedia Sensor Networks," *Journal of Computer Networks and Communications*, vol. 2015, pp. 1-29, 2015.

[15] S. A. Alghamdi, "Load balancing ad hoc on-demand multipath distance vector (LBAOMDV) routing protocol," *EURASIP Journal on Wireless Communications and Networking,* vol. 2015, 2015.

[16] S. Deshmane and P. Lambhate, "A Survey on Approaches for Extending Network Lifetime Topology," *International Journal of Advance Research in Computer Science and Management Studies,* vol. 2, 2014.

[17] J. H. Kleinschmidt, "Genetic Algorithms for Wireless Sensor Networks," ed, 2009, pp. 755-758.

[18] P. Sarisaray Boluk, K. Irgan, S. Baydere, and E. Harmanci, "IQAR: Image quality aware routing for Wireless Multimedia Sensor Networks," pp. 394-399, 2011.

[19] Y. BADDI. (2011). *Introduction au simulateur réseau NS2*. Available: http://y-baddi.developpez.com/tutoriels/ns2/#LV

[20] R. Patel, S. Pariyani, and V. Ukani, "Energy and Throughput Analysis of Hierarchical Routing Protocol (LEACH) for Wireless Sensor Network," *International Journal of Computer Applications,* vol. 20, 2011.

[21] I. F. Akyildiz and M. C. Vuran, "Factors Influencing WSN Design," ed, 2010, pp. 37-51.

[22] H.-C. Jang, H.-C. Lee, and J.-X. Huang, "Optimal Energy Consumption for Wireless Sensor Networks," in *Proceedings of the 9th Joint Conference on Information Sciences (JCIS)*, 2006.

[23] Z. Zhu, Y. Ding, R. Tang, H. Xu, and Y. Liu, "An improved service-aware multipath algorithm for wireless multimedia sensor networks," in *International Conference on Graphic and Image Processing (ICGIP 2012)*, 2013, p. 87684K.

[24] V. Kumar, A. Kumar, and A. Bhardwaj, "Performance evaluation of image compression techniques," presented at the Devices, Circuits and Systems (ICDCS), 2012 International Conference on, Coimbatore, 2012.

[25] O. Ghorbel, W. Ayedi, M. W. Jmal, and M. Abid, "Images compression in WSN: Performance analysis," presented at the Communication Technology (ICCT), 2012 IEEE 14th International Conference on, Chengdu, 2012.

[26] T. Ma, M. Hempel, D. Peng, and H. Sharif, "A Survey of Energy-Efficient Compression and Communication Techniques for Multimedia in Resource Constrained Systems," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS,* vol. 15, pp. 963-97

Dual Security Testing Model for Web Applications

Singh Garima
Research Scholar
Department of Computer Science and Engineering
JECRC University
Jaipur, Rajasthan, India

Kaushik Manju
Associate professor
Department of computer science and Engineering
JECRC University
Jaipur, Rajasthan, India

***Abstract*—In recent years, web applications have evolved from small websites into large multi-tiered applications. The quality of web applications depends on the richness of contents, well structured navigation and most importantly its security. Web application testing is a new field of research so as to ensure the consistency and quality of web applications. In the last ten years there have been different approaches. Models have been developed for testing web applications but only a few focused on content testing, a few on navigation testing and a very few on security testing of web applications. There is a need to test content, navigation and security of an application in one go. The objective of this paper is to propose Dual Security Testing Model to test the security of web applications using UML modeling technique which includes web socket interface. In this research paper we have described how our security testing model is implemented using activity diagram, activity graph and based on this how test cases are generated.**

Keywords—Web application testing; Security testing; UML modeling; Web socket programming

I. Introduction

In recent times web based applications are frequently used by all. There is no need to install these applications on each system, but they are installed on the web server. A web server is an internet information service on which web application is implemented. With a growing concern about the quality of web applications web application testing is again an area of research to explore. An effective modeling technique is required to know particular challenges of web applications for testing [1]. In the normal daily routine types of web applications the security feature is implemented to verify the e-mail ID, cell phone number, landline number and other government approved identification with social identifications like e-mail and Face book account etc. After the formality of registration along with the collection of the general information about a customer or user, registration ID and password are provided to access the whole activity of website excluding the web server admin authority. At this stage it is not possible to identify whether the site is being accessed by the 100% authentic user whose records have been recorded. This is one of the critical types of task which identifies the authentication of use of ID and password. And in this situation, loosing rights in favor of user are sometimes critically harmful to the company. In order to provide a solution and to overcome this kind of situation there is a need of a self managing web application which handles the control itself ,as per need of the user , essential for security purpose. The proposed model gives a way to generate test cases and helps in web application testing. In this study, Unified Modeling Language (UML) approach has been used. UML based approach for modeling web applications was used earlier by different researchers for testing contents, to navigate the model. This research extends to model based testing to test security of web applications along with content and navigation. Specific Testing Process Model (STPM) gives composite view of content, navigation and security model to test web applications. In this paper researcher has elaborated Specific Testing Process Model and has implemented Dual security testing model of the STPM, which is one of the aspect of STPM.

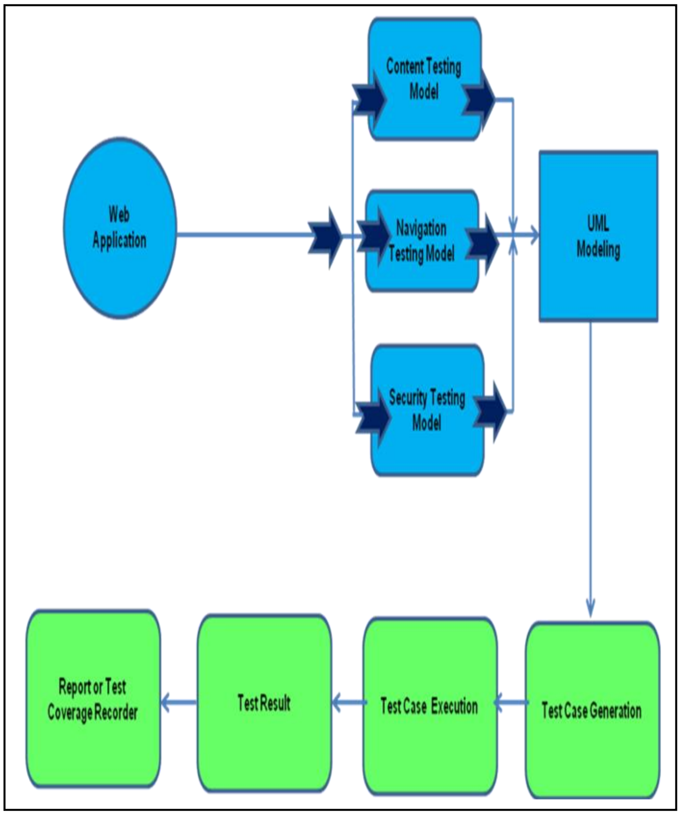


Fig. 1. Specific Testing Process Model

II. Objectives

A. To design a model (Dual Security Tesing Model) to test the security of web application by using UML modeling techniques including web socket interface.

B. To Implement Dual Security Tesing Model of STPM.

III. Literature Review

The motivation of present research is to work on three different perspectives of a web application by using a composite model called Specific Testing Process Model (STPM). This study proposes a secondary stage of testing navigation and security of web application by busing UML modeling techniques. Various models have already been proposed for testing web applications. According to the unique characteristics and challenges of web applications, these models have a different origin and test goals. Different methods using partial rewriting based specification language for both syntactic and semantic checking were developed [2] [3] for the static applications. Researchers focus on the content of web sites by correcting and reforming the syntax and semantic [4] [5]. In this study, UML diagrams have been used for content testing of web application. Based on these diagrams test cases are generated. Test cases can be generated efficiently for content testing by using UML modeling. For testing navigation of the applications, there are UML based models [6] [7] [8], graph based models [9] [10], state charts based models [11] [12]. Also, researcher proposed a novel approach to generate test cases from UML activity diagrams [13]. In this approach, navigation testing covers contextual and non-contextual hyperlinks of web application, Security aspects of the web application should be analyzed and modeled during entire development cycle to identify security requirement in the early stage of the development progress. There can be various security constraints like access control, availability, authentication, integrity, secrecy, etc. which should be taken care of. Numerous researchers have explored the use of the UML language for modeling, security aspects of web applications. Different security models [1] [14] [15] [16] used UML approach to understand security requirements. Web socket protocol can be used to develop web application [17]. Using HTML5 based web browser, web socket based applications can be executed. The creation of a real time application and live content facility can be done using web socket protocol. It gives more interaction between browser and application [18]. This study cover access control, security aspect of web application. It also provides the user with an interface by using web socket programming in order to have suggestions regarding the product purchase while using online shopping web application.

IV. Testing Model

Different modeling methods are available to test the web applications at different levels, i.e., content, navigation and security, but no model has yet been developed which can test all three levels of modeling in integration as discussed in the literature above. Here Fig.1 shows a Specific Testing Process Model to test web applications. Researcher has used three sub-models which are as follows:

A. Content Testing Model

Information displayed on the web application and its presentation plays a vital role as we say first impression is the last impression. If a user finds inaccurate information and an unstructured layout of an application its quality and users will be affected. Researcher extends the content model [7], this model tests the completeness and the correctness of web application information displayed in the form of web pages i.e., its outlay. The content testing model is important because it describes where the objects (text, button, Audio, Image, Form, Video, Frameset, Frame) are placed.

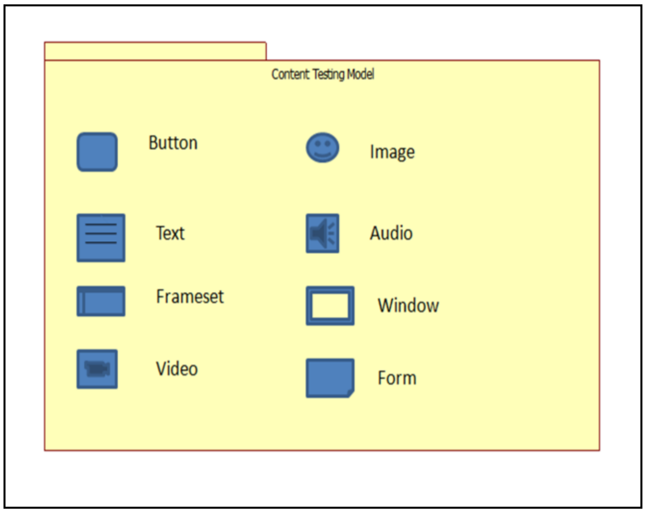


Fig. 2. Content Testing Model

B. Navigation testing model

Navigation of web application gives freedom to a user to move from one page to another within the same or different pages of an application, click on links, images etc. In other words navigation in context of a web application is the sequence of web pages that a user can browse to achieve a desired page or function. Here in this research, navigation testing model is a sub-model of Specific Testing Process Model, which allows a tester to test whether a user is able to reach information and navigate according to content testing model. The basic elements to test in the navigational testing model are contextual and non-contextual hyperlinks described in [19]. As shown in Fig. 3.

Contextual Hyperlinks: Link between objects is called contextual, as it carries information from its source to destination object. The contextual link can be the links provided within the web application having its source or a destination with in the application.

Non-Contextual Hyperlinks: The non-contextual link is the link which does not carry information within the application. The content of the required page does not depend on the content of its source page.

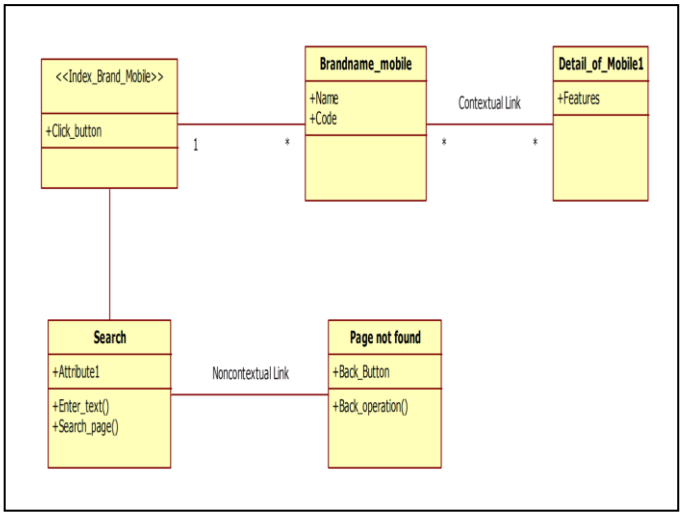


Fig. 3. Navigation model of contextual and non-contextual links

C. Security Testing Model

Model-driven security is an approach proposed by [16] and is used to simplify system design and generate artifacts. Security testing aims at verifying the effectiveness of the overall web applications defenses against undesired access by unauthorized users, its capability to prevent system resources from improper use and granting authorized users access to authorized services and resources. There has been only little work on the UML based security model. The focus of the security testing model is on Role Based Access Control (RBAC) [20]. It is an approach to restrict system access only to authorized user as shown in Fig4., One Time Password (OTP) [21] and Completely Automated Public Turing test to tell Computer and Human Apart (CAPTCHA) [22].

Researcher has elaborated dual security testing model which is one of the aspects of STPM. It provides a feedback on the users of online web application as shown in Fig. 5. Using the feedback facility current user can collect the views of previous users about the product they want to purchase. The feedback providers are already registered users of that application so the user can authenticate the validity of the product and also organizations can develop business intelligence. This process is accountable under the dual verification. Here, a registered user goes for dual verification so that organization is sure about the validity of the user and it can provide feedback details to the registered user. As soon as he/she logs in a One Time Password (OTP) SMS is sent to the registered mobile number or user registered mail account. Then the user fills an interface provided to him/her with the information about registration id ,OTP and CAPTCHA which are then verified in the database. Otherwise the user is asked to get registered first. After verification, if the user is found registered he/she can take the feedback information regarding the product and can also avail help of Customer Care Service (CCS). In CCS the browser control is given to the admin so that while using an interface he can help the user in buying the product by providing suggestions like product comparison.

In dual security testing model, the user's browser is controlled by admin of the web application when the user wants Customer Care Service (CCS) by using web socket programming. Using this dual security testing model, at the time of purchasing any product if user needs suggestions regarding the product price range, comparison with other similar product, then as he click on Customer Care Service (CCS), at this time the browser gets controlled by admin remotely using web socket programming interface, admin also sees the user browser activity.

As the user selects the product a suggestion message about the price range of the product with other similar product on the same online shopping application is displayed. By this model the customer is able to compare the price of the product on the visited online shopping site itself, he or she need not move on to another web site for a comparison of different range of the product. The dual security testing model helps in knowing the pattern of purchase for product line which includes ranges of products customer selects most, which is useful in data collection for companies.

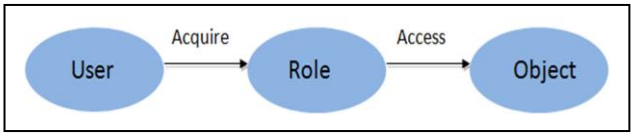


Fig. 4. Role Based Access Control

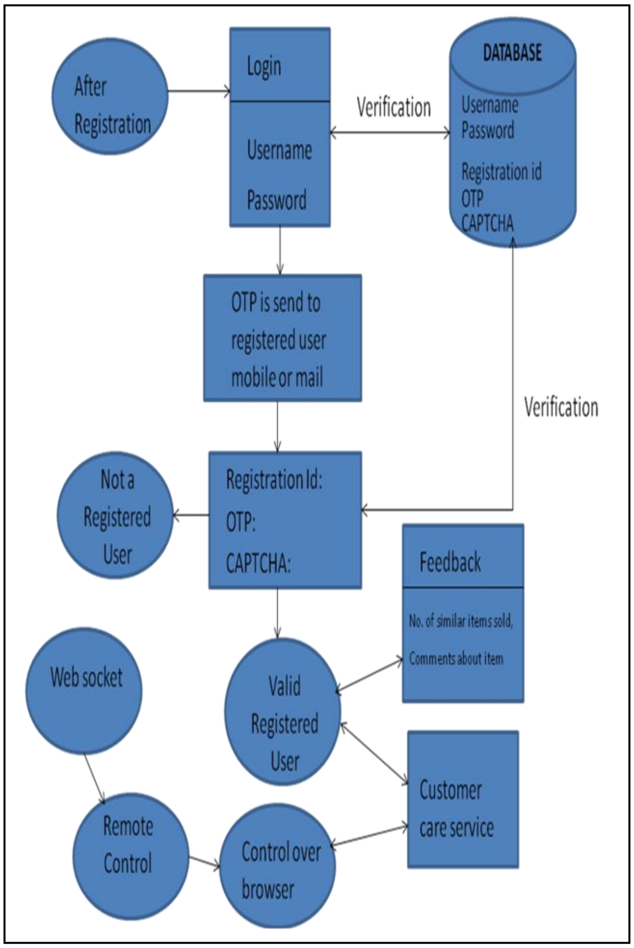


Fig. 5. Dual Security Testing Model

V. Proposed Implementation Methodology

In this section, researcher discusses proposes an approach towards implemention of the model and generation of a test case from an activity diagram. The proposed approach passes through three basic steps. These three basic steps are as follows:

A. Activity Diagram

Designing activity diagram with the required information. The activity diagram describes the guideline about the modeling and necessary information required. Here in the proposed methodology the activity diagram is designed to implement the modeling, which is shown in the above mentioned Fig. 5, about “Dual Security Testing Model”. The activity diagram changes the state of an object from previous object and the output at various stages. Activity diagram creates object during the execution according to decision and conditions. Finally, activity diagram Fig. 6 provides the highest level of the abstractions.

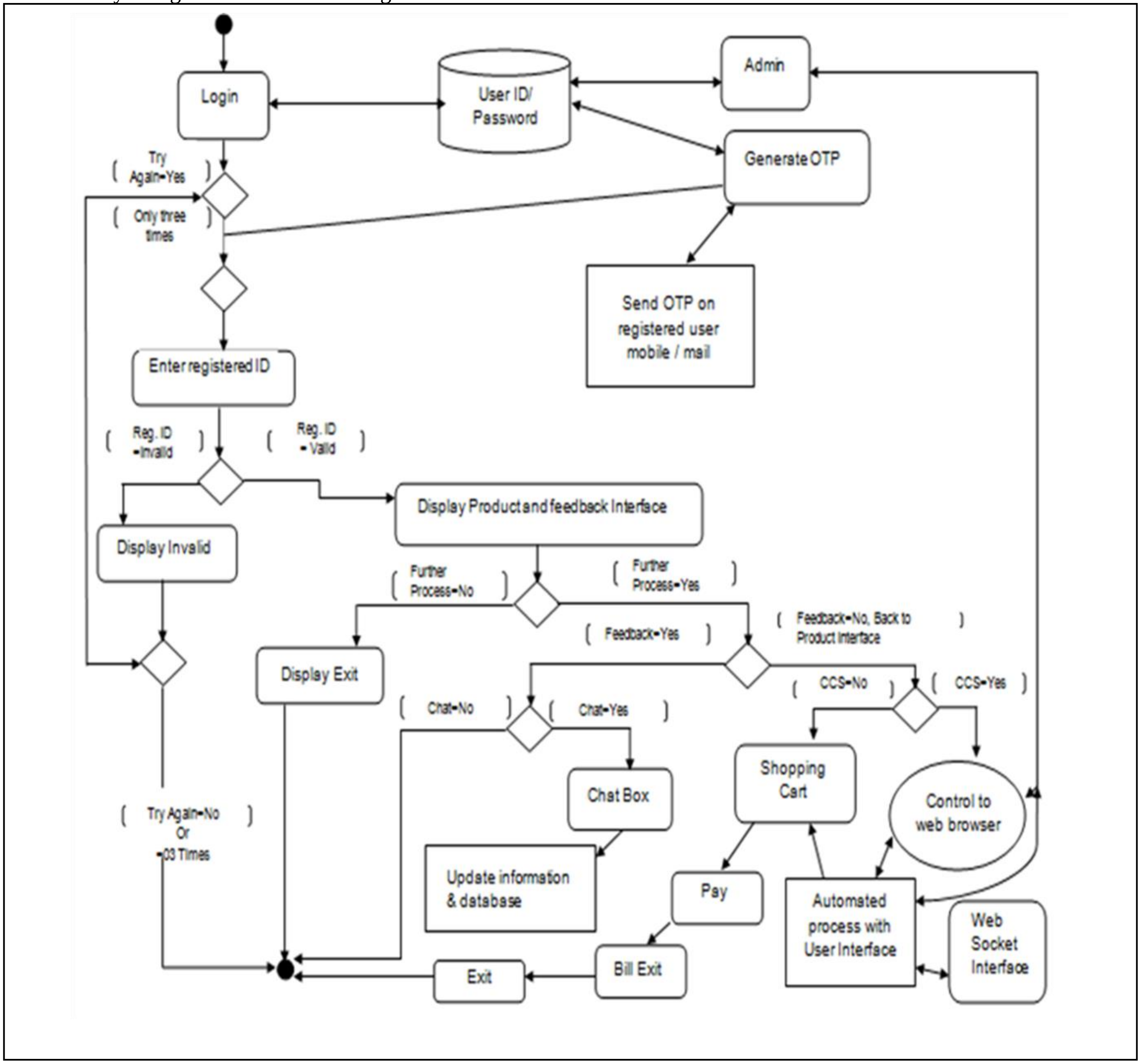


Fig. 6. Activity diagram of online shopping cart

B. Activity Graph

Converting the activity diagram into an activity graph.

The activity graph is a converted form of the activity diagram used for recognizing the activity and flow procedure of modeling. An activity graph is a directed graph where each node in the graph represents a construct for eg. initial node, flow final node, decision node, guard condition, join node, merge node etc., and each edge of activity graph represents the flow in the activity diagram. The activity graph encapsulates constructs of an activity diagram in a systematic and suitable manner. In the present study a set of rules for mapping constructs for an activity diagram is proposed. Nodes of an activity graph are as follows:

S: Start Node

E: End Node

A: Activity

D: Decision

C: Condition

T: Transaction

Fig. 7 represents the activity graph for the activity diagram in Fig. 6. To form edges, consider one-to-one mapping from an edge of the activity diagram into an edge between two nodes in the activity graph. The graph is labeled for better understanding the flow and activity. Each labeled node is acting as the storage of information as the data structure, which is known as the Node Description Table (NDT). To understand activity graph Fig. 7, refer Table I. Node Description Table.

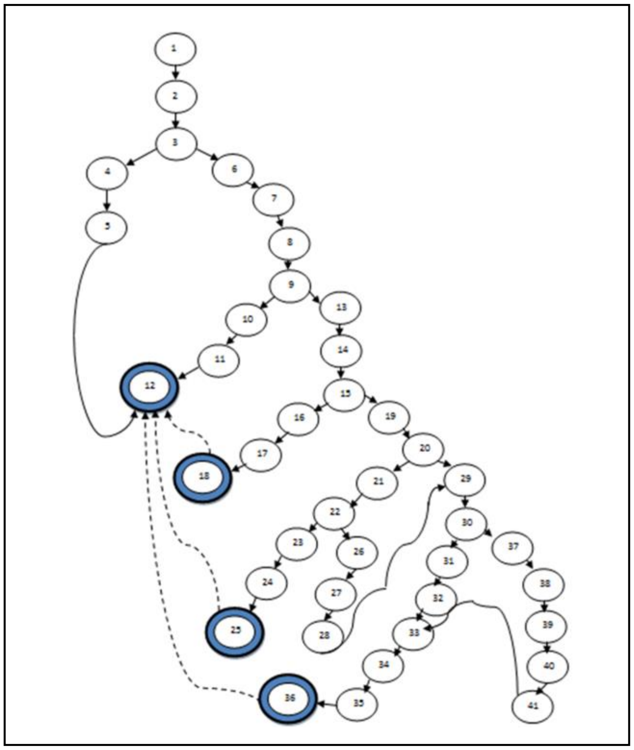


Fig. 7. Activity Graph

C. Generating Test Cases

Generating test cases from the activity graph.

The test case generation is basically an approach to cover all the coverage or the criterion of coverage of all the activities. The present approach of generating test cases from an activity graph is following the given test coverage criterion.

TABLE I. NODE DESCRIPTION TABLE

Node Index	Type of Node	Description of activity
1	S	Start
2	A	Enter Login ID / Password
3	D	Decision
4	C	LoginID=Not Valid (Try Again)
5	A	Display Not Valid
6	C	LoginID=Valid
7	T	Generate OTP
8	A	Enter Registration ID
9	D	Decision
10	C	Registration ID= Not Valid
11	A	Display Not Valid
12	E	End Node
13	C	Reg ID = Valid
14	A	Display Product & Feedback Interface
15	D	Decision
16	C	Further Process = NO
17	A	Display Exit
18	E	End Node
19	C	Further Process = YES
20	D	Decision
21	C	USER Feedback = YES
22	D	Decision
23	C	ChatBox = NO
24	A	Display Exit
25	E	End Node
26	C	ChatBox = YES
27	A	ChatBox Entry
28	T	Update Information
29	C	USER Feedback = NO
30	D	Decision
31	C	CCS = NO
32	A	Display Shopping Cart Interface
33	T	Payment
34	T	Billing
35	A	Display EXIT
36	E	End Node
37	C	CCS = YES
38	A	Display Control Transfer to Web Browser
39	T	Automated Controlled User Interacted Shopping Process
40	T	Control Activity with Socket Programming (Security+Admin)
41/32	A	Display Extended Shopping Cart
42/33	T	Payment
43/34	T	Billing

a) Basic Path Coverage Criterion: At this, firstly define the basic path in activity graph. A basic path is sequence of activities where an activity in that path occurs exaclty once.

b) Simple Path Coverage Criterion: A Simple path is considered for activity diagrams that contain concurrent activities. It is representative path from a set of basic path whwre each basic path has the same set of activities, and activities of each basic path satisfy an identical set of partial order relations among them.

c) Activity Path Covergae: The aim of this covergare os to cover both loop testing and concurrency among the activities of activity diagrams.

TABLE II. MACRO TEST CASES FROM ACTIVITY GRAPH

Test Case No	Sequence of Branch Condition	Activity Sequence	Object State	Expected Result	Actual Result
1	Login ID = Invalid TryAgain = Yes	Enter Login / Password, Display Invalid	False	True	True
2	Login ID = Valid TryAgain = No	Enter Login / Password, Generate OTP	True	True	True
3	RegID = Invalid	Enter RegID, Display Invalid	Invalid	True	True
4	RegID = Valid	Enter RegID, Display Interface	Valid	True	True
5	RegID = Valid Further Processing = No	Further Processing, Display Exit	Valid	True	True
6	RegID = Valid Further Processing = Yes	Further Processing, Display User Feedback	Valid	True	True
7	RegID = Valid Further Processing = Yes	Further Processing, Display User Feedback	Invalid	False	True
8	RegID = Valid Further Processing = Yes ChatBox = No	Further Processing, Display User Feedback Display Exit	Invalid	False	True
9	RegID = Valid Further Processing = Yes ChatBox = Yes	Further Processing, Display User Feedback	Valid	True	True
10	Further Processing = Yes ChatBox = Yes	Further Processing, Display User Feedback, ChatBox Entry	Valid	True	True
11	ChatBox = Yes	ChatBox Entry, Update Information	Valid	True	True
12	ChatBox = Yes	ChatBox Entry, Update Information,	Valid	True	True
13	Further Processing = Yes	User Feedback = No, CSS = No	Invalid	True	True
14	Further Processing = Yes CSS = NO	CSS = No, Display Shopping Cart Interface	Invalid	True	True
15	Further Processing = Yes CSS = NO	CSS = No, Display Shopping Cart Interface Payment	Valid	True	True
16	Further Processing = Yes CSS = NO	CSS = No, Display Shopping Cart Interface Payment +Billing + Display Exit	Valid	True	True
17	Further Processing = Yes CSS = Yes	CSS = Yes ,	Valid	True	True
18	Further Processing = Yes CSS = Yes	CSS = Yes , Display Control Transfer to WB	Valid	True	True
19	Further Processing = Yes CSS = Yes	CSS = Yes , Display Control Transfer to WB, Automated Control User Interaction Shopping	Valid	True	True
20	Further Processing = Yes CSS = Yes	CSS = Yes , Socket Security + Admin Controlled	Valid	True	True
21	Further Processing = Yes CSS = Yes	CSS = Yes , Socket Security + Admin Controlled, Display Extended Shopping Cart	Valid	True	True
21	Further Processing = Yes CSS = Yes	CSS = Yes , Socket Security + Admin Controlled, Display Extended Shopping Cart	Invalid	True	False
22	Further Processing = Yes CSS = Yes	CSS = Yes , Socket Security + Admin Controlled, Display Extended Shopping Cart,	Valid	True	True

These 22 test cases in Table II. Are the cases represent a broad canvas of activity based test case generation. Further, all the test cases can extend at micro level to generate the test case for content and navigation level.

VI. RESULT

a) Dual Security Testing Model has been implemented using an online shopping application. As shown in Fig 6. activity diagram is developed to describe user registration process through modeling.

b) An activity graph is constructed from activity diagram to understand the flow procedure of the application.

iii. Each node of the activity graph has been described in Table I.

c) The model has been implemented through different test cases generated from activity graph (Table II.)

d) The testing approach provide an easy way to find errors in web applications.

VII. Conclusion

Researcher has developed a model (Dual Security Tesing Model) to test the security of web application by using UML modeling techniques including web socket interface. The model has been implemented through test cases. The test cases are generated and system conformance can be checked with the system model. It is suitable with automated admin controlled customer care service system which is beneficial for the cutomers using web applications. The model helps in data collection for the organization which sales their product online. This helps in knowing the buyers pattern, so that the product range can be enhanced. The Future task would be validation of the model through automation and web engineering applications.

Acknowledgment

This research is supported by the JECRC University, Jaipur, Rajasthan, India.

References

[1] Alafi MH., Cordy JR., Dean TR. (2012), 'Recovering role-based access control security models from dynamic web applications', in Brambilla M., Tokuda T., Tolksdorf R. (eds.) , *Web Enginerring,12th International conference, ICWE 2012,* Berlin, Germany, July 23-27, 2012. Springer, pp. 121-136.

[2] Alpuente M., Ballis D., Falaschi M. (2005), 'A Rewriting-based Framework for Web Sites Verification', S. Abdennadher S., Ringeissen C. (eds.), *Proceedings of the 5th International Workshop on Rule-Based Programming (RULE 2004), Rule-Based Programming 2004,* June 1, 2004. Aachen, Germany, Elsevier, pp.41–61.

[3] Alpuente M., Ballis D., Falaschi M., Romero D. (2006), 'A Semi-Automatic Methodology for Repairing Faulty Web Sites', in Hung DV. And Pandya P. (eds.) *Software Engineering and Formal Methods, Fourth IEEE International Conference, SEM 2006,* Pune, India, September 11-15, 2006. IEEE, pp. 31–40.

[4] Coelho J, Florido M. (2006), 'VeriFLog: A Constraint Logic Programming Approach to Verification of Website Content', in Shen HT., Li J., Li M., Ni J. and Wang W. (eds.) *Advanced Web and Network Technologies, and Applications, International Workshops:XPA, IWSN, MEGA,and ICSE, APweb 2006*, Harbin , China, Jan 16-18, 2006. Springer Berlin Heidelberg, pp. 148–156.

[5] Coelho J. and Florido M. (2007), 'Type-Based Static and Dynamic Website Verification',in Werner B. (ed.) *Internet and Web Applications and Services, Second International Conference ICIW 2007,* Morne, Mauritius, May 13-19, 2007, IEEE, pp.32.

[6] Bellettini C., Marchetto A., Trentini A. (2004), 'WebUML: reverse engineering of web applications', in Liebrock LM. (ed.), *Applied Computing, the ACM Symposium SAC 2004,* Nicosia, Cyprus, March 1-17 2004. ACM, pp.1662–1669.

[7] Knapp A., Zhang G.(2006), 'Model Transformations for Integrating and Validating Web Application Models', in Heinrich C. and Ruth B. (eds.) *Modellierung, International workshop, MOD 2006*, Innsbruck, Austria; Springer, pp.115–128.

[8] Nora K., Baumeister H., Hennicker R. and Mandel L. (2000), 'Extending UML to model navigation and presentation in web applications', in Winters G. and Winters J. (eds.) *Modelling Web Applications in the UML Workshop, UML 2000*., England, UK, October 2-3, 2000, York, pp. 1.

[9] Sabharwal S., Bansal P., Aggarwal M. (2013), 'Modelling the navigation behavior of dynamic web application', *International journal of computer applications-Scholarly peer reviewed research publishing journal*, **65(13)**, pp. 20-27.

[10] **S**ciascio, D. E., Francesco D. M., Mongiello M., Piscitelli G. (2003), 'Web application design and maintenance using symbolic model checking', in Canfora G., Brand M.V.D., Gymothy T. (eds.) *Software maintenance and reengineering,* Seventh European conference CSMR 2003, Benevento, Italy, IEEE, pp.63-72.

[11] Han M, Hofmeister C. (2006), 'Modeling and verification of adaptive navigation in web applications', in Wolber D., Calder N. Brooks C and Ginige A. (eds.) *Web Engineering, 6th International Conference ICWE 2006*, Palo Alto, California, July 11-14 2006, ACM press, pp.329–336.

[12] Winckler M, Palanque PA. (2003), 'StateWebCharts: A Formal Description Technique Dedicated to Navigation Modelling of Web Applications', in Jorge J., Nunes N. and Cunha J. (eds.) *Interactive Systems. Design, Specification, and Verification, 10th International Workshop 2003,* Maderia Island, Portugal, June 11-13 ,2003, Springer, pp.61–76.

[13] Debasis Kundu and Debasis Samantha (2009), ' A Novel Approach to Generate test Cases from UML Activity Diagrams', *Journal of Object Technology*, Vol-8, No-3, May –June 2009, pp-65-83.

[14] Chehida S. and Rahmouni M.K.(2012), 'Security requirements analysis of web applications using uml', in Malki M., Benbernou S., Benslimane S. and Lehireche A. (eds.) *Web and information technologies, 4th International conference ICWIT 2012.,* April 29-30, 2012, Sidi Bel-Abbes, Algeria, IEEE, pp. 232-230.

[15] David B., Manuel C. and Marina E. (2011), 'A Decade of model driven security', in Ruth B., Cramton J. and Lobo J. (eds.) *Access control models and technologies, 16th ACM symposium*, Innsbruck, Austria, June 15-17, 2011, ACM, pp.1-10.

[16] Zhendong M., Wanger C., Woitsch R., Skopik F. and Bleier T. (2013), 'Model-driven security: from Theory to application', *International journal of computer information systems and industrial management applications,* **5(1)**, pp.151-158.

[17] Furukawa Y. (2011), 'Web based control application using web socket', in Robichon M., Cassady C., Finlay C., Graham Y. (eds.) *Accelerator and large experimental physics control systems, International conference, ICALEPCS 2011,* Grenobal , France, Oct-10-14 2011, pp. 673-675.

[18] Zhangling Y and Mao D.(2012), 'A real time group communication architecture based on web socket' *International journal of computer and communication engineering,* **1(4)**, pp. 408-411.

[19] Nora K., Andreas K. (2003), 'Towards a common meta-model for development of web applications', in Lovelle J., Rodriquez B., Aguilar L., Gayo J. and Ruiz M. (eds.) *Web Engineering, 3rd International conference, ICWE 2003,* Oviedo, Spain, July 14-18, 2003, Springer, pp.497.

[20] Sandhu R., Coyne E. Feinstein H. Youman C. (1996), 'Role-based access control', *Journal of IEEE computer,* **29(2)**, pp. 38-47.

[21] Leung M. (2009), 'Depress phishing by CAPTCHA with OTP', in Luk K. (ed.) *Anti-counterfeiting, security and identification in communication, 3rd International conference ASID 2009*, Hong Kong, August 20-22, 2009, IEEE, pp. 187-192.

[22] Graeme B. (2012), 'Strengthening CAPTCHA based web security', *First Monday peer reviewed journal on internet,* **17(2)**, pp. 1-33.

Automatic Keyphrase Extractor from Arabic Documents

Hassan M. Najadat
Department of Computer information Systems
Jordan University of Science and Technology
Irbid, Jordan

Mohammed N. Al-Kabi
Computer Science Department
Zarqa University
Zarqa, Jordan

Ismail I. Hmeidi
Department of Computer information Systems
Jordan University of Science and Technology
Irbid, Jordan

Maysa Mahmoud Bany Issa
Computer Science Department
Jordan University of Science and Technology
Irbid, Jordan

***Abstract*—The keyphrase is a sentence or a part of a sentence that contains a sequence of words that expresses the meaning and the purpose of any given paragraph. Keyphrase extraction is the task of identifying the possible keyphrases from a given document. Many applications including text summarization, indexing, and characterization use keyphrase extraction. Also, it is an essential task to improve the performance of any information retrieval system. The internet contains a massive amount of documents that may have been manually assigned keyphrases or not. The Arabic language is an important language in the world. Nowadays the number of online Arabic documents is growing rapidly; and most of them have no manually assigned keyphrases, so the user will scan the whole retrieved web documents. To avoid scanning the entire retrieved document, we need keyphrases assigned to each web document manually or automatically. This paper addresses the problem of identifying keyphrases in Arabic documents automatically. In this work, we provide a novel algorithm that identified keyphrases from Arabic text. The new algorithm, Automatic Keyphrases Extraction from Arabic (AKEA), extracts keyphrases from Arabic documents automatically. In order to test the algorithm, we collected a dataset containing 100 documents from Arabic wiki; also, we downloaded another 56 agricultural documents from Food and Agricultural Organization of the United Nations (F.A.O.). The evaluation results show that the system achieves 83% precision value in identifying 2-word and 3-word keyphrases from agricultural domains.**

Keywords*—*Arabic Keyphrase Extraction; Unsupervised Arabic Keyphrase Extraction; Information Retrieval

I. Introduction

The world witnessed during the last two decades an exponential growth in the size of the Internet, which represents the largest heterogeneous reservoir of information. Web documents contain information stored in this global system of interconnected computer networks which is called the Internet. Information stored in the Internet varies in their type, where we can find text, audio, video, images, and other formats.

The Arabic language is one of the six official languages adopted by the United Nations since it ranked the fifth largest natural language among the top 100 used natural languages worldwide. But Arab Internet users ranked 7th worldwide following the users of the following languages, English, Chinese, Spanish, Japanese, Portuguese and German. Arabs constitute 5% of the world population while their Arabic content constitutes only 1% of the Internet content. Although Arab contribution to the Web is one fifth of their population estimates, but on the Internet, there is a large number of Arabic textual documents stored in this giant reservoir of information. Keyphrase extraction is an essential process in information retrieval, document summarization, and clustering. We can extract keyphrases either manually or automatically. Some of the Web textual articles have manually extracted keyphrases. Also, the effectiveness of manual keyphrase extraction is higher than its counterpart automatic keyphrase extraction, but it is costly and slow about automatic keyphrase extraction.

Some studies are conducted to explore the automatic extraction of Arabic keyphrases. This study presents a new unsupervised algorithm to extract Arabic keyphrases from textual documents, where attributes such as Term Frequency-Inverse Document Frequency (TF×IDF), Phrase position, title threshold, terms frequency, phrase frequency, and phrase distribution are used by this novel algorithm to identify keyphrases.

This study is organized as follow: Section 2 presents an overview of the related work to Keyphrase extraction while Section 3 presents the methodology followed to accomplish this study Section 4 presents the results of the tests conducted on our new algorithm while Section 5 presents conclusion remarks and future work.

II. Related Works

First, this section presents a review of few numbers of related studies to our new algorithm. Witten, Paynter, Frank, Gutwin and Nevill-Manning study presents an automatic algorithm called Kea to extract keyphrases from textual documents. Kea uses lexical methods to identify candidate keyphrases, where a score is computed for each candidate keyphrase. Also, Kea adopts machine learning techniques to identify the good candidate keyphrases. Tests were conducted on their algorithm using a large dataset yield a good performance [6].

An interactive tool called PhraseRate to help human classifiers in the Infomine Project is presented by J.B. Keith Humphreys. This tool requires no training and uses Webpage structure to extract keyphrase from those Web pages, where tests on this tool prove its effectiveness [4].

A statistical language model is used by Takashi Tomokiyo and Matthew Hurst to extract keyphrases, where phraseness and informativeness unified into a single score to rank the automatically extracted keyphrases [5]. Turney et al. 2003 [13] exhibit an approach to extract keyphrases, where each document is decomposed into a number of phrases. Each of these phrases is considered as a candidate keyphrase. A supervised learning algorithm is used to identify keyphrases. Another study conducted by Medelyan et al. 2009 [9] shows that providing high- quality features to machine learning algorithm will lead to successfully extracting keyphrases.

Min Song et al. 2003 [8] demonstrate KPSpotter which provides flexible and web-enabled keyphrase extraction by combining the information-Gain data mining measure with multiple NLP methods. This algorithm processes multiple input text formats such as HTML or XML. TF×IDF and distance are measured from first occurrence. Then the attributes are discretized into ranges to calculate the probability of each candidate phrase to be a keyphrase. According to these values, the candidate phrases are ranked to select the most descriptive candidate phrase to be a keypharse. The algorithm was tested on a set of abstracts of some technical reports. Although the experiments showed that both KPSpotter and KEA perform poorly in terms of an average number of matches because of document length, both produce phrases with equal quality.

Quanzhi Li et al 2005 [11] provides a domain specific keyphrase extraction program called Keyphrase Identification Program (KIP). This program extracts a list of candidate noun phrases based on logic. Then, the algorithm sets a score for each term in each candidate phrase. A human-developed glossary database is used to store domain specific keywords and keyphrases and their initial weights. This database contains two tables, one for keyphrase and the other one for keyterm. Each table stores the keyphrase/keyterm and its weight. At first, the keyphrases and terms with their initial weights are defined manually. Then, the learning process takes its role which can be automatic or user-involved. By involving the user in the learning process, the quality of keyphrases can be controlled by the user of the program, he/she can add, remove and highlight any keyphrase he/she wants and the program will respond to that personalization feature.

Samahaa R. El-Beltagy and Ahmed Rafea 2009 [12] propose efficient extraction system for English language called KP-Miner, which uses the simplest version of Poter's stemmer, also they provide adaptation to the system to be able to work with Arabic documents. Although the system does not need training to achieve the extraction task, it was proved by experiments, that the system does good job that is comparable with KEA algorithm.

Also the study conducted by Jiang et al. 2009 [16] emphasize on the importance of using learning by rank techniques to extract keyphrases. Those researchers proposed casting the keyphrase extraction problem as ranking and learning, rather than casting it as a classification (keyphrases and non-keyphrases) using decision tree and Naive Bayes classifiers. Their experiments show that SVM significantly outperforms the others, where learning is exploited. Furthermore, Liu et al. 2010 [19] propose using a Topical PageRank (TPR) on word graph to determine the word importance with respected to different topics. Afterword the distribution of topics within each document is determined, and then the ranking scores of each extracted word are computed. Finally, the top ranked words are considered keyphrases by this method.

Liu et al. 2009 [18] propose unsupervised clustering based method for keyphrase extraction. Using clustering method on a document leads to a creation of different clusters, where the clustering starts with exemplar terms representing the centroid of each newly created cluster, and then all semantically related words and phrases are grouped into a single cluster. They claim that their newly proposed method outperform the sate-of-the-art graph-based ranking methods (TextRank) by 9.5% in F1-measure.

A study is conducted by Wan et al. 2010 [15] proposes the use of a few number of nearest neighbor documents to each document to enhance the process of document summarization and keyphrase extraction. To apply this cornerstone idea a graph-based ranking algorithm is used, where this algorithm uses local information extracted from the document under consideration, and global information extracted from neighbor documents. The tests show clearly the effectiveness and robustness of their proposed method.

According to Alexa, social networking sites like Facebook, Youtube, Twitter, LinkedIn are globally top ranked [1]. A huge number of messages, comments, and views are exchanged within social networking sites. Analyzing this huge number of messages and comments manually is tedious, slow, expensive, and impractical. A study by Zhao et al. 2011 [17] proposes a context-sensitive topical PageRank (cTPR) method to rank different keywords and extract topical keyphrases from Tweeter short messages (Tweets) [14]. This novel method uses a probabilistic scoring function to determine the relevance and interestingness of each keyphrase. Tests show the effectiveness of this method to extract topical keyphrases. Zhao et al. [17] represents an improvement to Liu et al. 2010 [19] study in which they propose using a Topical PageRank (TPR).

El-Beltagy et al. 2009 [12] exhibit in their study a new system to extract Arabic/English keyphrases from textual documents. Their system is called KP-Miner, which needs no training, and characterizes by an equivalent accuracy and sometimes superior to the accuracy of supervised machine learning systems [10, 14] used to extract keyphrases.

On the other hand El-shishtawy et al. 2009 [3] study used supervised learning techniques to extract Arabic keyphrases from Arabic documents. They used a method that does not rely on statistical information such as Term Frequency (TF) and term distances, but relies on linguistic knowledge, which includes syntactic rules based on part of speech (POS) tags. This helps to extract candidate keyphrases. Linear discriminant analysis (LDA) method is used to find a linear combination of linguistic features characterizing keyphrases, where ANOVA (analysis of variance) is used to evaluate each of the selected features. Tests show the effectiveness of this method to extract Arabic Keyphrases.

Al-Kabi et al. 2012 [2] study is based mainly on the Term Frequency (TF) to identify top frequent terms to build a co-occurrence matrix showing the occurrence of each frequent term. If the term is in the biasness degree, then the term is important, and could be considered as a candidate to be a keyword. Words with high $\chi 2$ could be considered a probable keyword, and $\chi 2$ proves it is better to identify keywords than a novel method based on term frequency - inverted term frequency (TF-ITF).

III. Methodology

B This part of the study presents the necessary steps followed to extract Arabic keyphrases extracted from the collected Arabic documents. In this study, around 200 Arabic Web documents collected from Wikipedia website (http://www.wikipedia.org/) and the Website of UN Food and Agricultural Organization (FAO) are used. Fig.1 presents the algorithm of our proposed System (AKEA) which used in this study to extract Arabic Keyphrases.

Consider the following notes related to algorithms shown in Fig.1: The Phrase (Ph) will be nominated as a candidate phrase if its frequency (PF) exceeds 2, since the Keyphrase in Arabic language must exist at least twice within a single paragraph.

After identifying each Arabic Keyphrase in the collection, the following attributes of each candidate Keyphrase are extracted: phrase frequency (PF), summation of phrase terms frequencies (Tf), PF×IDF (Phrase Frequency–Inverse Document Frequency), Phrase Position (Ph_Pos), Title Threshold (T_Thresh) and phrase distribution (Ph_Dist).

Eq. (1) represents PFScore which uses all the attributes mentioned in the previous paragraph. The equation is deduced empirically during conducting a series of tests to extract Arabic Keyphrases.

$$\mathrm{PF}_{\mathrm{Score}} = \left(\frac{1}{\mathrm{Ph\_Pos}+1}\right) + \mathrm{T\_Thresh} + \sum_{i=1}^{\mathrm{Ph\_Len}} \mathrm{TF} + (\mathrm{Ph\_Dist}) + \log_2 \mathrm{PhF} + (\mathrm{PhF} \times \mathrm{IDF}) \qquad (1)$$

Eq. (1) is a combination of adding a number of terms on the right- hand side of Eq. (1). The first term is (1/(Ph_Pos+1)), which represents the reciprocal of Phrase Position, Ph_Pos, plus one to avoid division by zero. This term yields the highest score to phrase at the beginning of each paragraph. This term is based on the idea that Arabic keyphrases lie in most cases at the beginning of each paragraph.

The second term on the RHS within Eq. (1) is T_Thresh. This term yields highest scores to those keyphrases which contain all the terms in the document title.

```
Algorithm: AKEA.
Input: Arabic Textual Document.
Output: List of the Extracted Arabic Keyphrases.
BEGIN
  WHILE Not EOF
    Remove Arabic Stop Words
    Stem Arabic Text
    Compute Term Frequency (TF) of each Arabic
    Identify each Paragraph P in the document
    WHILE NOT END of (P)
      Identify each Phrase Ph in the document
      Compute Phrase Frequency (PF)
      IF (PF) > 2
        Extract Phrase (Ph) attributes
        Compute Phrase score (Pscore)
        Save P, Ph, PF, and Pscore into (Phrases-List)
      END IF
    END WHILE
    WHILE NOT END of Phrases-List
      IF PF > 1
        Choose the highest frequency phrase
      END IF
      IF Ph is a Substring from any phrase in Phrases-list
        Remove Ph from the Phrases-List
      END IF
    END WHILE
  ENDWHILE
  Rank candidate phrases Ph in Phrases-List in descending order
  according to their PFScore
END
```

Fig. 1. Proposed AKEA Algorithm

The third term in Eq. (1) is the summation of term frequencies of the words which the phrase under consideration is consisting of summation keyphrases mostly contains high-frequency words. The expressive words are repeated over all the text. In this term, Ph_Len represents phrase length.

The fourth term in Eq. (1) is Phrase Distribution, Ph_Dist, which gives the probability of the phrase to be appearing in the i^{th} paragraph. So the phrase that has the highest distribution will be the most descriptive one to explain the idea of the paragraph. The frequency of the phrase helps in selecting the candidate phrases and keyphrases. For the keyphrases, they should repeat more than twice in the paragraph. All of the attributes are necessary and each one gives valuable information about the phrase, so that the output of the experiment will be more accurate.

The fifth term in Eq. (1) is $\log_2$PhF, where PhF represents a ratio computed according to Eq. (2):

$$\mathrm{PhF} = \frac{\mathrm{Doc\_PhF}}{\mathrm{Doc\_Total\_Ph}} \qquad (2)$$

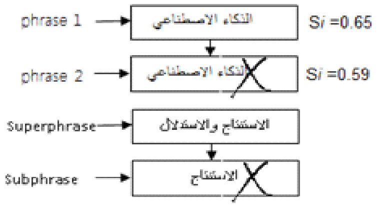


Fig. 2. Example of removing some phrases from candidate phrases list

Where Doc_PhF is a specific phrase frequency in a document, and Doc_Total_Ph is the total number of phrases in that document. The sixth term in Eq. (1) is PhF × IDF, which is the product of the previous ratio (PhF) used in the fifth term by inverse document frequency (IDF). After extracting the phrases of each paragraph and compute the score of each phrase. Some phrases may be repeated more than once if the system extracts the same phrase from different paragraphs. If the phrase exists in the phrases list more than once, the system will choose the highest score phrase and drop the duplicates. Also, it will drop the sub-phrases of some super-phrases to get the final candidate phrases list. Fig. 2 presents two examples that explain how to drop the duplicate phrases and sub-phrases.

IV. Experimental Results

Most of Keyphrase extraction systems must be trained before it can be applied to new documents. In our work, the system will not depend on training because of a large variety of subjects and we do not use domain-specific documents. In this section, we provide the results of our algorithm to extract keyphrases from Arabic documents. We will provide different combinations of the attributes that we used to define the score of the phrases and compare their performance. The performance of individual attributes differs completely from the performance of different combinations of attributes of AKEA system. This is what will be shown in the remaining of this section.

A. *Different Attributes Combinations*

Different combinations of the attributes are provided in this section. The individual attributes which were used in Eq. (1) are: phrase frequency, terms frequency, title threshold, TF×IDF, phrase position and phrase distribution. Using the attributes individually is not beneficial. Single attribute of a phrase does not give any indication about the importance of the phrase in the document. So we try many different combinations of these attributes and compare their results. For each combination of attributes, we compute the mean value of the results of the 100 documents of the dataset.

B. *Single Attribute Performance*

Table 1 shows the performance of different attributes individually in identifying different number of phrases. In this table, the column of number of correct keyphrases displays the fraction of automatic keyphrases over the manual keyphrases, while the column of number of phrases displays how many phrases that chosen from the top ranked phrases. Fig. 3 shows the random behavior for the system which tends to decrease in the average precision value. So we suggest new combinations of attributes that give better results. Now we give some examples of the different combinations and their results.

1) *Two Attributes Combinations*

In this section, we give the performance of different combinations of the five attributes: term frequency, title threshold, TF×IDF, position and distribution with phrase frequency as an example of combining two attributes at a time. Table 2 shows the details of combining phrase frequency with other attributes, one at a time. Fig. 4 shows the relationship between the number of phrases and the precision value for each combination mentioned in Table 2.

TABLE I. Performance of Different Attributes Individually (Experiment 1)

Attribute	**Number of keyphrases**	**Number of correct phrases**
Phrase Frequency	1	0.33
	5	0.41
	10	0.43
Terms Frequency	1	0.27
	5	0.35
	10	0.47
Title threshold	1	0.25
	5	0.34
	10	0.39
Pf×idf	1	0.37
	5	0.4
	10	0.41
Position	1	0.24
	5	0.29
	10	0.38
Distribution	1	0.35
	5	0.36
	10	0.44

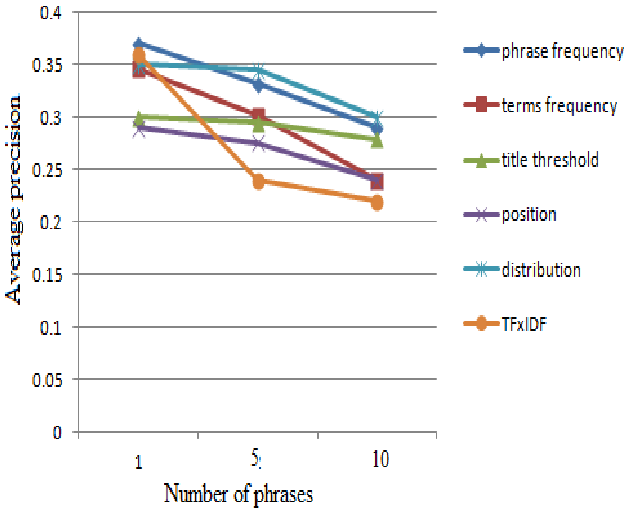


Fig. 3. Comparison of the individual performance of different attributes

TABLE II. Performance of Combining Two Attributes at a Time (Experiment 2)

Combination	**Number of keyphrases**	**Number of correct phrases**
Phrase frequency + term frequency	1	0.33
	5	0.41
	10	0.43
Phrase frequency +Title threshold	1	0.25
	5	0.34
	10	0.39
Phrase Frequency+PF×IDF	1	0.37
	5	0.4
	10	0.41
Phrase Frequency+Position	1	0.24
	5	0.29
	10	0.38
Phrase Frequency+Distribution	1	0.35
	5	0.36
	10	0.44

The information that presented by Fig. 4 confirms that we have to explore other combinations. The highest value of average precision appears when we take the top ten ranked phrases by using phrase frequency and distribution, but we may get a higher value of precision if we try other combination. If we try to combine two attributes at a time we need 15 experiments which are difficult to be explained.

2) *Three Attributes Combinations*

The example that we choose randomly to use here is to combine phrase frequency and phrase position with one attribute at a time from the following four attributes: term frequency, title threshold, TF×IDF and the phrase distribution. Table 3 shows the number of correct phrases for each combination. Keep in mind that the number of correct phrases is equal to the number of correct keyphrases that identified automatically divided by the number of manually identified keyphrases.

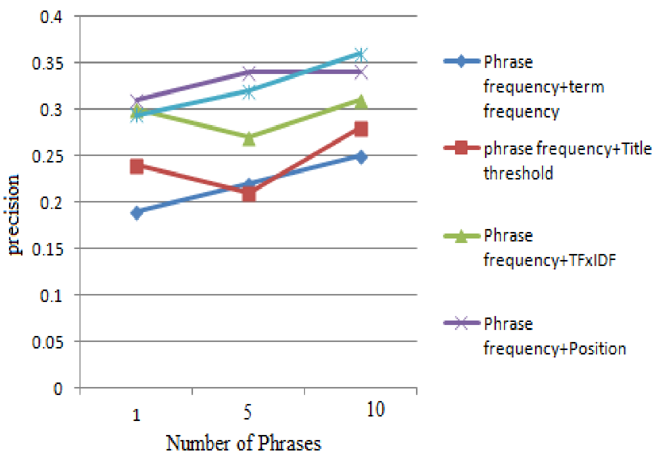


Fig. 4. Phrase frequency combinations performance

TABLE III. PERFORMANCE OF 3-ATTRIBUTE COMBINATIONS (EXPERIMENT 3)

Combination	Number of keyphrases	Number of correct phrases
Phrase frequency + position + term frequency	1	0.35
	5	0.4
	10	0.42
Phrase frequency + position + Title threshold	1	0.29
	5	0.35
	10	0.39
Phrase frequency + position + PF×IDF	1	0.39
	5	0.42
	10	0.42
Phrase frequency + Position + distribution	1	0.4
	5	0.43
	10	0.45

Fig. 5 shows a comparison between the precision values for each combination mentioned in Table 3. The experiments that we mentioned above shows a very convergent precision values except the combination phrase_frequency + position + distribution. This combination gives the highest precision value in increasing manner, but we still need a higher value for precision. For that reason, we try to find an equation that utilizes the advantages of all of the six attributes and combine them together, because all of the attributes are important. In this case no need to try different combinations.

3) *The Best Combination*

Each attribute has its own value that express information about the phrase. Phrase frequency gives the number of occurrences of the phrase in a given paragraph. It is common that the more important phrase will be redundant more than twice in the paragraph. Term frequency attribute represents the summation of phrase terms frequencies. Title threshold gives a value that expresses the relatedness between the phrase and the title of the document. PF×IDF is the combination between phrase frequency (PF) which is the number of occurrences of a specific phrase in a specific document, and inverse document frequency (IDF) which is the log of the ratio between a number of documents in the collection and number of the documents containing a specific phrase.

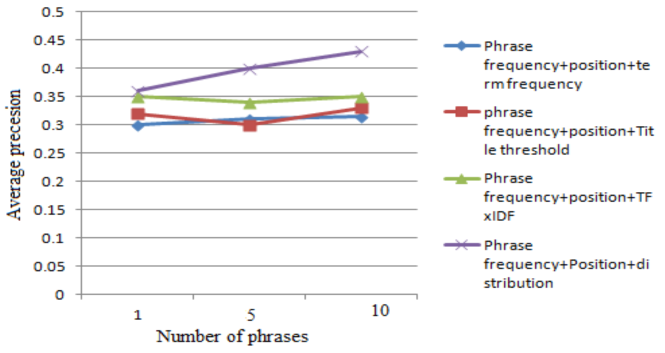


Fig. 5. 3-attribute combinations precision values

The value PF×IDF in our experiments is not very useful since we have non homogeneous document collection. The phrase position attribute is the number of words that precede the first appearance of the first word of the phrase in the paragraph. Lastly, the phrase distribution attribute is the possibility of the phrase appearing in the ith paragraph. We investigate the result of Eq. (1) and display them in Table 4 and Table 5. The value of phrase score PFscore represents the importance of the phrase in a specific paragraph. Fig. 6 presents the results of identifying 2-word keyphrases from stemmed and unstemmed text. Using Eq. (1) we get 0.7 average precision from stemmed text which is the best result of all experiments

TABLE IV. THE PERFORMANCE OF SI EQUATION FOR UNSTEMMED DOCUMENTS

Combination	Number of keyphrases	Number of correct phrases
Si	1	0.43
	5	0.47
	10	0.52

TABLE V. THE PERFORMANCE OF SI EQUATION GOR STEMMED DOCUMENTS

Combination	Number of keyphrases	Number of correct phrases
Si	1	0.54
	5	0.59
	10	0.67

Fig. 7 presents the average precision values the system achieved to identify 2-word and 3-word keyphrases from stemmed and unstemmed datasets.

It is clear that the number of correct phrases and precision values are raised obviously with the top 10 identified keyphrases. The AKEA system has achieved 70% accuracy using precision measure overall 100 test documents in identifying 2-word phrases. Also, it achieved 51% accuracy of precision measure in identifying 3-word keyphrases.

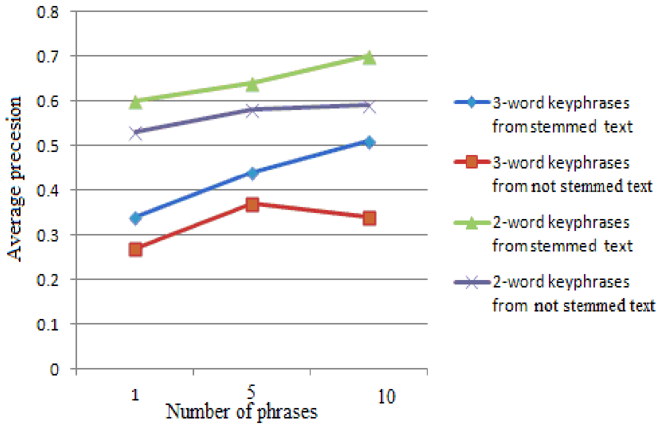


Fig. 6. Comparison between identifying 2-word and 3-word keyphrases from stemmed and not stemmed text

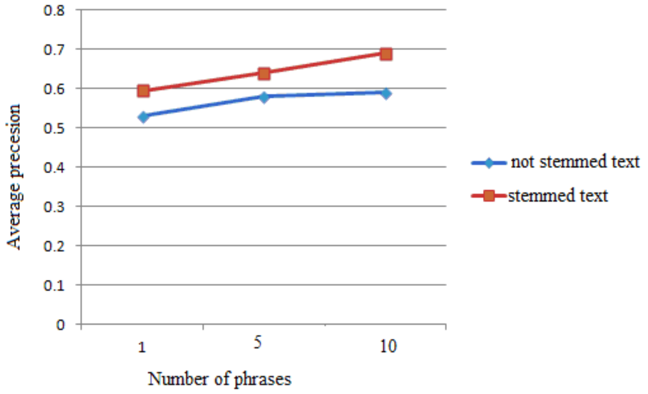


Fig. 7. Comparison between stemmed and unstemmed text output

The final results show that the AKEA system achieved 61% average accuracy of precision measure in identifying 2-word and 3-word keyphrases over all the 100 test documents.

The textual resources that had been used in our project were collected from Wikipedia website. The collection consists of 100 full-text documents and their abstracts that had been randomly downloaded from Arabic wiki. For each document, we run the system twice including using the stemmer [7] and without the stemmer in order to compare the behavior of the system in both cases. After getting the output for each document, we compare the results with the manually extracted phrases. The document collection that had been used to test the results of AKEA system was downloaded from www.ar.wikipedia.org. It contains 100 full-text documents with their abstracts from various domains. This document collection had been used to test KP-miner system [12]. A dataset of our documents and their manual keyphrases is available on www.claes.sci.eg/coe_wm/Data.htm. The average number of words per document in the dataset is in a range between 804 and 934 [12].

Majority of websites such as IEEE (Institute of Electrical and Electronics Engineers) that provides electronic documents provides only the abstract of the documents. AKEA system deals with the abstract like a paragraph, so it can identify keyphrases from any text regardless of the parts. Furthermore, the electronic documents provided by some websites from the types HTML and XML contain HTML/XML tags. These tags are removed by AKEA because they are non-Arabic letters and symbols provided that the input of our system must be a text file from utf-8 format. To investigate the behavior of the system when we provide it with an input that contains HTML/XML tags, a set of documents also downloaded from www.claes.sci.eg/coe_wm/Data.htm. We also test AKEA algorithm on another dataset contains 56 agricultural documents downloaded from FAO.

C. *Evaluation Criteria*

Using the author-assigned keyphrases as a gauge for assessing automatic-extracted keyphrases is logical suggestion because it eases the comparison between both keyphrases groups. Keep in mind author-assigned keyphrases are ranked by their importance, so it will help in evaluating the automatically extracted keyphrase quality. Table 6 shows examples to explain how to assess the keyphrase quality criteria. The column named system phrase contains the phrase identified by the system as a keyphrase, author phrase is the phrase that assigned manually as a keyphrase by the author of the document. The assessment column tells how to assess the system phrase, if the assessment is similar the system phrase is correct keyphrase, otherwise it is incorrect.

1) *Precision and Recall*

Precision and recall are the most famous measure to evaluate the information retrieval systems. When evaluating IR system, the precision is the fraction of retrieved document that are relevant, while recall is the fraction of all relevant documents retrieved. Table 7 explains all the possibilities of a given document in the dataset in an information retrieval system. The measures in Eq. (3) and Eq. (4) are used to evaluate the performance of information retrieval system. In keyphrase extraction system, any phrase might be keyphrase or non-keyphrase identified by the system. In addition, the document author might identify the phrase as keyphrase or non-keyphrase. So we have four possible cases of any phrase. Table 8 shows these possible cases.

According to Table 8, the definition of precision and recall will be as follows: Precision is the ability to retrieve top-ranked phrases that are most relevant. It is the proportion of extracted keyphrases that are correct. It can be calculated according to the following equation: P=A/(A+B), where A represents a number of keyphrases identified automatically and manually, and B represents a number of keyphrases identified automatically but not manually. Recall is the ability of the search to find all relevant phrases in the document. In keyphrase identification systems recall is defined as the proportion of correct keyphrases extracted.

TABLE VI. EXAMPLES OF ASSESSING THE SYSTEM IDENTIFIED PHRASES

System Phrase	Author Phrase	Assessment	Reason of assessment
حاسوب Computer	حساب Computing	Similar	Both phrases have the same stem (compute حسب).
حاسوب Computer	علم الحاسوب Computer Science	Different	The superphrase (علم الحاسوب) gives different meaning from the subphrase (الحاسوب).
علم الحاسوب Computer Science	علم- الحاسوب Computer-Science	Similar	The use of hyphen (–) and the slash (/) is allowed in the middle of the phrase.
علم الحاسوب Computer Science	علم، الحاسوب Computer, Science	Different	Using punctuation is not allowed in the middle of the phrase.
علم الحاسوب Computer Science	ع ح CS	Different	The abbreviation is different from the phrase.

TABLE VII. DOCUMENT CASES

	Relevant	Irrelevant
Retrieved	A	B
Not retrieved	C	D

$$\text{Precision} = A / (A+B) \qquad (3)$$

$$\text{Recall} = A / (A+C) \qquad (4)$$

TABLE VIII. PHRASE CASES

	Identified as keyphrases by the author	Identified as Non-keyphrases by the author
Identified as keyphrases by system	A	B
Identified as Non-keyphrases by system	C	D

The following equation calculates the recall value: R=A/(A+C), where A represents a number of keyphrases identified automatically and manually, and C represents a number of keyphrases identified manually but not automatically.

2) Results

In this section, we provide the results of our algorithm to extract keyphrases from Arabic documents according to our experiments that were explained in the previous subsections. Using the attributes individually is not beneficial. A Single attribute of a phrase does not give any indication about the importance of the phrase in the document. So we use many different combinations of these attributes and compare their results. For each combination of attributes, we compute the mean value of the results of the 100 documents of the dataset. The conducted tests on AKEA using the two types of Arabic documents (stemmed and not-stemmed), that it is better to stem Arabic text before using Arabic Keyphrase extractor as shown in Fig. 8. Fig. 9 presents a comparison between AKEA algorithm and KP-miner in extracting 2-word key-phrases using the same dataset.

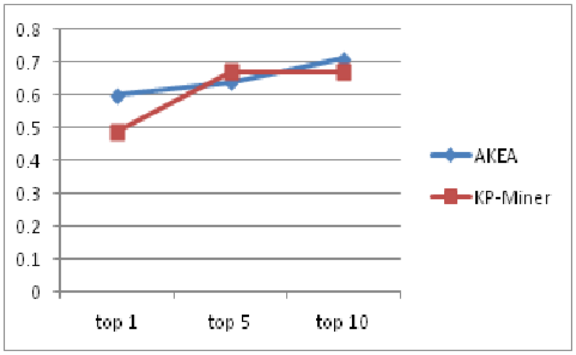


Fig. 8. Identifying 2-word key-phrases in AKEA and KP-Miner

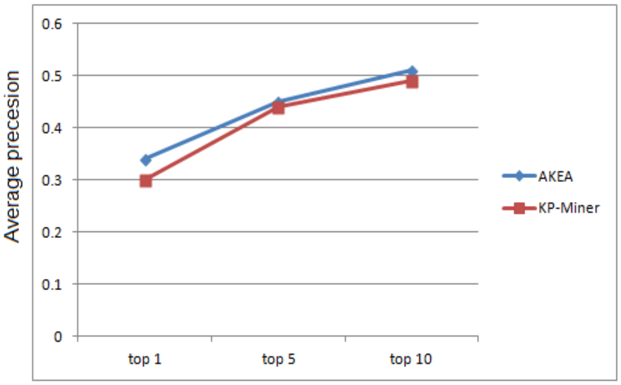


Fig. 9. Identifying 3-word keyphrases with AKEA and KP-Miner

To test the effectiveness of the AKEA algorithm to extract key-phrases from a domain-specific dataset, a collection of 56 various agricultural documents were collected from Food and Agriculture Organization of the United Nations (FAO) Website (http://www.fao.org). AKEA yields 83% precision to extract the top 10 key-phrases from this agricultural collection.

V. CONCLUSION

This study presents a novel supervised Arabic key-phrase detection algorithm using a limited dataset of around 200 Arabic Web documents collected from Arabic Wikipedia and Food and Agricultural Organization of the United Nations (FAO). This algorithm yields satisfactory accuracy results.

Future work includes the use of a larger dataset to test an enhanced version of our proposed algorithm, where new attributes will be adopted to improve the effectiveness of this algorithm.

REFERENCES

[1] Alexa Top 500 Global Sites, Available at: http://www.alexa.com/topsites (Accessed September 9, 2015).

[2] M. Al-Kabi, H. Al-Belaili, B. Abul-Huda, and A. Wahbeh, "Keyword extraction based on word co-occurrence statistical information for arabic text", Abhath Al-Yarmouk:Basic Science & Engineering., 22 (1), pp. 75- 95, 2013.

[3] T. El-Shishtawy and A. Al-Sammak," Arabic keyphrase extraction using linguistic knowledge and machine learning techniques", Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009.

[4] K. Humphreys, "PhraseRate: An HTML keyphrase extractor", Technical report, University of California, 2002.

[5] H. T. Tomokiyo and M. Hurst, “A language model approach to keyphrase extraction”, Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment, vol. 18, pp. 33-40, 2002.

[6] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G Nevill-Manning, “KEA: practical automatic keyphrase extraction”, Proceeding of the fourth ACM conference on Digital libraries, pp. 254-255, 1999.

[7] Java version of Shereen Khoja Arabic stemmer, Available at: http://zeus.cs.pacificu.edu/shereen/ArabicStemmerCode.zip (Accessed September 9, 2015).

[8] M. Song, I. Song, and X. Hu, “KPSpotter: a flexible information gainbased keyphrase extraction system”, Proceedings of the 5th ACM international workshop on web information and data management, pp. 50-53, 2003.

[9] O. Medelyan, E. Frank, and I. H. Witten, “Human-competitive tagging using automatic keyphrase extraction”, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: vol. 3. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1318-1327, 2009.

[10] P. D. Turney, “Learning algorithms for keyphrase extraction”, Information Retrieval, 2(3), pp. 303–336, 2000.

[11] L. Quanzhhi, W. Y. Brook, B. R. Stefan, and C. Xin, "Automatically finding significant topical terms from documents", AMCIS 2005 Proceedings, 2005.

[12] S. R. El-Beltagy and A. A. Rafea, “KP-Miner: A keyphrase extraction system for English and Arabic documents”, Inf. Syst., (34), pp. 132-144, 2009.

[13] P. D. Turney, “Coherent keyphrase extraction via web mining”, Proceedings of the 18th international joint conference on Artificial intelligence, pp. 434-439, 2003.

[14] Twitter, Available at: http://twitter.com/ (Accessed September 9, 2015).

[15] X. Wan and J. Xiao, “Exploiting neighborhood knowledge for single document summarization and keyphrase extraction”, ACM Trans. Inf. Syst, 28, 2, 2010.

[16] X. Jiang, Y. Hu, and H. Li, "A ranking approach to keyphrase extraction", Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09), New York, NY, USA, pp. 756-757, 2009.

[17] W. X. Zhao and et al., "Topical keyphrase extraction from Twitter", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 379-388, 2011.

[18] Z. Liu, P. Li, Y. Zheng, and M. Sun, "Clustering to find exemplar terms for keyphrase extraction", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09), vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 257-266, 2009.

[19] Z. Liu, P. Li, Y. Zheng, and M. Sun, "Automatic keyphrase extraction via topic decomposition", Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10), Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 366-376, 2010.

A Topic Modeling Based Solution for Confirming Software Documentation Quality

Nouh Alhindawi[1]
Faculty of Sciences and Information Technology, JADARA UNIVERSITY
Jordan

Obaida M. Al-Hazaimeh[2]
Department of Information Technology,
AL-BALQA' APPLIED UNIVERSITY
Jordan

Rami Malkawi[3]
Faculty of Sciences and Information Technology,
JADARA UNIVERSITY
Jordan

Jamal Alsakran[4]
King Abdullah II School for Information Technology,
THE UNIVERSITY OF JORDAN
Jordan

***Abstract*—this paper presents an approach for evaluating and confirming the quality of the external software documentation using topic modeling. Typically, the quality of the external documentation has to mirror precisely the organization of the source code. Therefore, the elements of such documentation should be strongly written, associated, and presented. In this paper, we use Latent Dirichlet Allocation (LDA) and HELLINGER DISTANCE to compute the similarities between the fragments of source code and the external documentation topics. These similarities are used in this paper to improve and advance the existing external documentation. Furthermore, these similarities can also be used for evaluating the new documenting process during the evolution phase of the software. The results show that the new approach yields state-of-the-art performance in evaluating and confirming the existing external documentations quality and superiority.**

Keywords—Software Documentation; LDA; Clusters; HELLINGER DISTANCE; and Information Retrieval

I. Introduction

Modern software often consists of thousands of software development artifacts, such as external documents, design documents, code, bug reports, and test cases. These different kinds of documents are used by different kinds of people, such as developers, testers and also the end customers or clients. Therefore, writing these documents in a clear, easy, and understandable way is considered as an attribute for ideal software development and maintenance processes.

Typically, Software Documentation faults and oversights can increase the errors caused by software engineers. Moreover, it wastes developer’s time and increases maintenance costs. For that reason, software engineers should pay much attention to documentation process. Moreover, Software Documentation quality is as significant as program quality. Any missing information about how to use the system, or how the system works, will cause the system to be degraded [1-3].

The external documentation describes each feature of the program, and assists the user in realizing these features, specially the new ones. Moreover, the external documentation can also go thus far as to supply thorough troubleshooting support. Generally, the external documentations are helpful in software engineering for development, maintenance, and evolution processes. Therefore, the external documentation should not be confusing, and they should be up to date. The assumption here is that external superiority documentation has to mirror precisely the organization of the source code. However, the external documentation and, where necessary, the system design and implementation, should be ideally modified and structured, so that changes can be easily documented and considered via external documentation correspondingly.

In this paper, a new methodology is presented that can be used to confirm the existing external documentation quality and superiority. The new approach for document assessment and confirmation consists of building models for source code and models for source code external documents using LDA. We compute the similarity between the documents distribution of the two models using Hellinger Distance.

Thus, we improve the techniques that were developed to deal with documentation quality assessment by integrating topic modeling with structural similarity measures to assess the quality of existing documentation.

In order to provide a base for our new external documentation confirming approach, we will now give more details about LDA modeling as well as a brief introduction to Hellinger Distance.

II. Extracting Topics with Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [4] is a popular technique for getting probabilistic topic models from textual corpora by means of a generative process. LDA model is based on a fully generative model; for each document in the entire corpus, LDA represents it as a mixture of linguistic topics.

That is, LDA use the probability distribution over the gained topics to represent each document. In other words, each document is modeled using LDA as multi-membership mixture of K-topics. Moreover, each topic is also represented as multi-Membership mixture of the corpus terms that exist in the vocabulary.

Using LDA, the corpus can be represented by a set of topics, and each document in the corpus can also be described by more than one of these topics. Moreover, each term from the corpus can be included in more than one of these topics. Therefore, any of corpus documents is not limited to being associated with a single topic, but as an alternative, it is modeled in a way that considers the possibility that document may address multiple topics.

Given *S* documents containing *k* topics stated over *u* unique words (w) the distribution of *i-th* topic to *i* over u words can be represented by φi and the distribution of *j-th* document ,document i (*doc* i) over k topics can be represented by *θj*.

The LDA assumes the following generative process for each document *doc i* in a corpus *D*:

- Choose N ~ Poisson distribution (ξ)
- Choose θ ~ Dirichlet distribution (α)
- For each of the N words w_i:
  - Choose a topic (k) to_k ~ Multinomial (θ).
  - Choose a word w_i from $P(w_i|z_n,\beta)$, a multinomial probability conditioned on topic to_k.

As conclusion, given a corpus of documents, LDA tries to discover the following:

- Recognizing a set of topics.
- Relates a set of words with a topic
- Specifies an exact mixture of these topics for each document in the corpus.

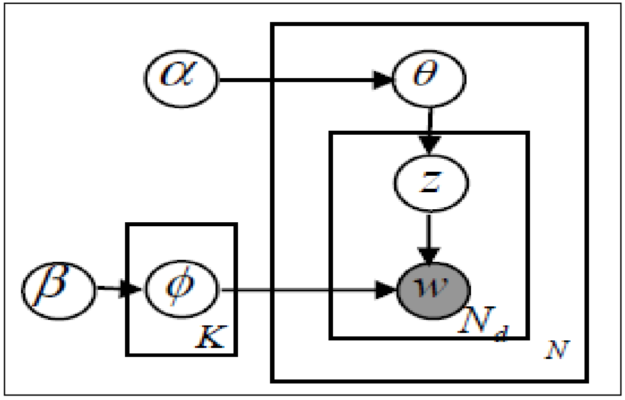


Fig. 1. LDA model. *K* is the number of topics; *N* is the number of documents; *Nd* is the number of word tokens in document *d*

For more details regarding LDA model, we refer the readers to Blei et al. work [4]. As mentioned before, LDA permits a document to have a combination of topics as we see in Figure 1. Moreover, the LDA model allows a document to exhibit multiple topics to different degrees, thus being more flexible than the cluster based techniques.

III. Hellinger Distance

Using HELLINGER DISTANCE with LDA modeling is our main contribution, as it achieves promising results. Using LDA proved its performance in locating and modeling any software artifacts, on the other hand, HELLINGER DISTANCE is also used in the literature as one of the standard methods that can compute the similarities between any dissimilar clusters probability distribution [5]. The main idea of our approach is to use the HELLINGER DISTANCE between document topics distributions to find the most likely similar and relevant topics from the two corpuses (SC and ED):

$$\mathrm{D}(\emptyset 1,\emptyset 2)=\sqrt{\frac{1}{2}\sum\nolimits_{\mathrm{t}=1}^{\mathrm{T}}\left(\sqrt{\emptyset 1,\mathrm{t}}-\sqrt{\emptyset 2,\mathrm{t}}\right)^{2}}$$

IV. Documentation Assessment and Confirming with LDA and Hellinger Modeling

The proposed methodology is based on a set of parallel and sequential steps, which are partially automated:

STEP1. Extracting source code artifacts

STEP2. Extracting documents from external documentation.

STEP3. Building a corpus for source code artifacts

STEP4. Building a corpus for external documentation

STEP5. Extracting source code corpus topics (SC)

STEP6. Extracting external documentation corpus topics (ED)

STEP7. Computing the HELLINGER DISTANCE between the documents of SC and ED

STEP8. Analyzing the topics documents similarities

As shown in Figure 2, the process is done in pipeline architecture, in other words, the output from one phase constitutes the input for the next phase. The source code artifacts and the external documentation are used to create the corpuses that are used to generate the semantic space for Information Retrieval (IR) (see steps 1, 2, 3, and 4). The semantic topics produced from LDA for both corpuses are automatically generated in phases 5 and 6. More details about this step can be found in [2, 5, 6]. Once the topics of both corpuses are generated, the HELLINGER DISTANCE between the two corpuses documents is computed. As a final phase, we analyze the topics documents similarities, we use the similarities between source code topics and documents to infer missing associations or cross-references between existing sections of a documentation or suggest relations for the new documentation and source code.

In the following paragraphs, we present with details the corpuses building steps, the topics generating procedure, and the HELLINGER DISTANCE method.

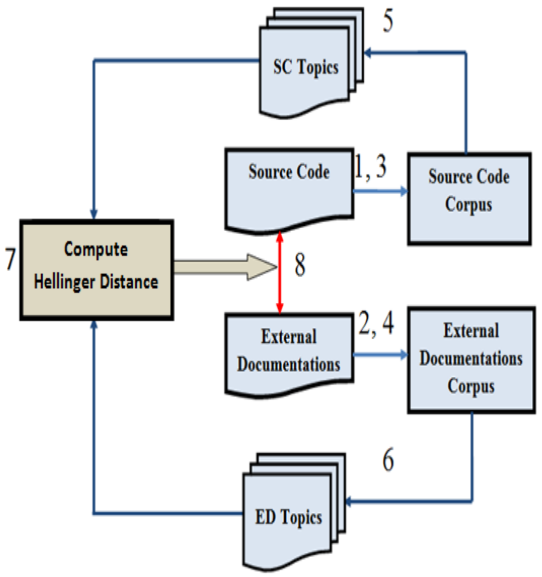


Fig. 2. Steps of Documentation Assessment with LDA and HELLINGER Modeling Approach

STEP 1: as an initial step for building source code corpus, we prepare the collections of artifacts which make up the corpus that LDA can process and infer. This is achieved by extracting all the textual information associated with a given source code; all the words used in comments or identifiers inside the method, class or package are extracted using our efficient corpus builder which was implemented in C++ to extract these important elements from source code that in XML format.

It takes less than 30 seconds to build both corpuses (corpora for of the two systems we used in the experiments). We use SrcML [7] tool to transform the C++ source code to XML format.

STEP 2: The same steps mentioned above for extracting source code artifacts are performed here to prepare the artifacts of the external documentation which make its corpus that LDA can infer. This is also done by extracting all the natural language information associated with a given source code; all the words used in include user documents (e.g., HTML,XML/docbook, LaTeX and Doxygen), build management documents (automake, cmake, and makefile), HowTo guides (e.g., FAQs), release and distribution documents (e.g.,ChangeLogs, whatsNew, README, and INSTALL guides), progress monitoring documents (TODO and STATUS), and extensible mechanisms (e.g., Python, Ruby, and Pearl bindings for an API) [2, 8, 9].

STEP 3: For both corpuses, we preprocess the words that can be found in both corpuses, starting by running them through a tokenizer. This allows us to split identifier names written with camel case or underscores (i.e., CamelCase or under_score) into their component words, giving us a better idea of what natural language topics and words are used in implementation.

STEP 4: In this step, we get rid of a set of reserved keywords and some other words that are very commonly used, such as “the” and “get”. Our approach allows the developer to specify easily any other stop words list.

STEP 5: The next step taken is stemming the words that make up our corpuses. Stemming includes removing the endings from words in order to recognize any corpus word despite what grammatical usage it appears in. We use porter default English stemmer [6].

STEP 6: After completing the previous steps, we are now able to generate and compute the topics with LDA. We use the LDA implementation provided by the Gensim library. Subsequently here, we choose the parameters to use in the computation, and then we extract topics from the documents. More details about this step are covered later in the following sub-sections.

STEP 7: In order to extract relevancy between the two corpuses linguistic topics, we use HELLINGER DISTANCE approach in two manners; in the first one, we compute the similarity between the topic i from ED topics and all the SC topics, while in the other one, we compute the similarities between all ED topics and all SC topics at the same query. Thus, we propose the following two methods for extracting and computing the two corpuses topics similarities: multi-topic and single-topic.

A. *Single- Topic (LDA-S)*

The LDA model is built based on all of the training documents of the source code. Given an ED test topic, we measure the HELLINGER DISTANCE between this topic distribution and the distributions of all SC topics. The SC topics with the lowest mean distance are returned as the most likely relevant SC topics to the taken ED topic. That is, the ED topic is queried over SC topics to retrieve the most similar topics.

B. *Multi- Topic (LDA-M)*

Here, the similarities between all SC and ED topics are measured, the result of this step is a ranked list that contains and shows any of SC and ED topics that have the maximum similarity percentage. Once the list is retrieved, the developer can distinguish and locate the related topics from both corpuses.

V. Experiments Setup and Datasets

In this section, we describe the experimental setup and datasets used in our experiments, followed by the evaluation of our new approach.

We conducted our experiments over KDE/KOFFICE open source system. We performed LDA topics modeling for both of KDE/KOFFICE source code system and over its external documentation. The evaluation of the new approach is done by comparing how many relevant topics from both corpuses were retrieved as relevant in the retrieved list, and the number of traceability links that exist between the two corpuses, which we found in our previous work [10].

Table I, shows the elements and the attributes for both of the two corpuses we built for KDE/KOFFICE system.

TABLE I. ARTIFACTS OF THE KDE/KOFFICE SYSTEM

KDE/KOffice	Count	Documents
Source Code Files	1057	11492
Non-Source Code Files	89	102
Total of External Documents		11594
Vocabulary	12839	_

The goal in [10], was to uncover traceability between source code and other artifacts using the TraceLab [11]. As mentioned before, this includes: user documents (e.g., HTML,XML/docbook, LaTeX and Doxygen), build management documents (automake, cmake, and makefile), How To guides (e.g., FAQs), release and distribution documents (e.g.,ChangeLogs, whatsNew, README, and INSTALL guides), progress monitoring documents (TODO and STATUS), and extensible mechanisms (e.g., Python, Ruby, and Pearl bindings for an API).

We performed the required preprocessing of the input texts. Both of the source code and the external documentation need to be broken up into the proper granularity to define the corpuses documents, which will be represented as vectors [2, 9, 12-14]. Therefore, we split up the source code into documents with function granularity level. As a result, each function has a corresponding document in the corpus of source code; this document contains the function name, local variable, global variable, function calls, and the internal comment of that function.

For external documentation, the paragraph is used as the granularity level. Table I contains the size of the system, as well as the dimensionality used for the LDA subspace and the determined vocabulary. For the LDA parameters, we can change the number of topics to be generated, as well as other LDA parameters, such as a number of iterations used and values of alpha and beta.

Typically, LDA model takes two parameters *Alpha* and *Beta,* where *Alpha* controls the division of documents into topics and *Beta* controls the division of topics into words. Larger values of *Beta* yields coarser topics, and larger values of *Alpha* yields coarser sharing of document into topics. For this reason the correct values of *Alpha* and *Beta* are required to obtain fine quality topics and to link topics to the original documents. A number of LDA implementations estimate these values on-the-fly while other implementations rely on the user to provide appropriate values [6, 15, 16].

We followed the recommendations in Gensim documentation, and set the Dirichlet hyper parameters to *Alpha*= min (0.1, 50/T) and *Beta* = 0.01, varying only the number of topics T. We ran the Gensim sampling process for *S* = 1000 iterations, and based the document representations on the last sample.

VI. EVALUATION AND DISCUSSION

The results are evaluated using categorization accuracy, i.e., the percentage of test documents topics that were correctly assigned to its corresponding source code topics. Moreover, we employ diverse accuracy series in the figures that reflect our results for precision of presentation.

The results show that using the LDA topic modeling along with the HELLINGER DISTANCE for confirming and linking the external documentation to its related source code fragments is working efficiently. As mentioned before, these outcomes have been proved using the already uncovered traceability links as shown in Table II.

In other words, for each of the extracted ED topics, we measured the HELLINGER DISTANCE between each of them and all of SC topics. We consider that a topic x from SD topics is related to set of topics from ED if the HELLINGER DISTANCE between them is the smaller. Thus, The SC topics with the lowest mean distance with respect to ED topic are returned as the most likely related topics. We called the related two topics as a pair. Next, we compared the pairs we have with the uncovered traceability links we found in our previous work [10].

TABLE II. DISCOVERED LINKS AND RECALL USING COSINE VALUE THRESHOLD

Cosine threshold	Total Links Recovered	Recall
0.60	184	84.2%
0.70	95	61.79%

In some cases, part of the documentation may refer to more than one source-code document, or a source-code document may be described by more than one external document. This fact has been proved in the results here, 103 ED topics based on the distance measure appear to be relevant to more than one SC topics, and this result confirms the efficiency of the proposed approach in spotting the relevancy between the source code fragments and between the significant external documentation.

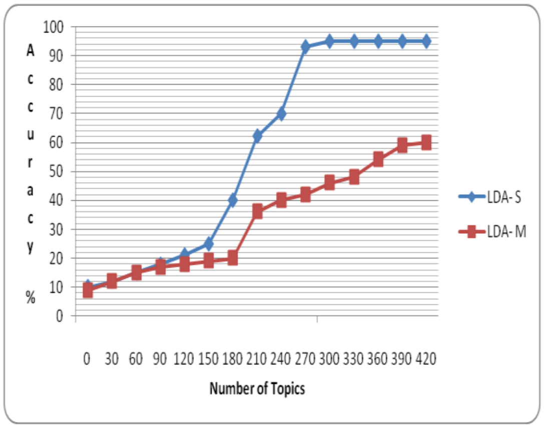


Fig. 3. Results of LDA-S and LDA-M

One notable result here, that 45 ED topics have poor relevancy with respect to all of SD topics. When investigating those topics, we found that most of them refer to authorship information and non-functional requirement information such as security recommendations. We argue here that labeling the external documentations that have such kind of information would be very efficient for developer's progression. For our experiments, we ran our LDA+Hellinger alternatives with 20, 40, 60 . . ., 300 and 400 topics. For LDA-M, the best accuracy we obtain is when the number of topics equal 400 as shown in Figure 3. However, LDA-S yielded a much higher accuracy than LDA-M.

Table III, shows the accuracy of investigated pairs matching compared with the recovered links. The second column in the table represents the number of pairs that were investigated, and the third column represents the percentage of accepted investigated pairs with respect to uncovered links. As we see in the table, LDA-S performs better accuracy that LDA- M with 226 investigated pairs. However, LDA-M only performs 0.30 as accuracy despite of the huge number of pairs that were retrieved within the specified threshold.

TABLE III. THE ACCURACY FOR BOTH TECHNIQUES (LDA-S AND LDA-M) WITH 0.25 AS THRESHOLD

Mechanisms	Number of Retrieved Pairs	Accuracy
LDA- S	226	0.80
LDA- M	391	0.30

When comparing the results of the two mechanisms (LDA-S and LDA-M), we note that LDA-S gives high precision even when only few topics are used, as we see in Table IV, The second column (Total links retrieved) represents the total number of recovered links (correct + incorrect), the third column (K value) represents number of topics that gives the best accuracy for each mechanism.

As we see the in the table, the difference between LDA-S and LDA-M is statistically significant. As we see, LDA- S discovered 181 traceability links, where LDA-M discovered 117 traceability links. The Table also shows the best K (number of topics) value where each mechanism gives the best accuracy.

TABLE IV. THE TOTAL NUMBERS OF LINKS WHICH DISCOVERED USING LDA-HELLINGER. K EQUALS THE NUMBER OF TOPICS THAT GIVES THE BEST ACCURACY FOR EACH MECHANISM

Mechanisms	Total Links Recovered via Matched Pairs	K- Value
LDA- S	181	300
LDA- M	117	420

An advantage of LDA-S over LDA-M is that LDA-S requires much less time to classify a test document when many SD per ED are available. However, this improvement in runtime may come at the punishment of accuracy and precision. The reason that LDA-M do better when more topics are considered may be that some important source code concepts are distributed to longer documents. That is, one concept/feature of source code fragments can be described by one or more external documentation. Furthermore, one source code concept/feature can usually be implemented by different parts of source code.

VII. RELATED WORKS

Several approaches have been developed in the past two decades to assist developers in obtaining an overview of the source code artifacts including the fragments of code, and the internal and the external documentation. However, the previous research in this area is limited. IR methods are considered as one of the most successful approaches in this field of research i.e., LSA and LDA [2, 8, 17].

There is a substantial amount of research which illustrates the relevance and the importance of documentation quality in the context of software evolution and development. Chen et al [18] presented the documentation quality problems as a major key problem in the domain of software engineering along with the main principles for writing the documentation for any software.

In [19], the author presented an automated quality assessment approach for software documentation using a developed document quality analysis framework and software document quality rules and principles.

Another framework for assessing documentation adequacy is also presented in [20], the authors mainly used a predefined taxonomic structure to assess a project documentation which funded by Naval Surface Warfare Systems (NSWC). Based on their findings of the authors, there is a need for a tool and method that can automatically evaluate any software documentation quality especially for large systems.

LDA was utilized for the first time to locate concepts in source code Linstead et al [21] by extracting the source code topics using LDA. Their approach can extract the concepts exist within the identifiers and the comments in the source code. Baldi et al [22] proposed a theory that software concerns are equivalent to the latent topics found by statistical topic models. They applied their approach to identify the global set of topics in many large systems.

In [16], LDA was utilized with the goal of enhancing and improving the process of analyzing the process of software evolution. Based on the results of the paper, the evolution process of software is more comprehendible when using the topics generated by LDA.

In [5], the authors use the HELLINGER DISTANCE between document topic distributions to find the most likely author of a specific document. Maskeri et al [23] considered the usage of the topics extracted with LDA from a software system.

Moreover, Classifying software systems into related groups in automatic way using LDA has been presented by Tian et al [24]. LDA was utilized to find traceability links between bug reports and program code by Lukins et al [25], their evaluation showed that LDA often drastically outperforms LSI.

In [8], Latent Semantic Indexing (LSI) was applied and utilized in order to find the similarities between fragments of code, the proposed approach aided the programmers when comprehending source code by clustering the similar and related fragments of source code. Moreover, LSI was also used in [13], the authors utilized and enhanced the usage of LSI to be used as a mapping technique for the concepts which expressed in natural language by relating them to their related fragments of code.

Topic Modeling was employed by the authors in [26], they used LSI to semantically cluster the artifacts which have similar or common vocabulary. The yielded clusters or groups are then linked based on the similarity between them along with visualization for these clusters. Moreover, labels are retrieved automatically for each cluster and for the linked ones. The visualization which provided by the authors can help greatly in program comprehension process.

A study on software documentation quality in practice was conducted and presented in [27]. The authors presented a survey which categorizes the current state of software documentation quality and employed analysis approaches for achieving software documentation quality checking process. Based on their findings, they confirm that the most significant quality characteristics for the documentation quality are precision, clearness, constancy, and readability.

VIII. Conclusion

In this paper, an approach to evaluate and confirm the existing external documentation quality and superiority is presented. The new approach uses Latent Dirichlet Allocation (LDA) along with HELLINGER DISTANCE to compute the similarities among the source code artifacts and its external documentation. A set of experiments was presented and the results validated by comparing them with uncovered links extracted in previous work over KDE/Koffice system.

The results show clearly that the new approach proved its efficiency in classifying and confirming the quality of source code external documentation. Moreover, based on the results, we argue here that labeling and grouping the external documentation would impact positively on the quality of the documentation. Based on the results we found, the needs for tools that can assess the software documentation quality in an automatic way are highly demanded.

Acknowledgement

The authors would like to thank faculty of sciences and information Technology, JADARA UNIVERSITY for support funding to carry out this research project.

References

[1] IEEE Standard for Software User Documentation. IEEE Std 1063-2001, 2001: p. 1-24.

[2] Marcus, A. and J.I. Maletic. Recovering documentation-to-source-code traceability links using latent semantic indexing. in Software Engineering, 2003. Proceedings. 25th International Conference on. 2003.

[3] Marcus, A. and J.I. Maletic, Recovering documentation-to-source-code traceability links using latent semantic indexing, in 25th International Conference on Software Engineering2003, IEEE Computer Society: Portland, Oregon. p. 125-135.

[4] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. J. Mach. Learn. Res., 2003. 3: p. 993-1022.

[5] Seroussi, Y., I. Zukerman, and F. Bohnert, Authorship attribution with latent Dirichlet allocation, in Proceedings of the Fifteenth Conference on Computational Natural Language Learning2011, Association for Computational Linguistics: Portland, Oregon. p. 181-189.

[6] Savage, T., et al. TopicXP: Exploring topics in source code using Latent Dirichlet Allocation. in IEEE International Conference on Software Maintenance (ICSM). 2010. IEEE Computer Society.

[7] Collard, M.L., M.J. Decker, and J.I. Maletic. Lightweight Transformation and Fact Extraction with the srcML Toolkit. in IEEE 11th International Working Conference on Source Code Analysis and Manipulation. 2011.

[8] Maletic, J.I. and A. Marcus. Using latent semantic analysis to identify similarities in source code to support program understanding. in 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI) 2000.

[9] Alhindawi, N., et al., A TraceLab-Based Solution for Identifying Traceability Links using LSI, in 7th ACM International Workshop on Traceability in Emerging Forms of Software Engineering (TEFSE)2013: California, USA. p. 79-82.

[10] Alhindawi, N., et al. LSI-Based Solution for Categorizing Software Repositories Commits for Maintenance in Working Conference on Reverse Engineering (WCRE). To Be Submmitted. 2013.

[11] Keenan, E., et al. TraceLab: An experimental workbench for equipping researchers to innovate, synthesize, and comparatively evaluate traceability solutions. in 34th International Conference on Software Engineering (ICSE) 2012.

[12] Alhindawi, N., et al. Improving Feature Location by Enhancing Source Code with Stereotypes. in International Conference on Software Maintenance (ICSM) Submitted. 2013.

[13] Marcus, A., et al. An Information Retrieval Approach to Concept Location in Source Code. in 11th Working Conference on Reverse Engineering. 2004. IEEE Computer Society.

[14] Poshyvanyk, D. and A. Marcus. Combining Formal Concept Analysis with Information Retrieval for Concept Location in Source Code. in 15th IEEE International Conference on Program Comprehension (ICPC). 2007.

[15] Tian, K., M. Revelle, and D. Poshyvanyk. Using Latent Dirichlet Allocation for automatic categorization of software. in 6th IEEE International Working Conference on Mining Software Repositories (MSR). 2009. IEEE Computer Society.

[16] Linstead, E., C. Lopes, and P. Baldi. An Application of Latent Dirichlet Allocation to Analyzing Software Evolution. in Seventh International Conference on Machine Learning and Applications. 2008. IEEE Computer Society.

[17] Binkley, D. and D. Lawrie, Information Retrieval Applications in Software Maintenance and Evolution, in Encyclopedia of Software Engineering, P. Laplante, Ed.2010: Taylor & Francis LLC.

[18] Chen, J.-C. and S.-J. Huang, An empirical analysis of the impact of software development problem factors on software maintainability. J. Syst. Softw., 2009. 82(6): p. 981-992.

[19] Dautovic, A., Automatic assessment of software documentation quality, in Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering2011, IEEE Computer Society. p. 665-669.

[20] Arthur, J.D. and K.T. Stevens, Document quality indicators: A framework for assessing documentation adequacy. Journal of Software Maintenance: Research and Practice, 1992. 4(3): p. 129-142.

[21] Linstead, E., et al. Mining Eclipse Developer Contributions via Author-Topic Models. in Mining Software Repositories, 2007. ICSE Workshops MSR '07. Fourth International Workshop on. 2007.

[22] Baldi, P.F., et al., A theory of aspects as latent topics, in Proceedings of the 23rd ACM SIGPLAN conference on Object-oriented programming systems languages and applications2008, ACM: Nashville, TN, USA. p. 543-562.

[23] Maskeri, G., S. Sarkar, and K. Heafield, Mining business topics in source code using latent dirichlet allocation, in Proceedings of the 1st India software engineering conference2008, ACM: Hyderabad, India. p. 113-120.

[24] Kai, T., M. Revelle, and D. Poshyvanyk. Using Latent Dirichlet Allocation for automatic categorization of software. in Mining Software Repositories, 2009. MSR '09. 6th IEEE International Working Conference on. 2009.

[25] Lukins, S.K., N.A. Kraft, and L.H. Etzkorn. Source Code Retrieval for Bug Localization Using Latent Dirichlet Allocation. in Reverse Engineering, 2008. WCRE '08. 15th Working Conference on. 2008.

[26] Rousidis, D. and C. Tjortjis. Clustering Data Retrieved from Java Source Code to Support Software Maintenance: A Case Study. in Software Maintenance and Reengineering, 2005. CSMR 2005. Ninth European Conference on. 2005.

[27] Plosch, R., A. Dautovic, and M. Saft. The Value of Software Documentation Quality. in Quality Software (QSIC), 2014 14th International Conference on. 2014.

A New Hybrid Network Sniffer Model Based on Pcap Language and Sockets (Pcapsocks)

Azidine GUEZZAZ, Ahmed ASIMI, Yassine SADQI, Younes ASIMI and Zakariae TBATOU
Laboratoire LabSiv: Systèmes d'information et vision
Equipe SCCAM : Sécurité, Cryptographie, Contrôle d'Accès et Modélisation
Department of Mathematics and Computer Sciences
Faculty of Sciences, Ibn Zohr University, B.P 8106, City Dakhla, Agadir, Morocco

***Abstract*—Nowadays, the protection and the security of data transited within computer networks represent a real challenge for developers of computer applications and network administrators. The Intrusion Detection System and Intrusion Prevention System are the reliable techniques for a Good security. Any detected intrusion is based on data collection. So, the collection of an important and significant traffic on the monitored systems is an interesting feature. Thus, the first task of Intrusion Detection System and Intrusion Prevention System is to collect information's basis to treat and analyze them, and to make accurate decisions. Network analysis can be used to improve networks performances and their security, but it can also be used for malicious tasks. Our main goal in this article is to design a reliable and powerful network sniffer, called PcapSockS, based on pcap language and sockets, able to intercept traffic in three modes: connected, connectionless and raw mode. We start with the performances assessment performed on a list of most expanded and most recently used network sniffers. The study will be completed by a classification of these sniffers related to computer security objectives based on parameters library (libpcap/winpcap or libnet), filtering, availability, software or hardware, alert and real time. The PcapSockS provides a nice performance integrating reliable sniffing mechanisms that allow a supervision taking into account some low and high-level protocols for TCP and UDP network communications.**

Keywords*—*Network Security; Intrusion Detection; Intrusion Prevention; Sniffing; Filtering; Network sniffer; Libpcap; Libnet; Sockets

I. INTRODUCTION AND NOTATIONS

The sniffing is a technique of monitoring every packet that crosses the network. A packet sniffer is a piece of software or hardware that monitors all network traffic. Network analysis is the process of listening and analysis of network traffic. It controls network communications to identify performance problems, locates security vulnerabilities, analyzes the behavior of the application, and performs capacity planning. The management and the supervision of exchanged data by network systems are a fundamental task that contributes to a reliable intrusion detection and analysis of intrusive activities detected. The sniffing tools are used to listen, monitor, capture, record, and analyze network traffic. They extract the necessary information's to make decision and implement the best strategies to improve the computer security. Many sniffers are available to capture packets circulated in wired networks (Ethernet sniffers, for example) and wireless network. They help network managers to assess and review the data over their networks, to indicate the network problems and to identify some failures monitored network [16]. The sniffing is also exploited by attackers to gather a database of information on victims' networks and hosts that constitute them. They can intercept the data and even the users' passwords [28] [31]. The goal of this paper is to describe a new hybrid sniffer for a relevant collection of data. For this, we proceed as follows: the second section cites an art state on the techniques of network sniffing; a related study will be developed on some sniffers. In the third section, the performances assessment is performed on a list of network sniffers based on various parameters to ensure the computer security objectives, mainly the type of used library, libpcap or libnet to establish the performances and limitations of each library and finally validate our choice. A classification is deducted at the end of the section. The proposed network sniffer works in raw, connected and connectionless mode; it will be described in part four. A detailed description of the new model and its characteristics will be discussed in fifth section. This work will be finished by a conclusion and perspectives. Trough this paper we use the following notations:

IP:	Internet Protocol.
IDS:	Intrusion Detection System.
IPS:	Intrusion Prevention System.
ARP:	Address Resolution Protocol.
NIC:	Network Interface Cards.
MAC:	Media Access Control.
CRC:	Cyclic Redundancy Check.
PDU:	Protocol Data Unit.
BSD:	Berkeley Software Distribution.
BPF:	Berkeley Packet Filter.
LSF:	Linux Socket Filter.
FFPF:	Fairly Fast Packet Filter.
CGF:	Control Flow Graph.
ATM:	Asynchronous Transfer Mode.
ISDN:	Integrated Services Digital Network.
FDDI:	Fiber Distributed Data Interface.
RFMON:	Radio Frequency MONitor.
SSID:	Service Set Identifier.
WEP:	Wired Equivalent Privacy.
WAP:	Wireless Application Protocol.
PSK:	Pre-Shared Key.
OSI:	Open Systems Interconnection.
LLC:	Logical Link Control

DOS:	Denial of Service.
SOCK_DGRAM:	Datagram Sockets.
SOCK_STREAM:	Stream Sockets.
SOCK_RAW:	Raw Sockets

II. Related Work

In this section, we discuss the sniffing types, components of the sniffing tools, used libraries to capture network packets and the used methods to filter the captured traffic.

The network monitoring is a difficult and demanding task. It is an essential part in the use of network administrators who are trying to maintain the good operating of their networks and need to monitor the traffic movements and the network performances [21] [29]. The sniffing listens to public conversations in computer networks. It is used by network managers to manage and ensure the network security. It can also be used by unauthorized users. Mostly, this device is placed between the server and the clients web pages, it listens and analyzes all sent and received requests by the server. Sometimes, a network sniffer is called a network monitor or a network analyzer [30]. There are different types of sniffing packets [3] [30]:

- IP sniffing: collects all IP packets traveled through a network corresponding to the IP addresses of supervised entities.
- MAC sniffing: captures the corresponding frames to supervised interfaces MAC addresses.
- ARP sniffing: intercepts the ARP packets used to query the ARP cache during network communication.

Generally, the sniffing is divided into two major classes: passive sniffing that collects raw traffic circulated in the network without treatments, and active sniffing that intercepts and treats the collected traffic [3][28]. The sniffers monitor a wide sent and received information by computer networks. There are many commercial and no commercial tools, hardware and software that enable to intercept packets [30]. The copies of captured packets are stored in a temporary (buffers) or permanent memory (database server). They are analyzed to extract the useful information or specific models (patterns). The amount of captured traffic depends on the location of the controlled host as well a primary server in a computer network intercepts a significant traffic than isolated system client. The sniffers operate in two distinct ways: with filtered way to capture the data containing a specific elements and an unfiltered way to collect all the raw network traffic. Some network topologies such as Ethernet are designed so that all machines connected to a network segment share the same transmission media, thus, the hosts that are connected to the same network segment will be able to see all traffic passing through that segment. Ethernet hardware is designed to filter the traffic passed, it captures the traffic which concerns it or has a broadcast addresses and ignores all other traffic. This is done using the MAC address. To copy all traffic, the host network cards have to be implemented in promiscuous mode [3] [16]. The hardware sniffers use the standard adapter’s NIC, otherwise they may face problems in the CRC error, voltage and cabling problem. The analysis of captured packets is often done in real time. The captured traffic may be submitted to a decoding operation to be descriptive and understandable text for easy interpretation. Sometimes, the sniffers edit the packets and transmit them to the network. The security aspect done by sniffers is represented by their availability to monitor and capture the traffic in and out of the network taking into account the clear text passwords and user names [4]. Besides, the network sniffers participate in detecting and identifying of the intrusions by monitoring the activities of networks and systems [14] [16] [31]. They are constituted by the components described by Clincy & Abi Halaweh in [3] [16]:

- Hardware: is represented by a NIC, activated in sniffing mode.
- Driver: starts the capturing data from the network cards, applies a number of filters on traffic and stores it in a memory.
- Buffer: stores the captured traffic or transfers it to permanent storage.
- Analyzer: is software responsible for analyzing the traffic in real time taking into accounts the criteria and analysis needs.
- Decoder: receives a stream of bits and interprets them to finally build a descriptive texts format.
- Editor: is available in some sniffers, it changes the traffic using a unified format and then converts it and retransmits it in the network.

The sniffers can be used effectively for teaching and learning networking concepts regardless of the technical context. They are presented to understand the model and protocols of network layers [26]. They allow to:

- Examine the format of a protocol data unit (PDU) to each layer in the network model.
- Examine the message exchanges for two TCP or UDP connections.
- Examine the messages transferred between a client application and a server.

The simulation with packet sniffers is used in learning of computer networking, allows a good understanding of network concepts, topologies and explains the functions and the roles of a hub, a bridge or switch and a router. It shows how a data packet is transmitted into LAN and illustrates the encapsulation and decapsulation operations while going through the protocol stack [31]. The main capture libraries are libnet and libpcap [1] [2] [22]:

TABLE I. LIBPCAP AND LIBNET LIBRARIES

	Libpcap	*Libnet*
Capture level	• Captures the packets in low level. • Extracts the packet so raw kernel without treatments.	• Manipulates a high level traffic. • Can manipulate a several low level networking routines.
Used mode	• Conected mode (TCP) • Connectionless mode (UDP)	• Conected mode (TCP) • Connectionless mode (UDP)
Filtering	• Compatible filtering with the BPF filter. • Initializes and configures filters. • Receives the packets using a loop.	• Doses not provide a packet filtering.
Supported protocols	• Supports almost of networking protocols.	• Injects any kind of an IP packet. • Manipulates a network firewall (IP filter, ipfw, ipchains, pf, PktFilter, ...). • Offers the addresses manipulation functions, the ARP cache and routing tables. • Manipulates an IP tunnel (tun BSD / Linux, Universal TUN / TAP device).
Errors management	• Provides the functions to manage errors.	• Provides the functions to manage errors.
Supported platforms	• Supported by all platforms.	• BSD (OpenBSD, FreeBSD, NetBSD, BSD / OS), Linux (Redhat, Debian, Slackware, etc.), MacOS X (Windows NT / 2000 / XP), Solaris, IRIX, HP-UX, Tru64.

The filtering is an essential operation to classify the captured packets using filters according to the needs of capture. When the packets are intercepted, a filtering is applied. The packet that respects the filter is stored. The capture filters are useful to limit the captured packets when concentrated on a specific packet type, the packets that meet the filter criteria are elected [32]. Among the criteria are used to filter a packet, we find: type packet used (IP, TCP, UDP, ICMP, ...), address of input or output interface, address of source or destination of packet, the number of source or destination port of application, …. The filter is a Boolean function which returns true if the traffic is accepted; otherwise, returns false (the traffic is ignored or rejected). For example, to apply the filtering, the operating system use a packet filter like the BSD Packet Filter for Open BSD systems [13] and the LSF filter for Unix platforms. To improve the filtering operation, several filters are implemented; we cite a result of a recent research on filters packets, the rapid filter FFPF (Now Streamline) [15]. Multiple filters can be loaded simultaneously in FFPF. To design a filter, two basic approaches are available: tree model and direct acyclic graph (CGF) model used by Berkeley Packet Filter [13]. The filtering can be classified into two types:

- The static filtering initializes the filter parameters to be applied in advance. It is provided for example by the pcap language [6] maintained and developed by researchers at the Lawrence Berkeley National Laboratory and enables the use of simple rules to remove the unwanted packets.
- The dynamic filtering implements the parameters that change during running. The filter Swift or Fast Dynamic Packet Filter is an example of dynamic filter [12] [27].

III. STUDY AND PERFORMANCES ASSESSMENT

This section will study a performance evaluation of a proposed list of sniffers setting up parameters related to computer security. It is completed by a classification.

A. *Assessment Parameters*

Normally, to realize an assessment performances and classify the various sniffers which use wired and/ or wireless networks, many criteria are available such as, supported platforms, operating systems and interfaces, user interface, number of protocols that the network sniffer can decode, available utilities to enable the user to personalize capturing and displaying network packets, support for customized protocol decodes, readability of captured data, provided statistical information, decoding captured data, …. Our main objective in this work is to propose an approach to improve the security level. So, our study is based on parameters related to computer security that test the sniffers availability and their reliability. It is useful to recall that the sniffing requires the activation of interfaces in promiscuous mode for wired networks and in rfmon mode for wireless networks.

To compare and evaluate the proposed tools, we focus on evaluation characteristics dependent on computer security cited in [16] [28] [33].

- Availability: to test the availability of a sniffer, three parameters are cited:
  - Operating time of the sniffer.
  - Memory size allocated to the implementation of the sniffer (as the size increases, the treatment requires a lot of time).
  - Maximum controlled flow by the sniffer [34]. The table II bellow illustrates the different network technologies with their supported maximum flows.

TABLE II. NETWORK STANDARDS AND MAXIMUM FLOWS

Network	Standard	maximum flow
Wired Networks	Ethernet	Megabits, gigabits
	ISDN	Low flow Services: 64Kbps to 2Mbps
	ATM	High flow Services: 10Mbps to 622Mbps.
	FDDI	100 Mbps
	Token Ring	4 Mbps to 16 Mbps
Wireless Networks	802.11a	54 Mbit / s with a range of 10 m.
	802.11b	11 Mbit / s with a range of 10 0 m.
	802.11g	54 Mbit / s with a range of 100 m.
	802.11n	Frequency 2.4GHz and 5GHz

- Filtering: verifies the existence of a filtering system to filter the traffic.
- Used library: determines the used library by the sniffer to capture traffic: libnet or libpcap.
- Supported protocols: means the number of protocols taken by a sniffer.
- Alert: an alert will be produced, if a problem exists in the controlled segment,
- Real Time: the treatment in real time is a parameter of an effective sniffing [3] [25].

B. Classification of the Sniffers

We refer to the study treated in [3] [4] [5] [6] [7] [8] [9] [10] [11] [16] [17] [18] [19] [25] [26] [28] and [32] and we deduce the classification of sniffers according to the characteristics and proposed parameters:

TABLE III. CLASSIFICATION OF SNIFFERS

Network sniffers	S/ H	Library	Filtering	Flow	Availability	Alert	Real time
Tcpdump	S	Libpcap (Winpcap)	++	Flow of Ethernet networks	Very economical installation file size: 484 KB	- -	- -
Wireshark	S	Libpcap (Winpcap)	++	Flow of Ethernet and wireless networks.	81 MB after installation.	- -	++
PACKETYZER	S	Libpcap (Winpcap)	++	Flow of Ethernet, FDDI, PPP, Token Ring and wireless networks.	-supports 483 protocols. -Decodes and edits packets.	- -	++
Netflow CISCO	S H	Libpcap	++	High flow networks (Gigabit).	Very high (provides valuable information about users, network applications, peak hours). -2GH Dual processor. -2GO Memory.	++	++
Colasoft Capsa	S	Libpcap	++	Flow wired and wireless networks over 802.11a, 802.22b, 802.11g and 802.11n	-No Tolerant with the attacks: ARP, TCP port scanning, -Signals DOS attacks - Free version is available with limited features.	++	++
PRTG Network Monitor	S H	Libpcap	++	High flow	-653 MB on after windows 7 installation. -Integrates SNMP, Packet (Sniffing and Net flow). -monitors 24/7 network. - Includes over 200 types of sensors. -Less than 30 protocols (Free). - More than 30 protocols (Com)	++	++
Kismit	S	Libpcap	++	Flows of wireless networks 802.11n, 802.22b 802.11g and 802.11a	High (supports any wireless card rfmon)	++	++
Scapy	S	Libpcap and Libnet	++	Injectes the 802 frames	- Generates and receives quick and accurate traffic. - Decodes packets of a number of protocols.	++	++
OmniPeek	S H	Libpcap	++	Ethernet, Gigabit, 10 Gigabit, 208.11 a / b / g / n / ac wireless, VoIP, Video, MPLS and VLAN	-captures on multiple networks simultaneously. - Several hundred protocols - WPA, WPA2 and PSK Decoding.	++	++
ETHERAP	S	Libpcap	++	Flows of Ethernet, FDDI, Token Ring, ISDN.	- Is only available for GNU / Linux systems.	- -	++
Soft Perfect Network Protocol Analyzer	S	Libpcap	++	Flows of Ethernet networks	-Analyzes of fragmented floors. -defragments and reassembles the packets. - Size of the installation file 4.87 Mb.	- -	++
Airodump	S	Libpcap	++	- Wireless networks 802.11. - Supports 4.2 GHz channels	-Identification the coordinated access points. -Writes the several files containing details of all seen access points and clients.	- -	++

Com: Commercial license. Free: Free license. ++: Available. - -: Not available.
H: Hardware. S: Software.

C. Discussion of the Results

This section cites the architecture of many sniffers, their characteristics and their operating. We assess their performances based on parameters related to security objectives: authentication, confidentiality, integrity, availability and rapidity. Really, it’s difficult to meet this assessment, because normally the goal of sniffing is not to indicate the problems and attacks but to collect the circulated traffic in the networks and sometimes to inform the state of the monitored network excepting some IDS sniffers that can detect intrusions. The majority of these sniffers use libpcap library to intercept traffic and include a filtering system. They are highly available to monitor wired and wireless networks with a high flows supporting a large number of protocols. The treatments are often in real time and the detected problems are alerted by some sniffers. On the other side, this study helps us to discover certain limitations of those sniffers. They are based on a passive sniffing. Sometimes, they are exploited for unauthorized uses, for example, Airodump that is designed to crack WEP and WPA encryption algorithms; it is used to encrypt traffic on wireless networks. The implementation of software sniffer by interpreted languages such as Python presents a slow in their performance and increases consequently the system. The encrypted and fragmented packets are intercepted by sniffers but they are not analyzed. The hardware sniffers have adaptation and compatibility problems [3]. In the next section, we describe in detail the new model of a network sniffer.

IV. Our Proposal Scheme Pcapsocks

In this section, the proposed model of sniffing is cited. We prove that our proposal takes into account the benefits of a reliable collection of traffic to satisfy the current expectations. It is time to formulate a new proposition of network sniffer. Our model, called PcapSockS, based on pcap language and sockets satisfies. It decodes the intercepted traffic to prepare it for the analysis step and finally built a collection database for automatic intrusion detection. Specifically, it ensures two major tasks:

- Collects the data traffic in high and low level.
- Builds a database for the new proposed IDS/IPS.

The new design focuses on the combination of current performances of high sniffers and minimization of various limitations. Thus, we propose a distributed model consisted by two main components:

- The kernel is composed by two processors, the first to capture the traffic and the second for filtering.
- The operator decodes the elected traffic using the functions and treatments of normalization.

These components are described in the figure1 below:

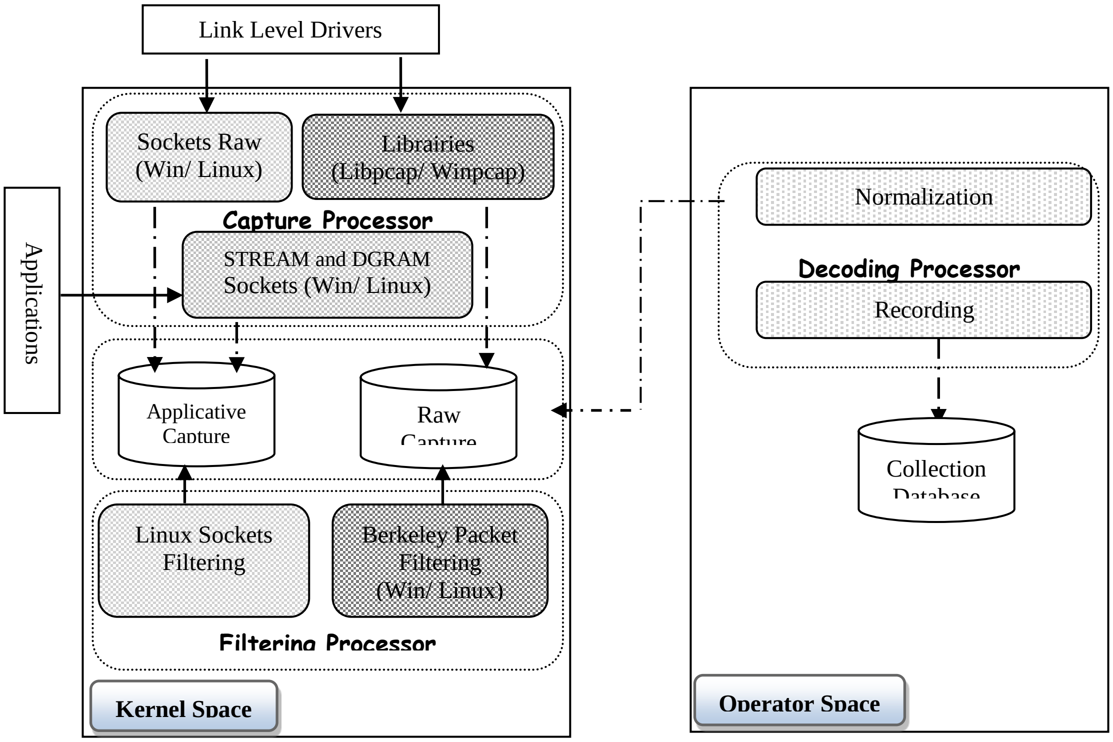


Fig. 1. Model of a network sniffer PcapSockS

The above design can be implemented in the Linux and Windows platforms, for the Berkeley Packet Filtering filter is an extension of Linux Sockets Filter [23]. So, the provided functions by the LSF are taken into account by the PBF in the case of windows. With this new design, we provide an optimal sniffer for capturing, filtering, optimization and decoding traffic while enabling the large satisfaction of various specificities and open the horizons for other works trying to improve the computer security techniques. The figure2 shows the flows exchanged between the various processors:

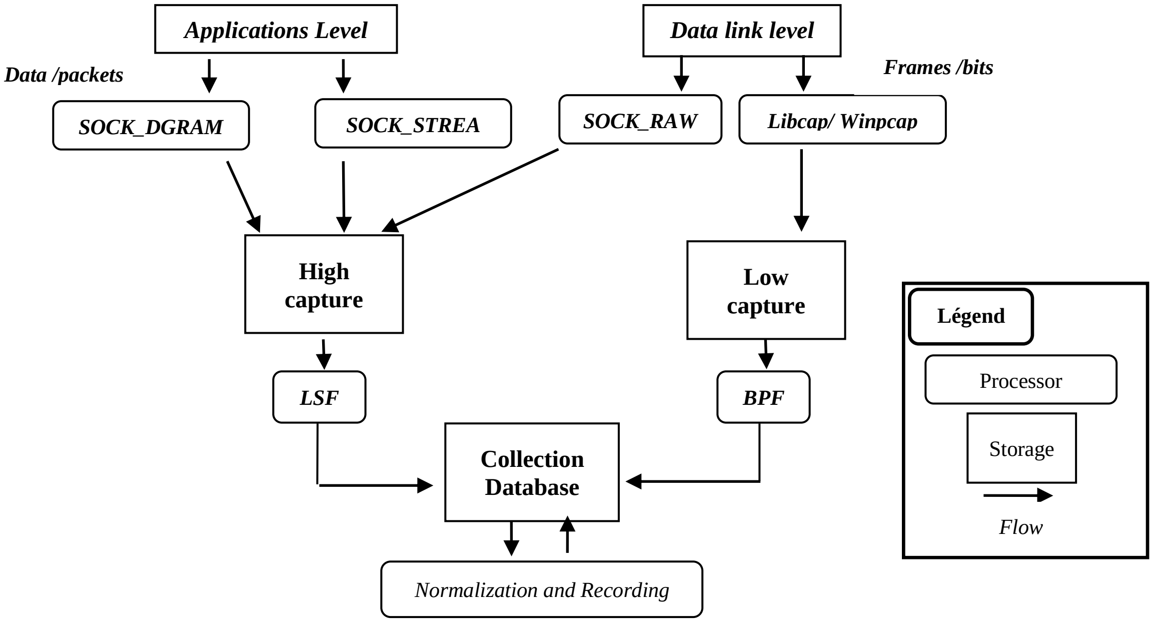


Fig. 2. Data flow diagram

Our new model provides different performances:

- Combining libpcap and sockets functions to capture the packets.
- Filtering traffic taking into account the capture needs.
- All treatments are in real-time.
- Encryption of transactions between the sniffer and Collection database.

The next section details the decoding operations used by the PcapSockS, it shows the performances provided by this new probe by comparing it with the Scapy and Wireshark which are considered the most famous sniffers in the moment.

V. Detailled Description

A. *Processing Operations*

The libpcap library is an open source library written in C that provides a programming interface from which the packets are intercepted [22]. It relies on a low level language, includes the functions that can be associated with the user request and provides a powerful and abstract interface for the capture process [24]. The process used by libpcap is defined by the following figure:

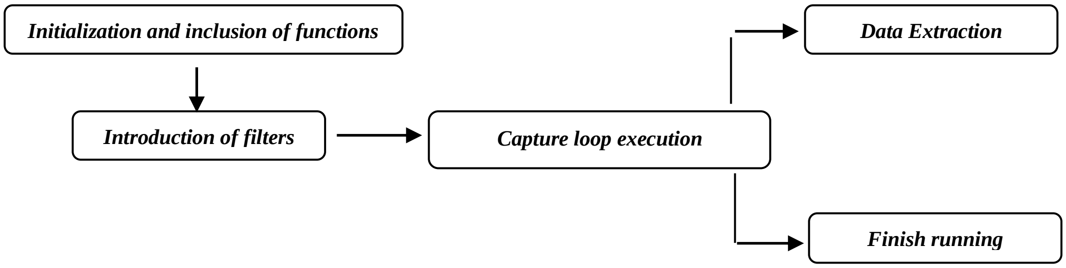


Fig. 3. Capture process provided by libpcap

The Sockets are the objects for sending and receiving messages between processes. They were developed by Berkeley in 1982 as part of the Berkeley version of Unix. The Sockets are the specific original Unix systems; they ensure the communication between various processes, applications and network layers. The main socket types are:

- SOCK_DGRAM: connectionless sockets (UDP messages).
- SOCK_STREAM: connection oriented sockets (TCP packets).
- SOCK_RAW or Raw sockets (frames and bits): The IEEE 802.2 protocol defines the sub layer LLC of the data link layer.

The sockets in connectionless and connection oriented mode are inserted between the layers 3 and 4 of the OSI model. The raw sockets are positioned in layers 1 and 2 [23].

The filtering is an essential process of checking the integrity of the kernel traffic. The copies of the collected traffic can be minimized by deploying a kernel agent called a packet filter that rejects unwanted packets [13]. The traffic can be ignored and blocked using one of the techniques used for the blocking of data [20].

B. Description of Solutions

The PcapSockS Sniffer integrates libpcap to intercept traffic from the low-level, physical and data link layers of the OSI model. This traffic is composed of a set of bits and frames, it's saved in a temporary basis to apply the BPF filter and then meet adequate collection conditions. Libpcap provides the possibility to introduce the filters to filter traffic: PBF, SWIF [12]. It applies the filters on traffic in the basis in order to choose the elected packets. This latter is redirected to the operator space. The decoding processor normalizes and stores the chosen traffic in the collection Database. In the high level, we use the sockets mechanism to ensure a reliable collection. The TCP and UDP sockets are implemented for this purpose. Raw sockets are used to reinforce the interception to the low level with libpcap. The collected traffic is saved in a temporary basis to apply the filter LSF and redirected directly to Collection Database. So, our sniffer collects data in three modes:

- Connection oriented mode requires a prior connection establishment between communicating entities; this connection is defined by a logical relationship between the parts which exchange data.
- Connectionless mode cannot guarantee a reliable connection, insertion errors, wrong delivery, duplication, or non sequencing delivery packets. These faults can be reduced by providing a reliable transmission service to a protocol layer of the highest level.
- Raw mode can provide both services in connection oriented and connectionless mode.

The filtering provides a considerable gain; it avoids the congestion and the saturation of memory. The filtering is a very useful to meet the various network services using mainly in intrusion detection [14] [31]. The PcapSockS Sniffer implements the filtering operations on collected traffic taking into account the parameters and attributes characterizing the monitored entities. The treatments are in real time. Take into account the time constraints which are as important as the accuracy of the results for this system synchronizes multiple tasks that take place and the possibility of including several shorter threads in a single process [25].

To show the performances provided by the PcapSockS Sniffer, it is very useful to compare it with other sniffers which have demonstrated their reliability, we cite Scapy and Wireshark.

TABLE IV. COMPARISON OF PCAPSOCKS WITH SCAPY AND WIRESHARK

Sniffer	Platforms	Low capture	High capture	Low filtering	High filtering	Network
Scapy	-Win -Linux -Mac OS	-Libpcap	-Libnet -Python Functions	-PBFfilter	-No	-Wired -Wireless
Wireshark	-Win -Linux	-Libpcap	-No	-PBF Filter	-No	-Wired -Wireless
Pcap.Sock Sniffer	-Win -Linux	-Libpcap -Raw Sockets	- Sock_Stream -Sock_Dgram	-PBF Filter	-LSF Filter	-Wired

VI. CONCLUSION AND PERSPECTIVES

There are many available tools used to capture network traffic that researchers use in their work, but there is a limitation in their functions. Some tools capture network traffic only without analysis. Therefore, the researchers have to use another tool for analysis to get the traffic feature like it is need of his work. In this article, we studied in detail the discipline of sniffing which is an interesting task but is difficult to put in place taking into account the various needs. The sniffing enables improved security of computer networks and systems that compose them.

This study provides a list of popular sniffers to evaluate and to deduct the existed limits. Thus, a classification is provided based on the parameters cited in the second section, related to computer security: availability, traffic filtering, real time, used library and flow.

We propose this software for the data collection part of our new intrusion prevention system approach based on neural network. So, we describe in detail its objectives, roles of its various components and nature of modes used for sniffing. Our future work is to implement and validate the steps of PcapSockS Sniffer and integrate this sensor in IDS/IPS.

REFERENCES

[1] Yan Grunenberger, Thesis "Réseaux sans fil de nouvelle génération : architectures spontanées et optimisations inter-couches", 7 Jun 2010.

[2] http ://libdnet.sourceforge.net/.

[3] ms.sonali, a.karale, ms.punam, p.harkut, "packet sniffing" international journal of pure and applied research in engineering and technology IJPRET, 2014; Volume 2 (9): 654-661.

[4] Pallavi Asrodia and Hemlata Patel, "Analysis of Various Packet Sniffing Tools for Network Monitoring and Analysis", Department of Computer Science and Engineering, Jawaharlal Institute of Technology, Borawan, Khargone, (M.P.), International Journal of Electrical, Electronics and Computer Engineering 1(1): 55-58(2012).

[5] All about Wireshark [Online] Available http://www.wireshark.org/.

[6] All about Tcpdump [Online] Available http:// www.tcpdump.org/.

[7] http://www.colasoft.com/capsa/.

[8] https://www.kismetwireless.net/.

[9] http://www.paessler.com/prtg.

[10] http://www.secdev.org/projects/scapy/.

[11] David Rideau, "Outils de collecte pour réseaux gigabits Une alternative à la technologie Cisco Netflow", Département Réseau du CICG (Centre Inter-Universitaire de Calcul de Grenoble).

[12] Zhenyu Wu, Mengjun Xie, Member, IEEE, and Haining Wang, Senior Member, IEEE "Design and Implementation of a Fast Dynamic Packet Filter", IEEE/ACM TRANSACTIONS ON NETWORKING, 2011.

[13] S. McCanne and V. Jacobson, "The BSD packet filter: A new architecture for user-level packet capture," in Proc. Winter USENIX Tech. Conf., 1993, pp. 259–269.

[14] V. Paxson, "Bro: A system for detecting network intruders in Real-Time", vol. 31, no. 23–24, pp. 2435–2463, Dec. 1999.

[15] H.Bos, W.de Bruijn, M. Cristea, T. Nguyen, and G. Portokalidis,"FFPF: Fairly fast packet filters," in Proc. USENIX OSDI, 2004, pp. 347–363.

[16] Clincy; Abu Halaweh. "A taxonomy of free network snifers for teaching and research", Journal of Computing Sciences in Colleges, Volume 21 , Issue 1, pp 64-75, 2005.

[17] All about soft perfect network protocol analyzer [Online] Available http://www.softperfect.com/products/networksniffer/

[18] http://etherape.sourceforge.net/

[19] http://www.aircrackng.org/doku.php?id=airodump-ng

[20] B.Suneel Kumar, S.V.V.D Venu Gopal, M.Satish Kumar, "Blocking Technique of Dataflow in Networks", ASR Engineering College, Tanuku, W.G Dist, and Andhra Pradesh.

[21] AlishaCecil, "A Summary of Network Traffic Monitoring and AnalysisTechniques", acecil19@yahoo.com

[22] Alejandro L´opez Monge , "Aprendiendo a programar con Libpcap", kodemonk@emasterminds.net, 20 de Febrero de 2005

[23] Christophe Gimenez,"MiniSniff Application de capture de trames,V.A.E Algorithmique N1/N2 – Réseaux", DESS C.C.I. 2003-2004

[24] Fulvio Risso and Loris "An Architecture for High Performance Network Analysis", Degioanni Dipartimento di Automatica e Informatica – Politecnico di Torino Corso Duca degli Abruzzi, 24 – 10129 Torino, Italy

[25] Manas Saksena, "Conception de logiciel en temps réel – Progrès actuels et défis à venir", Université Concordia, et Bran Selic, ObjecTime Limited

[26] Bruce P. Tis , "Using packet sniffing to teach networking concepts", Simmons College, Boston Ma Journal of Computing Sciences in Colleges, Volume 30 Issue 6, June 2015, Pages 67-74

[27] Zhenyu Wu, Mengjun Xie and Haining Wang "Swift: A Fast Dynamic Packet Filter", The College of William and Mary, NSDI '08:5th USENIX Symposium on Networked Systems Design and Implementation

[28] Dr. Charu Gandhi, Gaurav Suri, Rishi P. Golyan3, Pupul Saxena, Bhavya K. Saxena, "A Packet Sniffer – A comparative study", VOL. 2, NO. 5, MAY 2014, 179–187 Available online at: www.ijcncs.org ISSN 2308-9830

[29] SB .A. Mohammed, Dr.S.M Sani, Dr. D.D. DAJAB, "Network Traffic Analysis: A Case Study of ABU Network", Computer Engineering and Intelligent Systems, ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online), Vol.4, No.4, 2013

[30] Rupam, Atul Verma, Ankita Singh, "An Approach to Detect Packets Using Packet Sniffing", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.4, No.3, June 2013

[31] X.Yuan, P.Vega, Jinsheng Xu, Huiming Yu, and Stephen Providence "An animated simulator for packet sniffer", , Department of Computer Science, North Carolina A&T State University, 1601 East Market St., Greensboro, NC 27411

[32] L.Chappell, "Wirehark Network Analysis",The Official Wireshark Certified Network Analyst™ Study Guide Second Edition San (Version 2.1b).

[33] Y.Farhaoui, A.Asimi, "Performance method of assessment of the intrusion detection and preventionsystems", International Journal of Engineering Science and Technology (IJEST), ISSN : 0975-5462, Vol. 3 No. 7 July 2011.

[34] Claude Duvallet, "Les réseaux informatiques", Université du Havre UFR Sciences et Techniques 25 rue Philippe Lebon - BP 540 76058 LE HAVRE CEDEX Claude.Duvallet@gmail.com.

Choosing a Career Based Personality Matching: A Case Study of King Abdulaziz University

Nahla Aljojo
Faculty of Computing and Information Technology, Information Systems Department,
King Abdulaziz University, P.O. Box, 132139, Jeddah, 21382, Saudi Arabia

***Abstract*—Traditionally, selecting a career involved matching the specific aptitudes and characteristics of an individual with a career which required or involved such factors. This particular approach has as its foundation the fact that certain careers have need of individuals with certain skills and attitudes, that is to say the individual is a 'fit' for that particular career based on their knowledge, skill and disposition. Many students have problems determining their college majors and a suitable career. This paper shows how to find a career that fits one's personality, and aims to help students analyse their personalities based on the Holland personality test. This paper will identify the most suitable career for students by using a validated Arabic version of the Holland survey, which is one of the most popular models used for career personality tests. This study implements a new set of tasks, testing 117 students from King Abdulaziz University with the Arabic version of the Holland test. The test was applied to female students from three majors, computer science, information systems, and preparatory year distance learning students. The implications of the test results will help students to understand their personality types and determine a suitable career, as the results of the test suggest suitable careers for students which match their personalities. Ultimately, the results show a difference between computer science, information systems and preparatory year distance learning students with regard to personality types and suitable careers.**

Keywords—Holland's Theory of Vocational Personalities; RIASEC personality and environment types; occupational interests; Career change; hexagonal model

I. Introduction

Historically, the selection of an occupation or career might have been considered as less difficult and more straightforward many years ago, before the industrial revolution. Often an individual's career was simply determined by following in their father's footsteps, that is to say a carpenter's son would become a carpenter, and a cobbler's son would learn to be a cobbler like his father, while the nobility and landed gentry groomed their offspring to become future leaders. Over time, and especially since the industrial revolution, this process of young people choosing their occupation or career has changed dramatically and become far more complex. Wider access to higher levels of education, women entering the workforce and the swing from industrialization to service-based industries in many developed economies have all had an impact on how an individual chooses their occupation or career. In this post-industrialization period, where a higher level of skills is generally required due to the emergence of knowledge working and knowledge based industries, individuals might have greater choice than ever before regarding their careers, but the extensive opportunities when choosing the right career also present challenges. With this in mind, research which allows students to make informed decisions regarding their professional lives, and therefore their future personal lives (Zaidi et al., 2012) is essential in order to ensure that students fully understand this complex process and how it might impact their lives in the future.

This paper has been designed to find a career that fits one's personality. The Holland test was developed as a technique to help determine a suitable career according to the six aforementioned personality types (Holland`s theory). Furthermore, this paper compares personality types, and suitable careers that fit these personality types, of students from computer science (CS), information systems (IS), and first year distance learning students within Saudi Arabia.

The author chose female students from different departments at the Faculty of Computing and Information Technology, King Abdulaziz University. These students came from the departments of CS, IS, and preparatory year of distance learning, in order to determine suitable careers for them according to the six personality types. The author subjected the students to a test consisting of 106 questions, and then calculated and analysed the results by using SPSS. The results showed that the optimal solution was the Holland test, used by normal students to help them better understand their personalities. The advantages of this solution are the time and effort saved, its ease of use, and that it can be understood and applied anytime and anywhere

II. Literature Review

The aim of recruitment is to select the best possible applicant who has the capacities that are needed for the job and who will fit in well with the organization (Rynes & Gerhart, 1990). The most common theory used in employee selection processes is perhaps the theory of fit (Sekiguchi, 2004). During the past century, models of fit or congruence have achieved a significant role in the field of industrial and organizational psychology and human resources management (Saks & Ashforth, 1997; Schneider, 1987, 2001; Holland, 1997; Kristof, 1996; Pervin, 1968; Ekehammer, 1974; Lewin, 1935; Murray, 1938; Parsons, 1909). Employee selection processes have especially focused on achieving person-job fit (Werbel & Gilliland, 1999) which is the congruence between the abilities of a person and the demands of a job (Edwards, 1991; Kristof, 1996). During the past decade or so, several authors have recognized that the practitioner involved in personnel selection and those involved in scientific studies of this discipline have diverged and are moving further and further away from each other (e.g. Anderson, Herriot & Hodgkinson, 2001; Dunnette, 1990; Hodgkinson, Herriot & Anderson, 2001; Sackett, 1994). One example of this trend is that the American Uniform Guidelines on Employee Selection Procedures (1978), Harvey (1991) and Harvey and Wilson (2000) recommend that the traits and abilities of workers should be left out of selection processes. According to them, personal traits do not meet the requirements of verifiable and replicable job analysis data.

There are various factors which drive the decision to choose any particular career. Davidson (2010) identified environment as being of importance, as well as remuneration, proximity to family, and personality. Earlier, Borchert (2002) in a study of students at high school highlighted the factors of personality, opportunities, and environment as being the fundamental drivers of decision making regarding a career. However, the overriding factor determining career selection was in fact the personal desire and disposition of a student to select a particular career.

The view that individuals should select a career in line with their personality was also supported by Ferguson (2000) who listed six basic vocational interests which included social, investigative, realistic, enterprising, artistic and conventional (SIREAC types) referred to as Holland Typology. Ferguson claimed that individuals can typically be classified as one of these personality types. His recommendation was that an individual should consider careers which provided a good match in terms of the environment offered and their own personality type. His conclusion was that a greater level of congruence between personality traits, personal interest and the environment offered by a particular career would lead to a higher level of professional and personal satisfaction.

Parental profession affects the choice of career, as it is the closest influence an individual normally has. High school coursework, higher education and vocational training opportunities also influence career decisions. But most of all it is an individual's personality which plays the most important role when choosing a career. Students tend to opt for careers which are similar to their personalities (Schreiner, 2010). Gioia (2010) explained the reasons why people make bad career choices. She discussed various reasons why people selected unsuitable careers, such as parental expectations, peer pressure, uninformed decision making and poor self-image.

III. Holland Occupational Themes (Riasec)

Holland's (1985) theory of jobs describes a person's personality as being one of six main types – realistic, conventional, enterprising, investigative, social, and artistic. Holland suggests that most individuals belong to one of six main types. Along with these types, an individual has a singular identity, which might reflect interests, values, abilities, and/or fantasies (Miller, 1994). Holland expounded his hypothesis and focused around his experience as an academic and instructor. His experience allowed him to decide how people might be re-grouped into different types focused on their profession. This methodology created a relationship between different working environments, singular identities, and the career decision process and its advancement (John, 1997).

Holland's codes, both individual and environmental, are communicated in three-letter codes. A three-letter code is created by selecting the three main types among the most suitable categories for the selected person from Holland's six main types.

TABLE I. Holland's Personality Styles (Source: Adapted from Cruickshank (2005))

Type	Personality
Realistic(Adventuring/ Producing)	These people prefer manual occupations which require working with their hands, tools, machines, and technology, and they have a narrow scope of interests with a closed system of beliefs and values. In troubleshooting and problem solving, they prefer practical and structured solutions. Preferred vocation: automotive engineer , Boiler maker , Electrician, and Farmer
Investigative (Analytic)	These people prefer occupations which work with ideas, examination, watching, understanding, and controlling processes. This type of person does not enjoy social and business activities. In troubleshooting and problem solving, they depend on thinking, collecting data, and making careful analyses. Preferred vocation: computer operator, laboratory technician and mathematics teacher
Artistic (Creative)	These people prefer occupations requiring activities which are ambiguous, unsystematic, disordered, or use materials to produce creative art. This type of person does not enjoy systematic, orderly, or monotonous activities. In troubleshooting and problem solving, they demonstrate inventive, creative, and artistic competencies Preferred vocation: actor/actress, artist, interior decorator, photographer
Social (Helping)	These people prefer activities which control others in an attempt to cure, teach, develop, or train them. This type of person does not enjoy systematic, orderly, or monotonous activities. In troubleshooting and problem solving, they show dominant human interaction and social competencies. Preferred vocation: funeral director , librarian, minister/priest, social sciences teacher
Enterprising (Influencing)	These people prefer activities which control others in an attempt to reach organisational goals or achieve economic gain. This type of person does not enjoy scientific or intellectual tasks. Preferred vocation: contractor, lawyer, radio/TV announcer, real estate, sales person
Conventional (Organising)	These people prefer occupations which deal with data, ordering written or numerical data, organising things, or performing systematic activities. This type of person does not enjoy artistic, ambiguous, or exploratory tasks. In troubleshooting and problem solving, they follow established rules and look for advice or counsel. Preferred vocation: bookkeeper, key punch operator, post office clerk, typist

The individual's work nature is based on the three-letter code and gives a short summary of what the person is good at by explaining the level of similarity to the three words together. For example, the three-letter code 'CER' indicates that the individual has an overwhelming conventional identity and characteristics of enterprise and realism (Miller, 1994 (Table 1 summaries the six Holland types, according to Cruickshank (2005)).

Holland developed a hexagonal model in order to show connections linking personality types as well as define the notions of differentiation and consistency. The former is considered as the degree to which a personality pattern can be described. The latter is considered as the extent of association which exists between different personality types in an individual. An example of this would be personality pattern RI, which has greater consistency than CA. Additional high consistency patterns also include RC, IA, AS and SE. Medium consistency patterns include IS, IC, AR, and SC, while low consistency patterns include RS, EI, AC and SR. Patterns which are easy to distinguish are those which bear a high degree of resemblance to one single personality type (Kelso et al., 1986). Those which are not so easily distinguishable or more difficult to classify might bear equal resemblance to a number of the six types of personality. Based on this, Holland claimed that the main personality characteristic of a person, corresponding to their type, was the most important driver in terms of choosing their vocation. Holland's theory extended to claiming that personality types flourish in a congruent environment, that is to say a congruent environment offers prospects and incentives which fit well with a person's own likes, interests and skills, in other words, a Realistic type in a Realistic environment (Holland, 1985).

The notion of forecasting results based on the matching of personality type with environment was based on the fact that personality types and environments had a mutual range of factors which made it possible to forecast the outcome of a certain type of person in a certain type of environment (Holland, 1985, p.34). It was also suggested (Holland, 1985, p.35) that placing a specific personality type within an appropriate and matched environment would lead to a range of beneficial outcomes, for example, a higher degree of satisfaction at work, greater accomplishments, as well as vocational longevity.

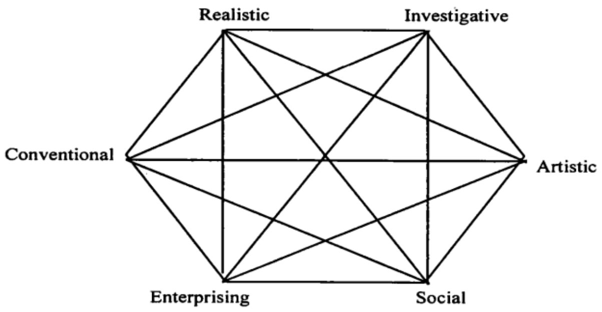


Fig. 1. Holland's Hexagonal Model for defining the psychological resemblances among personality types, environments and their interaction (Source: Holland 1985, p.29)

IV. Methodology

A. Participation

The participants in this study were 117 women, aged 20–40 years old. All participants were students from King Abdulaziz University, with 28 CS students, 27 IS students (Faculty of Computing and Information Technology), and62 preparatory year distance learning students.

TABLE II. RIASEC QUESTIONS

A	1	4	14	17	32	33	35	40	47	49	58	68	80	81
	90	91	102	105										
C	2	12	24	27	38	39	56	57	70	71	79	83	86	87
	92	103												
I	13	16	18	19	25	29	43	44	48	50	52	60	66	73
	76	77	93	98										
E	9	11	20	21	30	34	41	51	53	55	61	69	74	75
	89	95	97	106										
R	3	8	10	15	26	31	36	37	42	45	54	65	82	85
	96	99	100	101										
S	5	6	7	22	23	28	46	59	62	63	64	67	72	78
	84	88	94	104										

B. Intrsuments

The Holland test consists of 106 questions, each question requiring a simple yes or no response. Each of the sets of questions corresponds to one of the six personality types discussed above – Realistic (R), Investigative (I), Artistic (A), Social (S), and Enterprising (E). Conventional (C) includes only 16 questions. The division of questions can be seen below in Table II. For all questions, the students were given two possible answers - either yes or no - for each question.

The three highest categories represent the areas which have a high level of interest for a student and are compatible with a student's Holland Code. An example of this would be a student whose highest score is in the Social category, their second highest score is in the Artistic category, and their third highest score is in the Enterprising category: which would result in their Holland Code being SAE, indicating that they should focus their career choice in occupational areas related to this.

C. Procedure

A validated Arabic version of the Holland survey was uploaded to blackboard (Learning Management System) for students of preparatory year distance learning. Students then sent the completed Holland survey back by email. In the case of students from CS and IS, a validated Arabic version of the Holland survey was distributed manually to them and then collected manually once they had completed it.

V. Results

A. Descriptive Statistics

As seen in Table III, Descriptive statistics of the participants (students of CS, IS and preparatory year distance learning), the results according to personality styles were as follows:

- Social was the most frequent, representing 19% of the overall personality styles of students
- Enterprising was assigned the second rank, representing 16.8% of the overall personality styles of students
- Investigative was assigned the third rank, representing 16.4% of the overall personality styles of students
- Conventional was assigned the fourth rank, representing 16.2% of the overall personality styles of students
- Artistic was assigned the fifth rank, representing 16.1% of the personality styles of students
- Realistic was the lowest, representing 15.3% of the overall personality styles of students
- CES was the most frequent of the three highest, representing 44.4% of the overall personality styles of students
- ESA was assigned the second rank, representing 41.0% of the overall personality styles of students
- SAI was assigned the third rank, representing 32.5% of the overall personality styles of students
- IRC was assigned the fourth rank, representing 16.2% of the overall personality styles of students
- AIR was assigned the fifth rank of the three highest, representing 15.4% of the personality styles of students
- RCE was the least frequent, representing 12.8% of the overall personality styles of students

TABLE III. Descriptive Statistics of the Participants (Students of CS, IS and Preparatory Yeardistance Learning) According to Personality Styles

Personality Styles	Mean	Std. Deviation	Percent	N
S	13.34	2.72	19.12	117
E	11.76	3.23	16.84	
I	11.45	3.14	16.40	
C	11.32	2.76	16.21	
A	11.26	2.77	16.12	
R	10.70	2.69	15.32	
AIR	0.15	0.36	15.4	
ESA	0.41	0.49	41.0	
CES	0.44	0.50	44.4	
SAI	0.32	0.47	32.5	
RCE	0.13	0.34	12.8	
IRC	0.16	0.37	16.2	

As seen in Table IV, Descriptive statistics of the CS participants according to their personality styles, the results were as follows:

- CES was the most frequent, representing 48.1% of the overall personality styles of CS students
- SAI was assigned the second rank, representing 33.3% of the overall personality styles of CS students
- ESA was assigned the third rank, representing 29.5% of the overall personality styles of CS students
- IRC was assigned the fourth rank, representing 22.2% of the overall personality styles of CS students
- AIR was assigned the fifth rank, representing 14.8% of the personality styles of CS students
- RCE was the least frequent, representing 11.1% of the overall personality styles of CS students

TABLE IV. Descriptive Statistics of the CS Participants According to Their Personality Styles

Personality Styles (CS students)	Mean	Std. Deviation	Percent	N
CES	.48	.51	48.1	27
SAI	.33	.48	33.3	
ESA	.30	.47	29.6	
IRC	.22	.42	22.2	
AIR	.15	.36	14.8	
RCE	.11	.32	11.1	

As seen in Table V, Descriptive statistics of the IS participants according to their personality styles, the results were as follows:

- CES was the most frequent, representing 28.6% of the overall personality styles of IS students
- SAI, ESA and AIR were assigned the second rank, representing 25.0% of the overall personality styles of IS students
- IRC and RCE were the least frequent, representing 14.3% of the overall personality styles of IS students

TABLE V. Descriptive Statistics of the IS Participants According to Their Personality Styles

Personality Styles (IS students)	Mean	Std. Deviation	Percent	N
CES	.29	.46	28.6	28
SAI	.25	.44	25.0	
ESA	.25	.44	25.0	
AIR	.25	.44	25.0	
IRC	.14	.36	14.3	
RCE	.14	.36	14.3	

As seen in Table VI, Descriptive statistics of the preparatory year distance learning participants according to their personality styles, the results were as follows:

- ESA was the most frequent, representing 53.2% of the overall personality styles of distance learning students

- CES was assigned the second rank, representing 50.0% of the overall personality styles of distance learning students
- SAI was assigned the third rank, representing 35.5% of the overall personality styles of distance learning students
- IRC was assigned the fourth rank, representing 14.5% of the overall personality styles of distance learning students
- RCE was assigned the fifth rank, representing 12.9% of the personality styles of distance learning students
- AIR was the least frequent, representing 11.3% of the overall personality styles of distance learning students

TABLE VI. DESCRIPTIVE STATISTICS OF THE PREPARATORY YEAR OF DISTANCE LEARNING PARTICIPANTS ACCORDING TO THEIR PERSONALITY STYLES

Personality Styles (first year of Distance learning students)	**Mean**	**Std. Deviation**	**Percent**	**N**
ESA	.53	.50	53.2	62
CES	.50	.50	50.0	
SAI	.35	.48	35.5	
IRC	.15	.36	14.5	
RCE	.13	.34	12.9	
AIR	.11	.32	11.3	

B. *Correlation Analysis*

As seen in Table VII, inter-correlations among Holland's personality styles are as follows:

- Realistic which correlated positively with Artistic, Social, Enterprising and Investigative
- Artistic which correlated positively with Realistic, Conventional, Social, Enterprising and Investigative
- Conventional which correlated positively with Artistic, Social, Enterprising and Investigative
- Social which correlated positively with Realistic, Artistic, Conventional, Enterprising and Investigative
- Enterprising which correlated positively with Realistic, Artistic, Conventional, Social and Investigative
- Investigative which correlated positively with Realistic, Artistic, Conventional, Social and Enterprising

TABLE VII. INTER-CORRELATIONS AMONG HOLLAND'S PERSONALITY STYLES

Personality Styles	**R**	**A**	**C**	**S**	**E**	**I**
R	1	**.380****	.181	**.285****	**.307****	**.426****
A	.318**	1	**.318****	**.449****	**.487****	**.481****
C	.181	.318**	1	**.584****	**.523****	**.582****
S	.285**	.449**	.584**	1	**.748****	**.474****
E	.307**	.487**	.523**	.748**	1	**.568****
I	.426**	.481**	.582**	.474**	.568**	1

**Correlation is significant at the 0.01 level (2-tailed)

As seen in Table VIII, inter-correlations among the three highest of Holland's personality styles of students were as follows:

- AIR (Artistic, Investigative and Realistic) which correlated negatively with CES (Conventional, Enterprising and Social)
- ESA (Enterprising, Social and Artistic) which correlated negatively with IRC (Investigative, Realistic and Conventional)
- SAI (Social, Enterprising and Investigative)which correlated negatively with RCE (Realistic, Conventional and Enterprising)

TABLE VIII. INTER-CORRELATIONS AMONG HOLLAND'S PERSONALITY STYLES

Personality Styles	**AIR**	**ESA**	**CES**	**SAI**	**RCE**	**IRC**
AIR	1	-.163	**-.334-****	.109	.049	-.059
ESA	-.163	1	-.117	.052	-.060	**-.320-****
CES	**-.334-****	-.117	1	-.069	.069	-.161
SAI	.109	.052	-.069	1	**-.211-***	.041
RCE	.049	-.060	.069	**-.211-***	1	-.100
IRC	-.059	**-.320-****	-.161	.041	-.100	1

**Correlation is significant at the 0.01 level (2-tailed)

*Correlation is significant at the 0.05 level (2-tailed)

C. *Comparison of personality styles (percentage) between Information systems (IS), Computer science (CS) and preparatory yearof distance learning students*

Comparison of student personality styles (percentage) results from the Holland survey are shown in Table IX and Figure 2.

Based on the percentage of each personality style exhibited, it was found that students of CS and IS were different from preparatory year distance learning students in their personality styles (ESA). However, CSIS and preparatory year distance learning students were similar in terms of percentage of other personality styles represented (CES, SAI, IRC, RCE, and AIR).

TABLE IX. COMPARISON OF STUDENT PERSONALITY STYLES (PERCENTAGE)

Groups	**Personality Styles**					
	ESA	**CES**	**SAI**	**IRC**	**RCE**	**AIR**
CS Students	30%	48%	33%	22%	11%	15%
IS Students	25%	29%	25%	14%	14%	25%
Distance learning Students	53%	50%	36%	15%	13%	11%

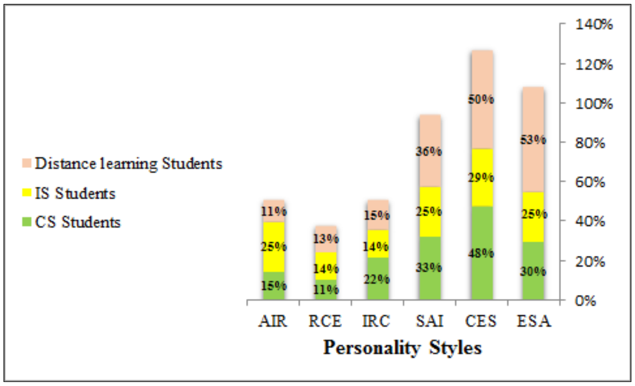


Fig. 2. Comparison of student personality styles (percentage)

Based on ANOVA, analyses were performed on the Holland survey with regard to differences in the six broad personality types between IS,CS and preparatory year distance learning students (Table X). IS,CS and distance learning students differed significantly on ESA factors (F (2, 114) = 4.32, P = 0.02< 0.05),indicating that IS, CS and preparatory year distance learning students differed in the ESA style, but were similar in the other broad personality categories AIR, CES, SAI, RCE and IRC.

TABLE X. COMPRESSION BETWEEN THREE GROUPS (ANOVA)

Personality Styles		Sum of Squares	df	Mean Square	F	Sig.
AIR	Between Groups	.36	2.00	.18	1.39	.25
	Within Groups	14.87	114.00	.13		
	Total	15.23	116.00			
ESA	Between Groups	1.99	2.00	1.00	4.32	**.02**
	Within Groups	26.32	114.00	.23		
	Total	28.31	116.00			
CES	Between Groups	.93	2.00	.47	1.90	.15
	Within Groups	27.96	114.00	.25		
	Total	28.89	116.00			
SAI	Between Groups	.21	2.00	.11	.48	.62
	Within Groups	25.44	114.00	.22		
	Total	25.66	116.00			
RCE	Between Groups	.01	2.00	.01	.06	.94
	Within Groups	13.06	114.00	.11		
	Total	13.08	116.00			
IRC	Between Groups	.13	2.00	.06	.45	.64
	Within Groups	15.79	114.00	.14		
	Total	15.91	116.00			

VI. DISCUSSION

This study deals with many important issues regarding the results of the Holland test and can help students to identify the career cluster(s) in which they would have the most interest and satisfaction. In discussing the results cited from Table3to Table 9, the following themes were evident:

- The three most frequent Holland codes for students (IS, CS and preparatory year distance learning students) were CES (Conventional, Enterprising and social). RCE was the least frequent of the overall personality styles of students. Suitable careers for students with high scores in code CES: Cost Accountant, Congressional-District Aide.

- The three most frequent Holland codes for CS students were CES (Conventional, Enterprising and Social). RCE was the least frequent of the overall personality styles of CS students. The three highest frequency Holland codes for IS students were similar to the three highest of CS students and three lowest frequency Holland codes for CS students were similar to the lowest frequency Holland codes for IS students. Suitable careers for CS and IS students with high scores in code CES: Cost Accountant, Congressional-District Aide.

- The three most frequent Holland codes for distance learning students were ESA (Enterprising, Social and Artistic. Suitable careers for preparatory year distance learning students: Customer Service Manager, Entrepreneur, Foreign Service Officer, Politician, Sales Manager, Advertising Executive, Branch Manager, Buyer, Social Service Director. AIR (Artistic, Investigative and Realistic) was the least frequent of the overall personality styles of distance learning students.

- It was noted from the inter-correlations among Holland's personality styles that Realistic correlated positively with Artistic, Social, Enterprising and Investigative, but not with Conventional. The results have several significant implications for students, for example, the career of Realistic students suitable for Artistic, Social, and Enterprising students.

- It was also noted from inter-correlations among Holland's personality styles that AIR (Artistic, Investigative and Realistic) correlated negatively with CES (Conventional, Enterprising and Social), ESA (Enterprising, Social and Artistic) correlated negatively with IRC (Investigative, Realistic and Conventional), and SAI (Social, Enterprising and Investigative) correlated negatively with RCE (Realistic, Conventional and Enterprising). The results have several significant implications for students. A negative correlation indicates an inverse relationship whereas AIR (Artistic, Investigative and Realistic)styles increase, the CES and ESA styles decrease. This means that the career of AIR students is not suitable for CES and ESA students.

- The personality type categories between IS, CS and preparatory year distance learning students were different in ESA style from the ANOVA test, but similar in the other broad personality types AIR, CES, SAI, RCE and IRC.

VII. Conclusion

This paper has discussed the Holland test in relation to observations indicating that some students have difficulty in determining a suitable career. As this affects their performance, productivity and satisfaction, it is critically important to understand how to find a career that fits their personality. The results may be useful for evaluating their problem in order to determine a suitable career according to the six personality types of Holland's theory. This paper also compared personality types of female IS students with the personality types of CS students and distance learning students in Saudi Arabia. It was noted from the results the personality type categories between IS, CS and preparatory year distance learning students were different in ESA style, but similar in the other broad personality types AIR, CES, SAI, RCE and IRC.

References

[1] Anderson, N., Herriot, P. &Hodgkinson, G. P. (2001). The practitioner-researcher divide industrial, work and organizational (IWO) psychology: Where we are now, and where do we go from here? Journal of Occupational and Organizational Psychology, Vol. 74 No 4, pp. 391-411.

[2] Cruickshank, C. S. (2005). A Test of Holland's Congruence Assumption Using Four-Year Public College Students in Ohio. Ohio: The University of Toledo.

[3] Davidson M (2010). Factors Affecting Career Choices. Retrieved January 20, 2011, from www.ehow.com: http://www.ehow.com/list_6685028_factors-affecting-careerchoices.html

[4] Dunnette, M. D. (1990). Blending the science and practice of industrial and organizational psychology: Where are we and where are we going? In M. D. Dunnette and L. M. Hough (Eds.) Handbook of industrial and organizational psychology (Vol. 1, pp.127). Palo Alto, CA: Consulting Psychologists Press.

[5] Edwards, J.R. (1991). Person-job fit: a conceptual integration, literature review and methodological critique. International Review of Industrial/Organizational Psychology, Vol. 6, pp. 283-357.

[6] Ekehammer, B. (1974). Interactionism in personality from a historical perspective. Psychological Bulletin, Vol. 81, pp. 1026-1048.

[7] Equal Employment Opportunity Commission (1978). The Office of Personnel Management, U.S. Department of Justice and U.S. Department of Labor (1979). Uniform Guidelines on Employee Selection Procedures.41 CFR Part 603 (1978).

[8] Ferguson, I. (2000). Holland's Theory Discussed. Retrieved September 28, 2010, from suite101:http://www.suite101.com/article.cfm/foundations_of_psychology/48558/1

[9] Gioia, C. (2010). Why people make bad career choices. Retrieved September 27, 2010, from http://www.helium.com: http://www.helium.com/items/1903238-why-people-make-bad-careerchoices

[10] Harvey, R. J. (1991).Job analysis.In M. D. Dunnette& L. M. Hough (Eds.), Handbook of industrial and organizational psychology (2nd ed., Vol. 2, pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.

[11] Harvey, R. J. & Wilson, M. A. (2000). Yes Virginia, there is an objective reality in job analysis. Journal of Organizational Behavior, Vol. 21, pp. 829-854.

[12] Holland, J. L. (1973).Making vocational choices: A theory of careers. Englewood Cliffs, NJ: Prentice-Hall. Johansson, C.B. & Campbell, D.P. (1971).Stability of the SVIB for men. Journal of Applied Psychology, 55, 34-36.

[13] Holland, J. L. (1985). Vocational Preference Inventory manual (1985 ed.). Odessa, FL: Psychological Assessment Resources, Inc.

[14] Holland, J.L. (1997). Making vocational choices (Third Edition); a theory of vocational personalities and work environments. Odessa, FL: Psychological Assessment Resources, Inc.

[15] Kelso, S. R., Ganong, A. H. and Brown, T. (1986).HebbianSynapses in the Hippocampus. proc. natl. acad. sci. usa 83: 5326-5330.

[16] Kristof, A. (1996). Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. Personnel Psychology, Vol. 49, pp. 1-49.

[17] Lewin, K. (1935). Dynamic theory of personality. New York: McGraw-Hill.

[18] Miller, M. J. (1994). A "circuitous" test of Holland's theory. Journal of Employment Counselling, 31, 137-143.

[19] Murray, H.A. (1938). Explorations in Personality. Boston, MA: Houghton Mifflin.

[20] Parsons, F. (1909).Choosing a vocation. Boston: Houghton-Mifflin.

[21] Pervin, L.A. (1968). Performance and satisfaction as a function of individual-environment fit. Psychological Bulletin, Vol. 69, pp. 56-68.

[22] Ryan, A.M. &Snackett, P.R. (1987). A survey of individual assessment practices by I/O psychologists. Personnel Psychology, Vol. 40, pp. 455-488.

[23] Rynes, S.L. &Gerhart, B. (1990). Interviewer assessments of applicant "fit": An exploratory investigation. Personnel Psychology, Vol. 43, pp. 13-35.

[24] Sackett, P. R. (1994) Integrity testing for personnel selection. Current Directions in Psychological Science,Vol. 3, pp. 73-76.

[25] Saks, A.M. &Ashforth, B.E. (1997). A longitudinal investigation of the relationships between job information sources, applicant perceptions of fit, and work outcomes. Personnel Psychology, Vol. 50, pp. 395-426.

[26] Schneider, B. (1987). The people make the place. Personnel Psychology, Vol. 40, pp. 437-454.

[27] Schreiner, E. (2010). Factors Influencing Employment Choices after High School. Retrieved January 20, 2011, from www.ehow.com: http://www.ehow.com/list_6578688_factors-choices-after-high school.html.

[28] Sekiguchi, T. (2004). Person-organization fit and person-job fit in employee selection: A review of the literature. Osaka KeidaiRonshu, Vol. 54 No. 6, pp. 179-196.

[29] Werbel, J.D. & Gilliland, S.W. (1999). Person-environment fit in the selection process. In Ferris G.R. (Ed.).Research in Personnel and Human Resource Management, vol. 17, pp. 209-243. Stamford, CT: JAI Press.

[30] Zaidi. F. B. Iqbal, S. (2012), Impact of career selection on job satisfaction in the service industry of Pakistan, African Journal of Business Management Vol. 6 (9), pp. 3384-3401.

Tracking Items Through Rfid and Solving Heterogeneity Problems During a Collaboration Between Port Companies

Mehdi ABID
Laboratoire d'informatique et d'aide à la décision
Faculty of science, Hassan 2 University
Casablanca, Morocco

Yassine SERHANE
Laboratoire d'informatique et d'aide à la décision
Faculty of science, Hassan 2 University
Casablanca, Morocco

Benayad NSIRI
Laboratoire d'informatique et d'aide à la décision
Faculty of science, Hassan 2 University
Casablanca, Morocco

Haitam AGHARI
Laboratoire Réseaux Informatique, Télécommunication et Multimédia
ENSEM, Hassan 2 University
Casablanca, Morocco

***Abstract*—In this article, we are proposing an architecture that enables improvements in various steps of the collaboration process between different port companies, based on the use of ontologies, multi-agent systems and RFIDs. This approach allows us to collect and present all the data stored in each information system, by exchanging and incorporating any data to facilitate its processing while respecting the territory regulation and compliance, and data confidentiality between all these port companies in a cooperative environment.**

Thanks to the use of RFID (radio frequency identification), this architecture can also deal with the process of tracking commodities belonging to any company that is included in this collaboration process, while each item will be monitored and tracked in real time.

***Keywords*—Ontologies; Multi-agent systems; RFID; Port Information System; Collaboration**

I. Introduction

Several financial losses are caused each year because of errors due to lack of or miscommunication between different companies [1], [2]; these losses engaged many studies based on the use of information and communication technologies between two firms or more to overcome this issue.

Extracting data remains an important step in a collaboration process, as well as tracking every commodity owned by a firm within the collaboration. The use of RFID technology has been proven to be a credible and reliable way to exploit any location of each item belonging to a firm in this collaboration.

The use of RFID technology becomes increasingly exploited in stock management [3], it allows the acceleration and simplification of the inventory process, (annual inventory, perpetual inventory, rolling inventory, auditing…), it also anticipates any warehouse articles out of stock. This technology also provides us with a better tracking, different pathing for each product owned by each firm during shipment from one firm to another.

RFID has a unique number determined by its manufacturer, thus offering a solution to customize an electronic circuit chip Fig.1. The use of this radio-tags provides several advantages in logistics security domain, such as anti-stealing, authentication, and counterfeit detection of a given product [4], through a discreet and different placement of tags implemented in each item.

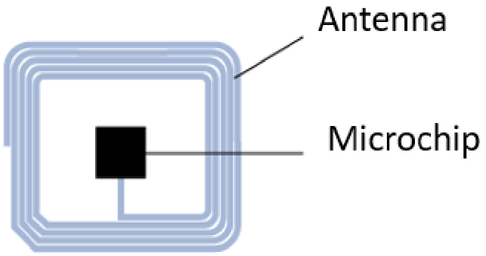


Fig. 1. Logistic chain architecture based on RFID technology

Nowadays, the evolution of technology has pushed the creation of a significant number of information systems dedicated to the port sector, which made the tendency of collaboration more difficult, besides collaborations have to be indispensable to share risk, reduce cost and decrease delay [5].

From an interoperability point of view, during a data exchange in a collaboration between two or more port firms, several problems occur, due to incompatibility between their information systems at the technical level (communication protocols), structural level (variety of database management system) or the semantic level (linguistic differences: name conflicts, synonyms, homonyms or polysemy issues).

This article utilizes several existing logistics' technologies to improve the collaboration process between port companies by tracking merchandise belonging to them using RFID technology.

Our approach also relies on the three-tier architecture, which bases on multi-agent systems, mediators, adapters and various local and global ontologies, to facilitate the exchange of data and overcome all types of heterogeneity (structural, technical and semantic) between all these heterogeneous information systems.

The remainder of the article is as follow: in the second chapter –Background– the state of the art on the use of RFID systems in the field of logistics and architectures designed to resolve semantic conflicts during data exchange, Section 3 –Method–approach definition, Section 4 –Implementation–simulation results, and Section 5 – conclusion.

II. Background

In the early 21st century, discovering RFID technology has driven business logistics especially port companies to use this technology, for instance the use of RFID tag, by providing each item with a RFID tag (Containers, pallets, crates, trolley, package…) facilitate identification task of these items inside warehouses, as well as traceability of any item already tagged.

Nowadays, the affordable price of different RFID tags allow it to be used increasingly in the logistic sector, especially for traceability of goods [6], while the overall idea is to make the RFID reader simultaneously interact with a group of RFID tags as in Fig. 2.

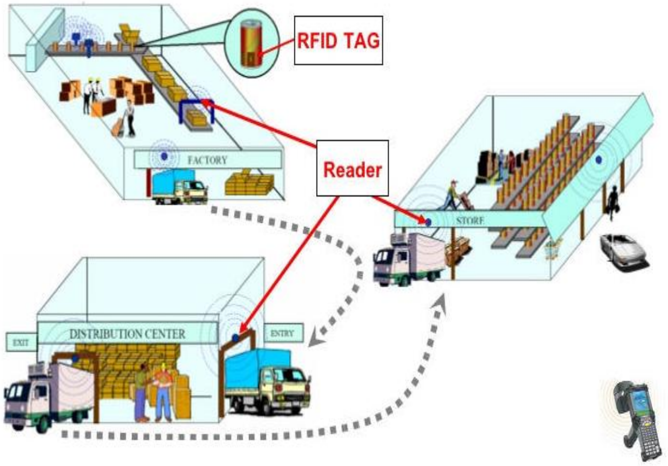


Fig. 2. Logistic chain architecture based on RFID technology

This technology makes the port company management more efficient and transaction more profit-making (cost management in supply chain, storage efficiency).

Currently, there are several types of RFID frequency, classified according to their frequency band (low, high and ultra-high frequency) [7], they depend upon the size of waves projected on the different component of the RFID reader as in Fig. 3.

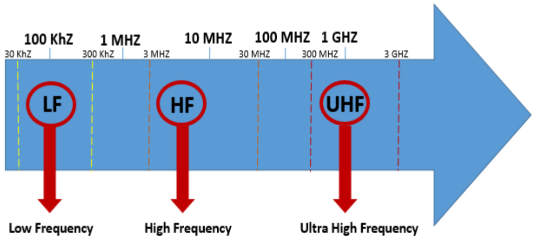


Fig. 3. RFID ranges (LF, HF, and UHF)

- Low frequency: provides a frequency band 30 kHz to 300 kHz (read range between 0 cm and 100 cm), reading is too slow but the data rate of piracy is less high compared to other frequencies during playback.
- High frequency: frequency band 3 MHz to 30 MHz (read range between 10 cm and 1m), read speed is fast but the piracy rate is higher compared to the low frequency.
- Ultra-high frequency: frequency band 300 MHz to 3 GHz (the reading distance can exceed 15 meters), playback speed is faster, but the piracy rate is still higher compared with the high frequency, this frequency is used nowadays in different logistics firms, particularly in the area of storage, inventory and stock management.

A powerful RFID tag can prevent RFID readers without special permission (access control or right frequency) to read the content.

There are 2 main categories of RFID tags, they can be either passive or active[8]:

- Passive tag: generally, it uses waves in order to transmit information through the energy transmitted by the reader, which supplies the onboard electronic circuits.
- Active Tag: usually embeds a source of internal energy (battery with up to 10 years autonomy), it sends the various information stored on the electronic circuit to the RFID server.

Furthermore, the use of RFID has a major role in reducing costs and response time between different companies.

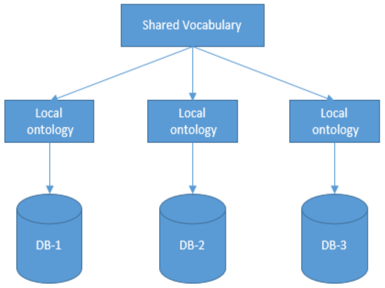


Fig. 4. Collaboration knowledge sharing model between different databases

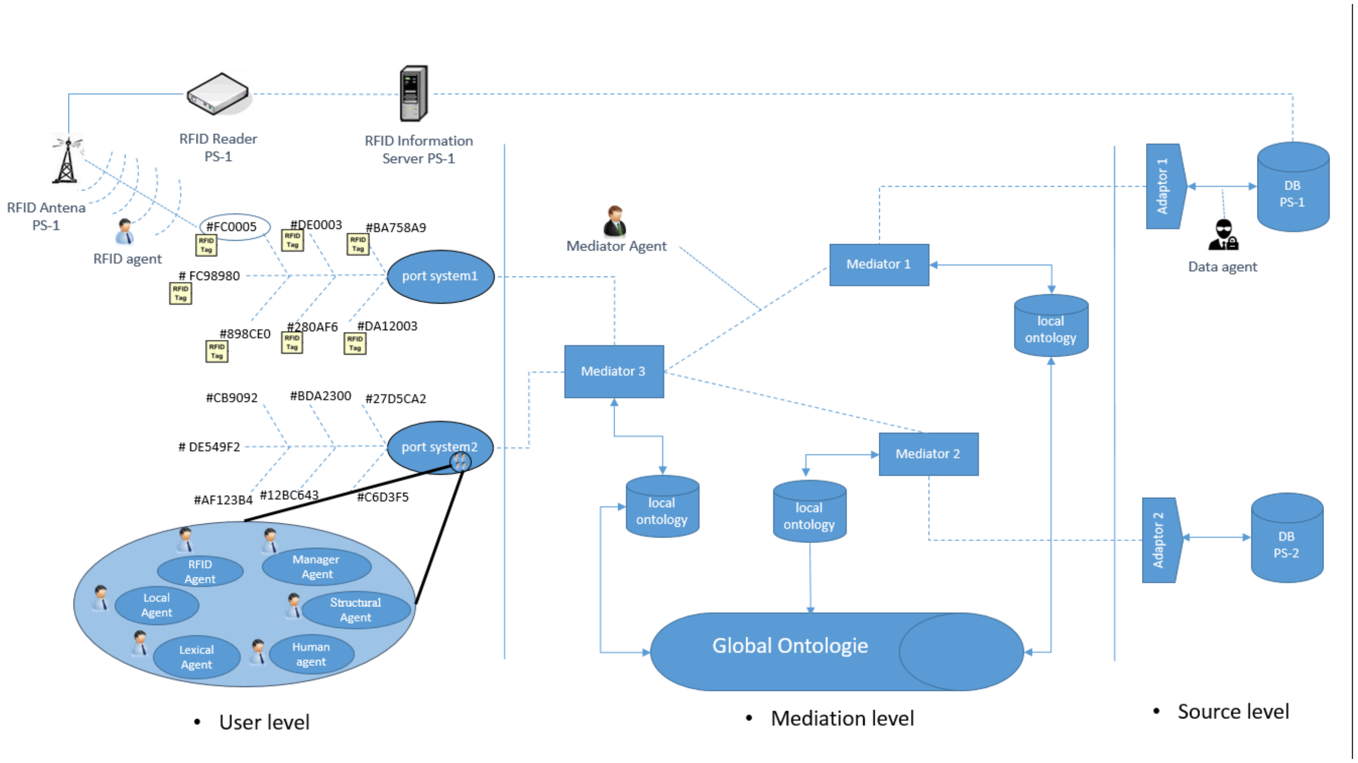


Fig. 5. Our Mediation architecture based on mediator, MAS, RFID, local and global ontologies

Many studies have proposed different approaches which drive to create an architecture used to establish interoperability during a collaboration between all information systems, by fixing the majority problems of heterogeneity [9], [10].

This approach involves several problems (e.g. structural and technical), including but not limited to misinterpretations during information systems data exchange due to linguistic differences, as well as slow data processing due to data redundancy (same data stored in many databases).

Several studies [11], [12] are using multi-agent systems in the collaboration between companies. Multi agents systems (MAS) consist of several groups of agents designed to operate and interact in an organized environment defined as:

MAS = Agent + Environment + interaction + Organization.

Several approaches have been used to solve semantic heterogeneity issues (synonyms, homonyms, and polysemy), such as the ontologies approach see Fig. 4. This approach describes each data source by its own ontology (local ontology), these local ontologies merge together in a single comprehensive ontology (global ontology) sharing a common vocabulary [13], this can occur during a data exchange between systems by making data comprehension easier, such as the consistency of data descriptions exchanged between information systems.

III. METHODE

The architecture that we propose is based on 3 levels, it allows easier data exchange between information systems, as well as access to external information sources (external data sources) to subtract data belonging to heterogeneous information systems, it also allows items (products/articles/goods) traceability all along the supply chain by using RFID technology see Fig. 5.

A. Source level

This level contains the databases belonging to all the different port information systems, it also detects the presence of multiple adapters, where each adapter is positioned between a mediator and a given database which is bound to a specific information system. The adapter is used to establish a unified interface while exchanging information during a collaboration between the information systems, in order to overcome the heterogeneity of any connected sources.

Source level also contains a group of data agents, where each agent has the role to review the different queries issued by an external information system and checks the integrity of assigned privileges, as well as the access rights to an adequate data extraction for each specific port information system.

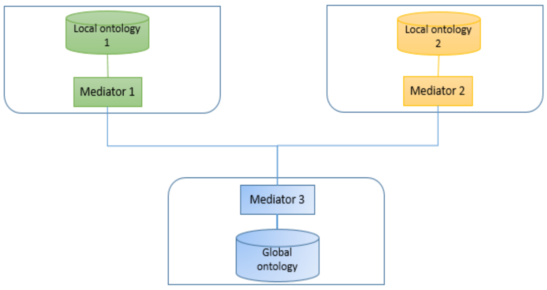


Fig. 6. Data integration architecture based on the use of ontologies

B. Mediation level

This level is characterized by the presence of several mediators, local ontologies, global ontology and mediator agents.

The role of a mediator is to facilitate interconnection between all the information systems even if there are differences at the technical and structural level, and thanks to the mediator agents presence, that have the right permissions of transmitting requests sent out by an external information system user, this mediator also functions as an intermediary between all the information systems databases.

The use of a global ontology in our architecture, is due to defining a specific representation of all the data related to each database as in Fig. 6, this global ontology is a unification (merger) of all the local ontologies. However, local ontologies are intended to classify various data information within each local information system.

During a data exchange, the use of a global ontology allows to unify and translate all data using the existing knowledge databases, in order to overcome the semantic heterogeneity and to ensure an effective collaboration and comprehension between all information systems.

We define the ontological characteristic of an information system space containing different data information assigned to each item, by the following notation:

$$O = <L, S, A^S, R>$$

- L: Language used
- S: Specificity of an article
- A^S: Collection of concept attributes.
- R: Relationships between concepts of the set S (Specificity).

1) Language used:

Language used by the system to distinguish the treatment of using different data, and their meanings depending on the country or region.

2) Specificity:

Si is a set of specificities of each data related to an article. i = {1, 2, 3, 4…, n} where i is the number of considered specificity.

The set S is defined as:

$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5$
$S_1 = \{A | A \in \text{Article}\}$
$S_2 = \{AC | AC \in \text{ArticleCategory}\}$
$S_3 = \{R | R \in \text{RFID\_Tag}\}$
$S_4 = \{D | D \in \text{Data}\}$
$S_5 = \{TA | TA \in \text{TrackArticle}\}$

Article: Essential component of a company's warehouse.

ArticleCategory: different items categories existing in a stock of a port factory warehouse. RFID_Tag: Radio frequency identification tag ID, a unique device identification for each item based on radiofrequency transmission technology.

Data: information stored in the databases.

TrackArticle: information on an article location, to help determine its position and maintain its traceability, during an expedition (export/import) outside the company.

3) Collection of concept attributes:

ASj, is a set of specificities of each article related data, where j= {1, 2, 3, 4…, n} is the number of considered specificity.

The set ASj is defined as:

$A^S = A^S_1 \cup A^S_2 \cup A^S_3 \cup A^S_4 \cup A^S_5$
$A^S_1 = \{\text{Ref} | \text{Ref} \in \text{Reference}\}$
$A^S_2 = \{T | T \in \text{Type}\}$
$A^S_3 = \{IR | IR \in \text{IdentificationRadioFrenquency}\}$
$A^S_4 = \{ID | ID \in \text{Identification}\}$
$A^S_5 = \{L | L \in \text{Location}\}$

Reference: references assigned to each product designated by a unique identifier.

Type: determines the type of products which is based on its features and group belonging (radioactive – dangerous – organic…).

IdentificationRadioFrenquency: Tag ID identification of each article belonging to a System.

Identification: consists of a unique chain of characters that differentiate each items from another and allows extracting a specific database information based on it.

Location: set of connector's id planted in the supply chain in order to determine the real time location of a RFID tagged article during an expedition.

4) Relation:

R is the set of relationships defined in our ontological characteristic space:

$R = R_1 \cup R_2 \cup R_3$
$R_1 = \{\text{Article, Article category, Article Data}\}$
$R_2 = \{\text{Article, read}\}$
$R_3 = \{\text{Article, trace}\}$

R_1: Relation defined by an article, its category and its data. To establish an appropriate and effective structure that enables proper management of stocks inside the warehouse.

R_2: Relation between an article and its RFID Tag.

R_3: Relation between each item and its location, during an expedition for traceability.

C. User Level

This level consists of a large number of port information systems, it is based on the monitoring of different products owned by an inside company or an outsider, each item labeled with a tag is connected to an antenna and also has a memory, which stores its product electronic code, in order to seek for RFID readers and send data about each item, which then stores its information into a RFID server.

This process ensures a better tracking management of each product related to any Port Information system.

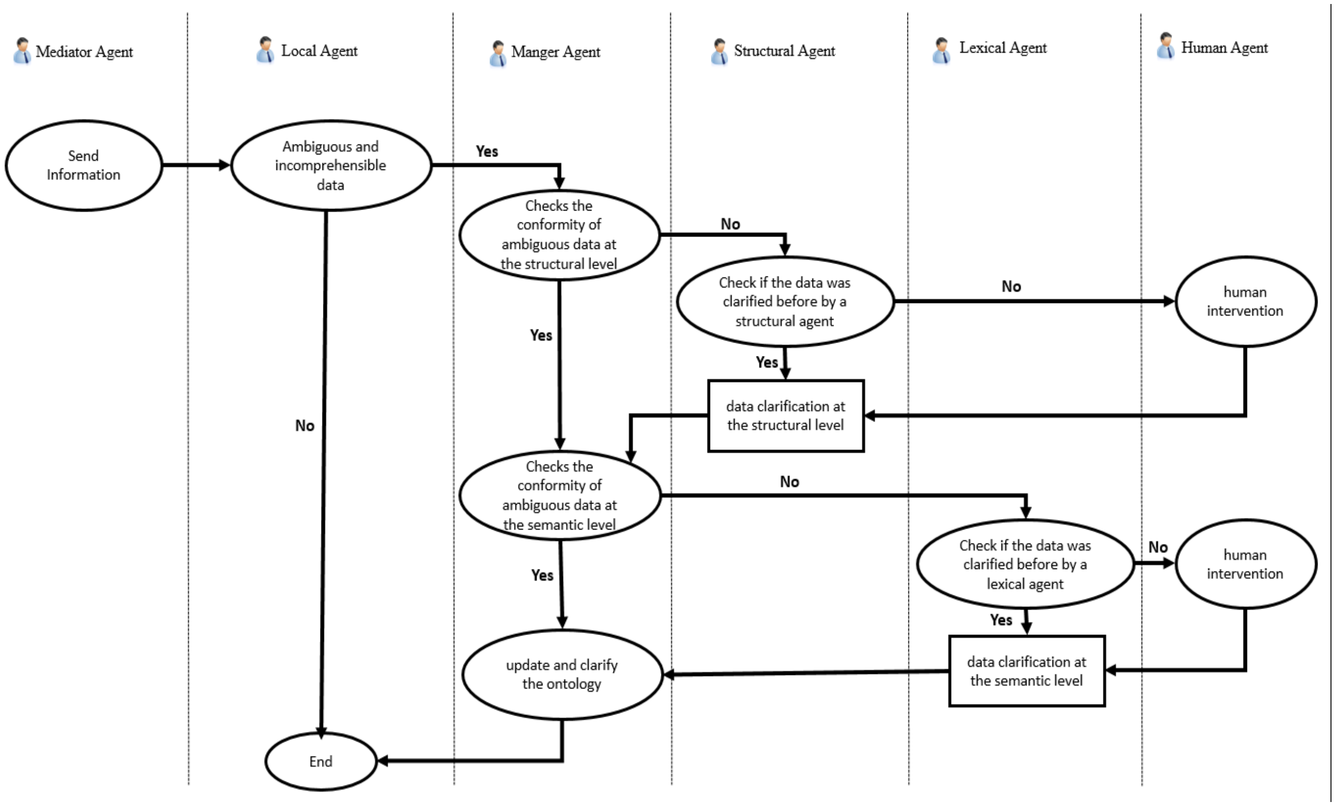

Mediator Agent
Local Agent
Manger Agent
Structural Agent
Lexical Agent
Human Agent
Send Information
Ambiguous and incomprehensible data
Yes
Checks the conformity of ambiguous data at the structural level
No
Check if the data was clarified before by a structural agent
No
human intervention
Yes
Yes
data clarification at the structural level
No
Checks the conformity of ambiguous data at the semantic level
No
Check if the data was clarified before by a lexical agent
No
human intervention
Yes
Yes
data clarification at the semantic level
update and clarify the ontology
End

Fig. 7. Communication prototype between agents during a data exchange

Each information system is also characterized by the presence of a group of agents that interact in an organized manner, to solve any problem of heterogeneity at the semantic level, while exchanging data between different information systems.

This level is composed of six types of agents: Mediator Agent, Local Agent, Manager Agent, Structural Agent, Lexical agent, human agent

Mediator Agent: Sends product information and data requested by an external system.

Local Agent: Gathers and transfers to the Manager Agent any ambiguous or incomprehensible coming from a Mediator Agent.

Manager Agent: Checks the conformity of ambiguous data on structural and semantic level, in case of failure the Manager Agent sends the unintelligible data to the Lexical agent, and also sends the ambiguous data to the structural Agent to resolve any structural or semantic conflicts see Fig. 7.

Both Agents (Structural Agent and Lexical Agent) check if the data has already been processed manually through human intervention before they update and clarify the ontology.

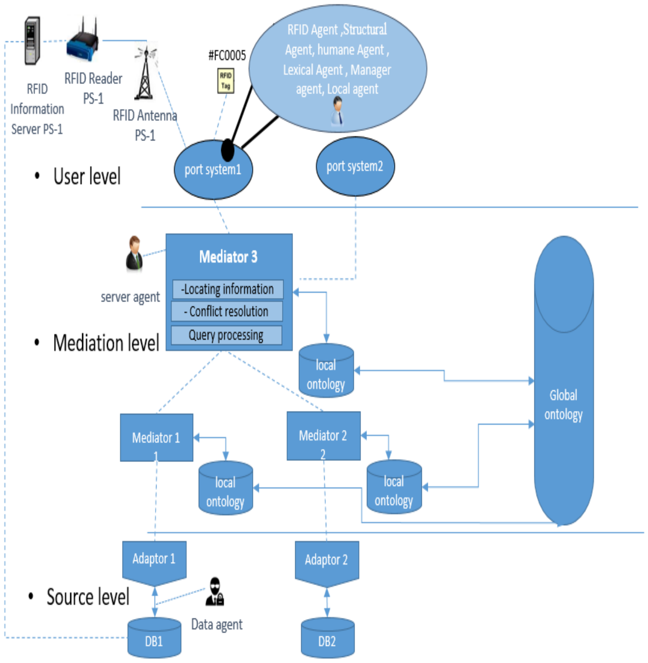
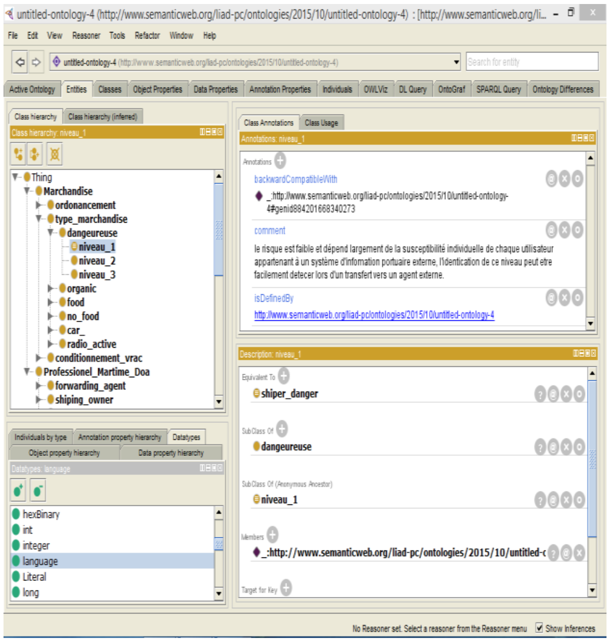


Fig. 8. Data mediation model applied between two Port Information

IV. Implementation

In this chapter, we decided to adopt a data mediation prototype, as well as individual item tracking between two port information systems {PIS1, PIS2} that are different at the technical, the structural and the semantic levels. This prototype has been developed through our approach defined in the preceding chapter, which aims to facilitate solving heterogeneity problems cooperation between firms in different locations that use their own architecture as well as their own data model.

Fig.8 describes the 3-tier architecture which we use for data exchange, access to different sources of information and tracking various items, in order to facilitate the communication between the two systems.



Annotations: niveau_1

backwardCompatibleWith
_:http://www.semanticweb.org/liad-pc/ontologies/2015/10/untitled-ontology-4#genid884201668340273

comment
le risque est faible et dépend largement de la susceptibilité individuelle de chaque utilisateur appartenant à un système d'infomation portuaire externe, l'identication de ce niveau peut etre facilement detecer lors d'un transfert vers un agent externe.

isDefinedBy
http://www.semanticweb.org/liad-pc/ontologies/2015/10/untitled-ontology-4

Description: niveau_1

Equivalent To: shiper_danger

SubClass Of: dangeureuse

SubClass Of (Anonymous Ancestor): niveau_1

Members: _:http://www.semanticweb.org/liad-pc/ontologies/2015/10/untitled-

untitled-ontology-4 (http://www.semanticweb.org/liad-pc/ontologies/2015/10/untitled-ontology-4) : [http://www.semanticweb.org/li...
File Edit View Reasoner Tools Refactor Window Help
Active Ontology Entities Classes Object Properties Data Properties Annotation Properties Individuals OWLViz DL Query OntoGraf SPARQL Query Ontology Differences
Class hierarchy Class hierarchy (inferred)
Class hierarchy: niveau_1
Thing
Marchandise
ordonancement
type_marchandise
dangeureuse
niveau_1
niveau_2
niveau_3
organic
food
no_food
car_
radio_active
conditionnement_vrac
Professionel_Martime_Doa
forwarding_agent
shiping_owner
Class Annotations Class Usage
Datatypes: language
hexBinary
int
integer
language
Literal
long
No Reasoner set. Select a reasoner from the Reasoner menu
Show Inferences

Fig. 9. An ontology Sample built using "Protégé" (open source ontology editor)

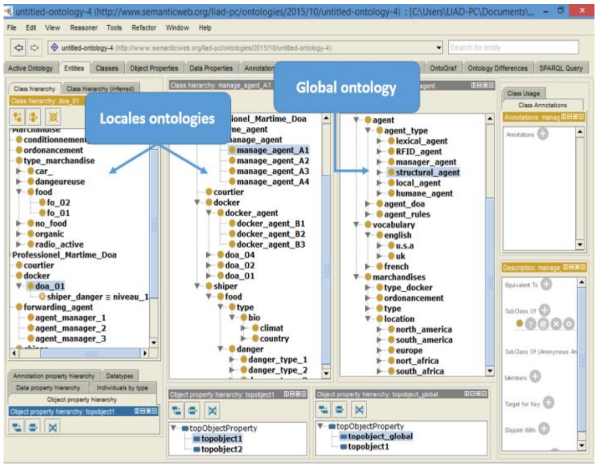


Fig. 10. Two ontologies unification under "Protégé" editor

The first level, which is the source level, consists of two different databases, one has been modeled with "UML" and we chose MySQL as a database management system, the second one was modeled with "Merise" using SQL Server as a database management system, this level also contains adapters displaying the results of queries issued by a user under the appropriate format for each information system.

The second level, which is the mediation level, contains tools that enable query processing from an information system, it also allows heterogeneity problems resolving through the use of mediators, local ontologies and the global ontology.

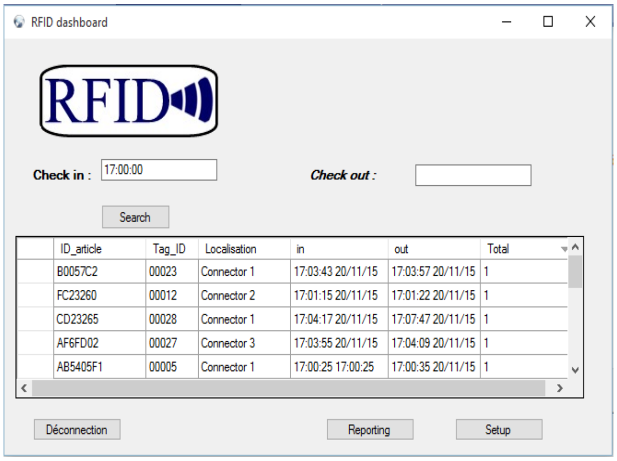
Each mediator aims to ensure communication and data exchange between each information system by solving structural and semantic conflicts. The semantic conflict is resolved using local ontologies associated with each mediator, these ontologies, which have been implemented through the editor Protégé (Fig. 9), contain different knowledge databases for each information system, they are built to overcome any type of semantic heterogeneity conflicts such as synonymous paronym, antonyms and homonyms problems.

All the local ontologies knowledge databases are merged into a single global knowledge base linked to the global ontology see Fig. 10, which encompasses the entire local ontologies to strengthen the semantic manual of the data exchanged.

The third level, which is the user level, contains two different user interfaces of the two Port Information Systems {PIS1, PIS2}, where each system is implemented differently, (Fig. 11 - 12).

The first system was developed using .NET WPF (Windows Presentation Foundation) for the user interface as for the second it was developed using Java Enterprise Edition 1.5.

The aim of the collaboration is to consult and extract data belonging to each external information system, and locate each product or item within the company or during the expedition, through the use of the RFID tag planted in each item see Fig. 13. Each product location is stored in an RFID server, and then transferred to the database of its information system.



RFID dashboard

Check in : 17:00:00

Check out :

	ID_article	Tag_ID	Localisation	in	out	Total
	B0057C2	00023	Connector 1	17:03:43 20/11/15	17:03:57 20/11/15	1
	FC23260	00012	Connector 2	17:01:15 20/11/15	17:01:22 20/11/15	1
	CD23265	00028	Connector 1	17:04:17 20/11/15	17:07:47 20/11/15	1
	AF6FD02	00027	Connector 3	17:03:55 20/11/15	17:04:09 20/11/15	1
	AB5405F1	00005	Connector 1	17:00:25 17:00:25	17:00:35 20/11/15	1

Fig. 13. Item tracking process dashboard

According to the mediation architecture that we adapted in the previous section, which is based on the use of the multi-agents system, the local ontologies and the global ontology, we were able to come up with a prototype to establish a successful data exchange between these two information systems by solving the majority of the heterogeneity problems.

V. Conclusion

This paper proposes a 3-tiers architecture approach based on the use of multi-agent systems, ontologies and mediators, this approach makes it possible to solve many conflicts (structural, technical and semantic) during a data exchange or information sharing between all the different port information systems databases, we were able to establish the process of using RFID technology, which allows to track articles belonging to each port company during their expeditions, from the starting until the arrival, through the different steps of the supply chain, this process will ensure strengthening the collaboration between these firms.

This approach establishes a mechanism of data information managing and decision making between all port information systems, in order to work collaboratively with all these port companies in a more efficient manner.

References

[1] L. R. Varshney and D. V. Oppenheim, “On Cross-Enterprise Collaboration,” in Business Process Management, S. Rinderle-Ma, F. Toumani, and K. Wolf, Eds. Springer Berlin Heidelberg, 2011, pp. 29–37.

[2] P. Friedl, R. Biloslavo, and others, “Association of Management Tools with the Financial Performance of Companies: The Example of the Slovenian Construction Sector,” Manag. Glob. Transit., vol. 7, no. 4, pp. 383–402, 2009.

[3] B. Sheng and C. C. Tan, “Group authentication in heterogeneous RFID networks,” in Homeland Security (HST), 2012 IEEE Conference on Technologies for, 2012, pp. 167–172.

[4] P. Tuyls and L. Batina, “RFID-Tags for Anti-counterfeiting,” in Topics in Cryptology – CT-RSA 2006, D. Pointcheval, Ed. Springer Berlin Heidelberg, 2006, pp. 115–131.

[5] M. ABID, B. NSIRI, and Y. SERHANE, “Interoperability between different port information systems,” Int. J. Math. Comput. Simul., vol. 8, pp. 156–161, 2014.

[6] G. R. Ram, N. R. Babu, N. P. Sudhakar, B. Raviteja, and K. Rammohanarao, “TRACKING OBJECTS USING RFID AND WIRELESS SENSOR NETWORKS,” Int. J. Eng. Sci. Adv. Technol., vol. 2, no. 3, p. 515, 2012.

[7] G. Swagarya, L. Boaz, and M. Kisangiri, “International Journal of Scientific Engineering and Research (IJSER) Designing of UHF-Radio Frequency Identification (RFID) Antenna,” 2013.

[8] H. M. Quraishi and F. Farheen, “Curtain Raiser on Gen-Next of RFID Technology,” IOSR J Ournal Electr. Electron. Eng. IOSR - JEEE, pp. 01–05, 2014.

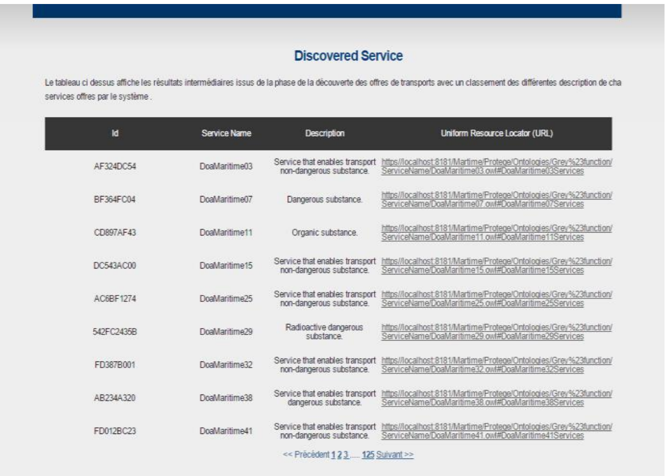
[9] R. Hammami, H. Bellaaj, and A. H. Kacem, “Interoperability for medical information systems: an overview,” Health Technol., vol. 4, no. 3, pp. 261–272, May 2014.

[10] R. Jardim-Goncalves, A. Grilo, and K. Popplewell, “Novel strategies for global manufacturing systems interoperability,” J. Intell. Manuf., pp. 1–9, Aug. 2014.

[11] F.-S. Hsieh and J.-B. Lin, “A self-adaptation scheme for workflow management in multi-agent systems,” J. Intell. Manuf., pp. 1–18, Aug. 2013.

[12] M. Miranda, M. Salazar, F. Portela, M. Santos, A. Abelha, J. Neves, and J. Machado, “Multi-agent Systems for HL7 Interoperability Services,” Procedia Technol., vol. 5, pp. 725–733, 2012.

[13] F. Freitas, H. Stuckenschmidt, and N. F. Noy, “Guest editor’s introduction: Ontology issues and applications,” J. Braz. Comput. Soc., vol. 11, no. 2, pp. 5–16, 2005.

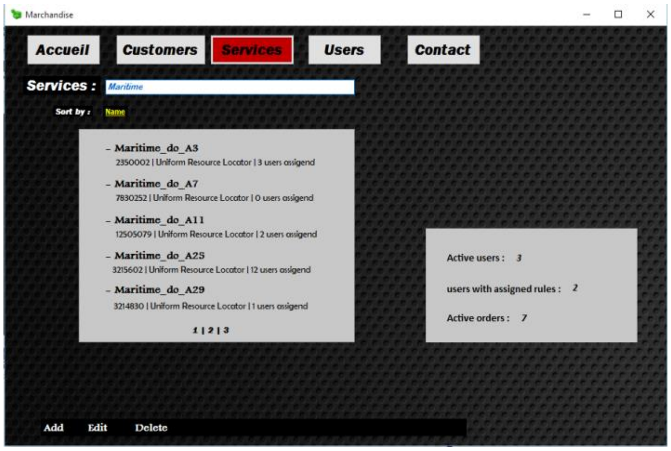


Discovered Service

Le tableau ci dessus affiche les résultats intermédiaires issus de la phase de la découverte des offres de transports avec un classement des différentes description de cha services offres par le système .

Id	Service Name	Description	Uniform Resource Locator (URL)
AF324DC54	DoaMaritime03	Service that enables transport non-dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime03.owl#DoaMaritime03Services
BF364FC04	DoaMaritime07	Dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime07.owl#DoaMaritime07Services
CD897AF43	DoaMaritime11	Organic substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime11.owl#DoaMaritime11Services
DC543AC00	DoaMaritime15	Service that enables transport non-dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime15.owl#DoaMaritime15Services
AC6BF1274	DoaMaritime25	Service that enables transport non-dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime25.owl#DoaMaritime25Services
542FC2435B	DoaMaritime29	Radioactive dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime29.owl#DoaMaritime29Services
FD387B001	DoaMaritime32	Service that enables transport non-dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime32.owl#DoaMaritime32Services
AB234A320	DoaMaritime38	Service that enables transport dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime38.owl#DoaMaritime38Services
FD012BC23	DoaMaritime41	Service that enables transport non-dangerous substance.	https//localhost:8181/Martime/Protege/Ontologies/Grey%23function/ServiceName/DoaMaritime41.owl#DoaMaritime41Services

Fig. 11. Different description services Screenshots provided by the first information system developed under Java Enterprise Edition



Services : Maritime

Sort by : Name

- Maritime_do_A3
  2350002 | Uniform Resource Locator | 3 users assigend
- Maritime_do_A7
  7830252 | Uniform Resource Locator | 0 users assigend
- Maritime_do_A11
  12505079 | Uniform Resource Locator | 2 users assigend
- Maritime_do_A25
  3215602 | Uniform Resource Locator | 12 users assigend
- Maritime_do_A29
  3214830 | Uniform Resource Locator | 1 users assigend

Active users : 3

users with assigned rules : 2

Active orders : 7

Fig. 12. Services Screenshots provided by the second information system developed under. .NET WPF

Frequency Domain Analysis for Assessing Fluid Responsiveness by Using Instantaneous Pulse Rate Variability

Pei-Chen Lin
Institute of Biomedical Engineering
National Chiao Tung University
Hsinchu, Taiwan R.O.C.

Chia-Chi Chang
Department of Electronics Engineering and Institute of Electronics, Institute of Biomedical Engineering,
Biomedical Electronics Translational Research Center
National Chiao Tung University
Hsinchu, Taiwan R.O.C.

Hung-Yi Hsu
Department of Neurology
Chung Shan Medical University
Section of Neurology, Department of Internal Medicine
Tungs' Taichung Metro Harbor Hospital, No.699, Sec. 8,
Taiwan Blvd., Wuqi Dist., Taichung City 435
Taichung, Taiwan R.O.C.

Tzu-Chien Hsiao*
Department of Computer Science, Institute of Biomedical Engineering, Biomedical Electronics Translational Research Center
National Chiao Tung University
Hsinchu, Taiwan R.O.C.

Abstract*—In the ICU, fluid therapy is conventional strategy for the patient in shock. However, only half of ICU patients have well-responses to fluid therapy, and fluid loading in non-responsive patient delays definitive therapy. Prediction of fluid responsiveness (FR) has become intense topic in clinic. Most of conventional FR prediction method based on time domain analysis, and it is limited ability to indicate FR. This study proposed a method which predicts FR based on frequency domain analysis, named instantaneous pulse rate variability (iPRV). iPRV provides a new indication in very high frequency (VHF) range (0.4-0.8Hz) of spectrum for peripheral responses. Twenty six healthy subjects participated this study and photoplethysmography signal was recorded in supine baseline, during head-up tilt (HUT), and passive leg raising (PLR), which induces variation of venous return and helps for quantitative assessment of FR individually. The result showed the spectral power of VHF decreased during HUT (573.96±756.36 ms^2 in baseline; 348.00±434.92 ms^2 in HUT) and increased during PLR (573.96±756.36 ms^2 in baseline; 718.92±973.70 ms^2 in PLR), which present the compensated regulation of venous return and FR. This study provides an effective indicator for assessing FR in frequency domain and has potential to be a reliable system in ICU.

Keywords—fluid responsiveness (FR); instantaneous pulse rate variability (iPRV); head-up tilt (HUT); passive leg raising (PLR)

I. Introduction

In intensive care unit (ICU), most of patients shock due to lack amount of blood who are after surgery or during injury situation. Fluid therapy is a frequent therapeutic strategy for the shock. However, fluid loading in the non-fluid response patient will delay definitive therapy and may be harmful. Furthermore, only half of patients have well fluid responsiveness (FR) in the ICU. Based on these for FR in patients is important in clinical. It is important to develop the reliable prediction method.

Conventional methods usually predict FR by using time domain analysis, such as impedance cardiography (ICG) and pleth variability index (PVI). ICG is a non-invasive method to detect electrical and impedance changes in the thorax by using dual sensors. The electrical and impedance changes are used to calculate hemodynamic parameters for evaluating fluid response. The reliability of hemodynamic parameters in ICG is based on placed position of dual sensors, which needs to operate by professional paramedic. PVI is a time domain analysis for predicting FR in mechanically ventilated patients. It adopts pulse oximeter waveform and then calculates dynamic change in perfusion index (PI) during respiratory cycle. Nevertheless, time domain analysis effects by motion artifact easily. For avoiding inaccuracy in time domain analysis by motion artifact, this study proposed a reliable method which predict FR by frequency domain analysis.

Heart rate variability (HRV) can be measured by interbeat intervals (RRi) on electrocardiogram (ECG), which provides time domain and frequency domain analysis to assess autonomic nervous system (ANS). Especially, frequency domain analysis is used to adopt fast Fourier transform (FFT) for spectral analysis. Furthermore, spectral is divided into low-frequency (LF) range (0.04-0.15Hz) to present sympathetic nervous system activities, and high-frequency (HF) range (0.15-0.4Hz) to present parasympathetic nervous system activities mainly. However, HRV studies are restricted by the feasibility and the reproducibility with inconvenient measurement [1]. Therefore, pulse rate variability (PRV) was proposed as a substitute measurement of HRV. PRV uses pulse wave, which collected from photoplethysmography (PPG), to replace ECG recording in HRV and has been examined as a surrogate of HRV during non-stationary conditions in previous study [2]. Moreover, the arterial pulse wave is regulated from complex physiological controls which make PRV provide much more information than HRV.

However, frequency domain analysis of HRV and PRV are both limited by the timescale of RRi and pulse wave time intervals. Since timescale limitation, the indication of high frequency range in spectral analysis was restricted by time resolution. For breaking timescale limitation, a novel adaptive method, named instantaneous pulse rate variability (iPRV), was proposed [3]. It adopted the frequency range extension method based on ensemble empirical mode decomposition (EEMD) [4] and instantaneous period (iPR) projecting technique [5] help for PRV spectral analysis. Therefore, iPRV provides a new indication, named very high frequency (VHF) range (0.4-0.8Hz) for the neural regulatory estimation and peripheral responses [6]. The literature has proposed that VHF of HRV is as a novel index of left ventricular function evaluation [7], which further indicates the cardiac function, venous return, and FR. However, the variation and interpretation of VHF of iPRV still needs further exploration and examination. There is a common clinical experiment, named passive leg raising (PLR), which induces the increase of venous return and helps for the quantitative assessment of FR [8]. Previous study has examined that iPRV is reliable by using PPG during non-stationary condition, such as head-up tilt (HUT) [9]. This study performed the clinical experiment, known as HUT and PLR, for the further exploration and examination. Since several studies revealed that the VHF index of HRV is a reliable evaluation of cardiovascular diseases [7]. VHF of iPRV has potential to indicate more physiological responses.

The aim of this study is to 1) explore the potential indication of VHF of iPRV during HUT and PLR, and 2) interpret the physiological meaning of VHF of iPRV in different non-stationary conditions.

II. Methods

A. Subjects and Data Collection

Twenty six healthy subjects (male: 14; age: 24±1), who had no history of cardiovascular disease, participated this study. All recruited subjects performed four trials in whole experiment. First, subjects were rest in supine position with 10-minute recording as baseline. Second, subjects were tilting up passively (HUT) on the automatic tilting table and kept in tilt-up position for 10 minutes. Then, subjects were back to the supine position with 5 minutes for recovering to baseline. Finally, subjects were raising leg passively (PLR) for 10 minutes. All measurements were performed in a quiet temperature controlled room and the experiment was approved by institutional review board of the hospital. This study was approved by institutional review board of Tungs' Taichung Metro Harbor Hospital. Informed consent was obtained from all participants before the experiment.

The ECG signal was recorded by BEST-C-04056 (BioSenseTek Corp., Taiwan) and the PPG signal was recorded by Nonin 8500 (Nonin Medical Inc., Plymouth, MN) with a sampling frequency 200Hz.

B. Instantaneous Pulse Rate Variability

The algorithm of iPRV analysis shows at Fig. 1. At first, the blood pulse signal was extracted from PPG signal as the pulse wave component by sifting process in EEMD. Sifting process is an iteratively detrending operation which uses to compute finite set of components, named intrinsic mode functions (IMFs), from source non-stationary data. Moreover, before sifting process, EEMD provides noise-assisted method into original data for eliminating multiple characteristic problem in IMFs. After mixtures of added noise and source data, detrending operation contains several steps. First, local extrema of data $x(t)$ are identified by peak-valley detection. The upper envelope $U(t)$ and lower envelope $L(t)$ are generated by cubic spline interpolation according to the local maxima and local minima. The trend in current timescale is computed by calculating the mean of $U(t)$ and $L(t)$, as $M(t)$.

$$M(t) = (U(t) + L(t))/2 \tag{1}$$

The new timescale $H(t)$ is representation after detrending operation by data $x(t)$ subtracting the trend.

$$H_k(t) = H_{k-1}(t) - M_k(t), k \geq 1 \tag{2}$$

Where $H_0(t) = x(t)$. After k times detrending operation, if the trend of $H_k(t)$ satisfies the criterion as the steady constant trend, then the components $H_k(t)$ were extracted from $x(t)$ as IMF. After n sifting process, $x(t)$ was decomposed into n IMFs, $IMF_1(t) \sim IMF_n(t)$, and one residue $r(t)$.

$$x(t) = \sum_{i=1}^{n} IMF_i(t) + r(t) \tag{3}$$

Since IMFs were decomposed from different mixtures, the ensemble IMFs are computed by averaging each corresponding IMF. However, the resolution of timescale still limit spectral analysis. For breakthrough timescale limitation, iPRV adopts iPR of blood pulse signal for proposing variation by using normalized direct quadrature (NDQ) [5]. NDQ contains several steps. First, the amplitude modulation of main component IMF_{main} was eliminated by iteratively normalization. Then, the empirical frequency modulation (FM) signal $F(t)$ of IMF_{main} is assumed to be cosine function, and its quadrature $sin\emptyset(t)$ can be computed directly.

$$sin\emptyset(t) = \sqrt{1 - F^2(t)} \tag{4}$$

The instantaneous phase $\emptyset(t)$ is calculated by taking $arctangent$ of FM signal and its quadrature, then the iPRinstan is obtained from inverse of the derivative of instantaneous phase.

$$\emptyset(t) = tan^{-1}(\sqrt{1 - F^2(t)}/F(t)) \tag{5}$$

Finally, fast Fourier transform was performed as the spectral analysis in each frequency band of IP. The spectral power of LF, HF and VHF were calculated by spectral integration as the clinical indicators. The spectral analysis programs in this study was developed by using commercial software platform (LabVIEW version 2013, National Instruments Corp., Austin, USA).

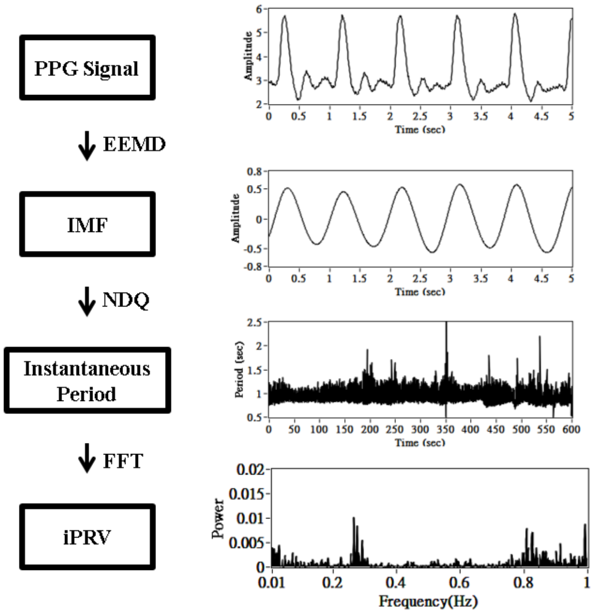


Fig. 1. The flow illustration of the algorithm of instantaneous pulse rate variability (iPRV)

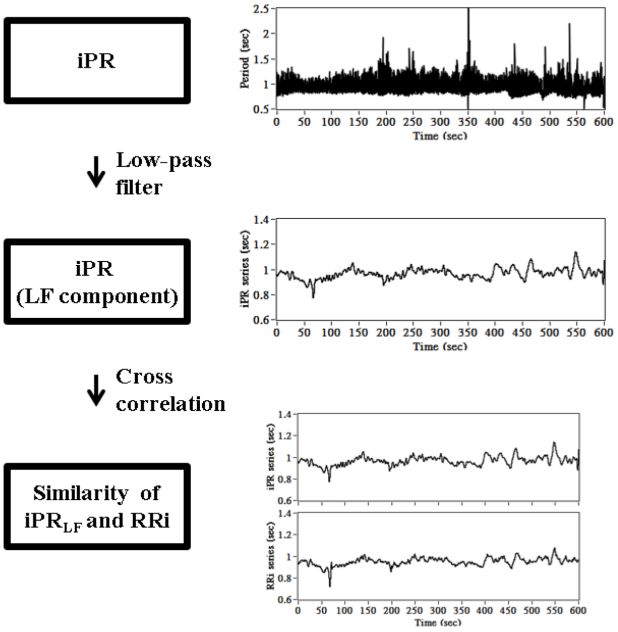


Fig. 2. The flow illustration of the algorithm of correlation analysis

C. Time Domain Sequential Analysis

This study used a correlation analysis to time domain sequential for ensuring similarity between iPRV and HRV during non-stationary conditions. Correlation analysis has several steps as follows (Fig. 2). First, low frequency band of iPR was filtered by low-pass filter in order to receive instantaneous pulse rate (iPR$_{LF}$) series $f_{iPR}(t)$ for comparing with interpolated time series of RRi $g_{RRi}(t)$. Then, cross correlation $R(d)$ and mean square error between iPR$_{LF}$ and RRi were calculated to measure the similarity between iPRV and HRV as time domain sequential analysis.

$$R(d) = \frac{\sum_{t=0}^{N}[(f_{iPR}(t)-\overline{f_{\iota PR}})*(g_{RRi}(t)-\overline{g_{RR\iota}})]}{\sqrt{\sum_{t=0}^{N}(f_{iPR}(t)-\overline{f_{\iota PR}})^2} * \sqrt{\sum_{t=0}^{N}(g_{RRi}(t)-\overline{g_{RR\iota}})^2}} \qquad (6)$$

Where N presents data length number of iPR$_{LF}$ and RRi. d is the time shift. $\overline{f_{\iota PR}}$ and $\overline{g_{RR\iota}}$ are the means of the corresponding series.

D. Statistic Analysis

Variation of spectral power in each frequency band were compared between different analysis method and different condition using paired-sample t test for significant difference. P value less than 0.05 was considered statistically significant. Results of spectral power are reported as mean ± standard deviation. Statistical analysis was performed using commercial statistics software.

III. Results

A. Time Domain Sequential Analysis Between Baseline and Non-stationary States

The comparison of time series between iPR$_{LF}$ and RRi was presented in Fig. 3 in one of the participants as an example. The iPR$_{LF}$ and RRi were similar fluctuation in different conditions sequentially in time domain.

The results of the time domain sequential analysis were summarized in Table I. The results of all participants' cross correlation between iPR$_{LF}$ and RRi were high correlation (0.667±0.109 in baseline; 0.672±0.096 in HUT; 0.675±0.105 in PLR). Mean square error between iPR$_{LF}$ and RRi were quite small (0.005±0.004 in baseline; 0.004±0.005 in HUT; 0.005±0.005 in PLR). There is no significant difference between each condition.

TABLE I. The Result of Cross Correlation and Mean Square Error in Different Conditions

	Baseline	Head-up tilt (HUT)	Passive leg raising (PLR)
Cross correlation	0.667±0.109	0.672±0.096	0.675±0.105
Mean square error	0.005±0.004	0.004±0.005	0.005±0.005

The form is (mean ± standard deviation).

B. Spectral Analysis of Different Conditions in HRV and iPRV

The results of the spectral analysis in different conditions were summarized in Table II. In HRV spectrum, the power of LF increased both in HUT and PLR. The power of HF decreased in HUT. The power of VHF in each condition was small in HRV spectrum. In iPRV spectrum, the power of LF increased both in HUT and PLR, which were the same variation as HRV. The power of HF decreased in HUT but increased in PLR. The power of VHF decreased significantly during HUT and increased during PLR. The illustration of the iPRV spectrum were summarized in Fig. 4 in one subject for example. The results of all participants' iPRV spectrum were similar with subtle change of the frequency peaks' locations. In LF band, there is a spectral peak around 0.1Hz in each experiment. In HF band, the peak is around 0.3 Hz in each experiment. In VHF band, there is a peak around 0.7 to 0.8 Hz when subject during supine and PLR position.

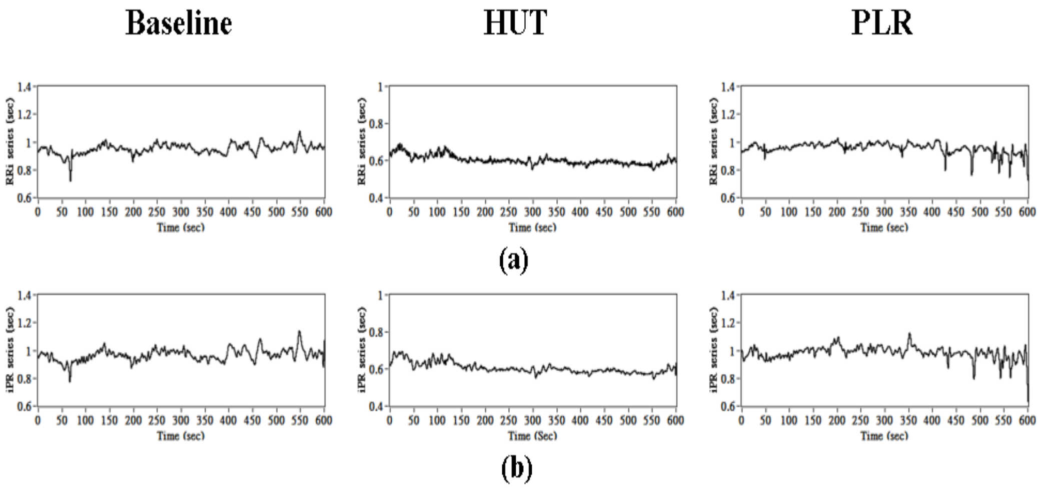


Fig. 3. The comparison of fluctuation of time series during different conditions. (a) RRi from ECG, (b) iPR_{LF} from PPG

TABLE II. THE RESULT OF HRV AND IPRV SPECTRUM

		Baseline	***Head-up tilt (HUT)***	***Passive leg raising (PLR)***
HRV	***LF***	229.73±249.77	303.92±531.35	268.54±412.64
	HF	228.23±192.75	203.23±184.64	226.58±183.88
	VHF	22.38±27.75	16.00±10.87	28.27±30.34
iPRV	***LF***	324.58±393.67	401.85±771.51	427.50±691.18
	HF	474.23±430.27	363.00±284.71	515.50±448.09
	VHF	573.96±756.36	348.00±434.92*	718.92±973.70

The form is (mean ± standard deviation); * means $p<0.05$ compared with HRV. LF denotes low frequency band (0.04-0.15Hz); HF denotes high frequency band (0.15-0.4Hz); VHF denotes very high frequency band (0.4-0.8Hz).

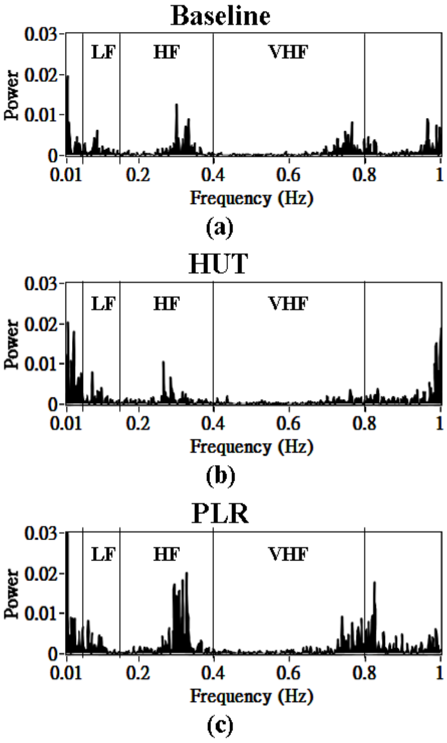


Fig. 4. The illustration of the iPRV spectrum during (a) baseline, (b) head-up tilt (HUT), and (c) passive leg raising (PLR) in one of the participant as an example

IV. DISSCUSSION

The component represents pulse wave was extracted from PPG signal by using EEMD. As a result, NDQ is a reliable technique to transform pulse wave component into iPR, which indicates the regulation of ANS in spectral analysis [3]. The reliability of iPRV in frequency domain had been examined in the literature [9]. In addition, the information from iPR is more than from RRi by using frequency extension method to break limitation of time resolution in RRi. For the verification of the reliability of iPR in time domain, this study performed time sequential comparison of iPR_{LF} and RRi for further examination. The result illustrated that the time series of iPR is similar with RRi and contains much more intrinsic components with high frequency component, which provides much more physiological information for the assessment. Moreover, iPRV has high positive correlation with HRV in time domain. It has the potential usefulness as an indicator for cardiovascular circulation assessment.

Previous study had examined iPRV spectrum assessed the new indicator (VHF) to show more physiological information [9]. Though some literatures investigated that VHF of HRV is a reliable evaluation of left ventricular function [7], VHF of iPRV still needs further examination. It had been examined that VHF contains parasympathetic activities and peripheral responses, which are influenced by venous return and cardiac function. The influences of respiration on VHF were examined by paced respiration study [6]. The mechanism of the VHF indication needs more exploration.

This study applied the clinical experiment, known as HUT and PLR, for the further examination. HUT and PLR served as the simple clinical experiments for cardiovascular circulation evaluation. HUT causes temporarily decrease of blood volume in upper body and then causes the decrease of venous return. These changes induce the auto-regulation for the compensation. The sympathetic activities increased during HUT, and the power of LF in HRV also increased, which quantitatively assessed the sympathetic activation. On the other hand, the parasympathetic activities decreased, and so did the power of HF in HRV. The LF and HF in iPRV shows the same variation with HRV during HUT. Furthermore, it offers information in the power of VHF decreased when venous return decreased. In another experiment, PLR causes the increase of blood volume in upper body and then causes the increase of venous return. The cardiac function was increased temporarily and induces the peripheral FR. The PLR induces sympathetic activation [10] and the power of LF both in HRV and iPRV also increased. The power in VHF in iPRV increased when venous return increased during PLR. However, the power of HF in iPRV during PLR had different variation with HRV. This is probably due to effect of respiratory frequency. Even respiratory influence, the iPRV spectrum still demonstrated similar variation along with HRV spectrum in LF and HF during HUT and PLR. Besides, the results showed that VHF has potential to indicate the relevant change of venous return and monitor the FR. It is reliable to observe FR in frequency domain analysis during different status of venous return. Predict FR also can be simple and intuitive by data acquisition from PPG and using method of iPRV to analyze.

Though iPRV analysis provided more information of venous return and FR in VHF, it has some limitations. First, it needed to combine with clinical trial (PLR) for observing variation, and it is improper as a real-time application. Second, the data acquisition depends on PPG sensor, which is sensitive and easy to be influenced by body movement and unstable measurement.

However, arterial blood pressure signal can substitute as another source signal for iPRV analysis [3], but the relavant measurement instrument is expensive and is not simple for usage. Third, the iPRV analysis based on the EEMD method which needed to set appropriate parameters and the results are mainly influenced by the parameters setting, which is different while the source signal and the signal properties, such as sampling rate, are different. In addition to these limitations, VHF in iPRV is reliable indicator for FR. The indicator would be examined on non-responders to verify the effectiveness and quantify level of variation of VHF, such as the threshold for the diagnostic reference that evaluate patient who has well FR or not.

V. Conclusion

This study provides an effective indicator for assessing FR in frequency domain. It has potential to be a reliable system for ICU which avoid delaying definitive therapy or additional damage to patient. In the future, this indicator needs to implement on the patients for further exploration.

Acknowledgment

This work was fully supported by the Taiwan Ministry of Science and Technology under grant numbers MOST-103-2218-E-009-016, MOST-103-2221-E-009-139, and MOST 104-2220-E-009-003, and in part by 'Aim for the Top University Plan' of the National Chiao Tung University and Ministry of Education, Taiwan, R.O.C. This work was also supported in part by the UST-UCSD International Center of Excellence in Advanced Bioengineering sponsored by the Taiwan National Science Council I-RiCE Program under grant number NSC-101-2911-I-009-101.

References

[1] M. V. Højgaard, N. H. Holstein-Rathlou, E. Agner, and J. K. Kanters, "Reproducibility of heart rate variability, blood pressure variability and baroreceptor sensitivity during rest and head-up tilt," *Blood Pressure Monitoring*, vol. 10, pp. 19–24, February 2005.

[2] E. Gil, M. Orini, R. Bailón, J. M. Vergara, L. Mainardi, and P. Laguna, "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," *Physiological Measurement*, vol. 31, pp. 1271–1290, August 2010.

[3] C. C. Chang, T. C. Hsiao, and H. Y. Hsu, "Frequency range extension of spectral analysis of pulse rate variability based on Hilbert–Huang transform," *Medical & Biological Engineering & Computing*, vol. 52, pp. 343–351, April 2014.

[4] Z. Wu, and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, pp. 1–41, January 2009.

[5] N. E. Huang, Z. Wu, S. R. Long, K. C. Arnold, X. Chen, and K. Blank, "On instantaneous frequency," *Advances in Adaptive Data Analysis*, vol. 1, pp. 177–229, April 2009.

[6] C.C. Chang, H.Y. Hsu, and T. C. Hsiao, "The interpretation of very high frequency band of instantaneous pulse rate variability during paced respiration," *Biomedical engineering online*, vol. 13, pp. 46–56 , April 2014.

[7] S. Babaeizadeh, et al., "A novel heart rate variability index for evaluation of left ventricular function using five-minute electrocardiogram," *Computers in Cardiology*, vol. 34, pp. 473–476, September 2007.

[8] X. Monnet, et al., "Passive leg raising predicts fluid responsiveness in the critically ill*," *Critical care medicine*, vol. 34, pp. 1402–1407, May 2006.

[9] P. C. Lin, K. C. Hsu, C. C. Chang, and T. C. Hsiao, "Reliability of instantaneous pulse rate variability by using photoplethysmography," *Workshop on Biomedical Microelectronic Translational Systems Research (WBMTSR 2014)*, August 2014.

[10] P. G. Guinot, et al., "Passive leg raising can predict fluid responsiveness in patients placed on venovenous extracorporeal membrane oxygenation, " *Critical Care*, vol. 15, R216. September 2011

Arabic Studies' Progress in Information Retrieval

Essam Hanandeh
Computer Information System, Zarqa University,
Zarqa, Jordan

Hayel Khafajah
Computer Information System, Zarqa University,
Zarqa, Jordan

Abstract—**The field of information retrieval has witnessed tangible progress over the past decades in response to the expanded usage of the internet and the dire need of users to search for massive amounts of digital information. Given the steady increase of Arabic e-content, excellent information retrieval systems must be devised to suit the nature and requirements of the Arabic language. This paper sheds light on the current progress in the field of Arabic information retrieval, identifies the challenges that hinder the progress of this science, and proposes suggestions for further research. This paper uses the descriptive analytical method to examine the reality of Arabic studies in the field of information retrieval and to study the problems that are being faced in this area. Specifically, the previous literature on information retrieval is reviewed by searching the related databases and websites.**

Keywords—Information retrieval; Arabic information retrieval; Indexing; Query reformulation

I. Introduction

The amount of global digital content has been increased by the continuous information flow from websites, company and government records, e-books, e-newspapers, e-magazines, and other online media. Retrieval systems have become imperative for users to extract information from huge amounts of text, images, and digital sounds. Information retrieval refers to the study of searching for information inside documents or for documents themselves [2]. Such a discipline becomes more important as the number of global Internet users increases and their independence on search engines as a major source for information is strengthened [1].

Text information retrieval involves the processing of natural languages and the retrieval of documents that contain the information that is needed by the user from huge databases. Any classical information retrieval system comprises three basic stages, namely, indexing, query reformulation, and matching. First, all extracted documents are indexed by using the best words or expressions that represent or have an actual indication of each document. Second, the query that is entered by the user to access the required information is reformulated to comply with the information retrieval model as well as to add other keywords or modify the weights of the existent words to achieve better search accuracy. Third, the entered query is matched with the existing index, and the most similar documents are retrieved and arranged in a descending order [2, 5, 22, 15]. By determining how documents are represented in the index, information retrieval models can control how the reception is represented. Many information retrieval models exist, of which the most common models include the Boolean model, fuzzy model, and vector space model [2, 5, 22, 15].

Given the increasing amount of Arabic digital content on the Internet and other electronic devices, the need to create information retrieval systems and engines that pay special attention to the peculiarities of Arabic—the language of the Noble Qur'an and Prophet Mohammad's traditions as well as one of the most widespread Semitic languages in terms of native speakers—continues to increase [7]. Arabic differs from English and other languages in several aspects. First, Arabic text is read and written from right to left. Second, Arabic forms vary based on their position and adjacent letters. Third, the diacritics in Arabic change the pronunciation of letters, meaning, and case of words [9]. Fourth, Arabic is a derivative—rather than inflectional—language with one of the most sophisticated morphological systems. This language divides the stems based on a specific set of weights to develop words of different meanings from the same stem. All these considerations present challenges to the mechanization of the morphological, syntactic, and semantic analyses of the Arabic language and to the retrieval of Arabic texts.

II. Types of Indexing

A. Automatic Indexing

In automatic indexing, an index is built to describe the content of each document in the database in a way that best accelerates and facilitates the search process [2]. This index is any kind of data structure that is used for storing words, keywords, or the general description of any document. An information retrieval system depends on matching the query of the user with all the inputs in the index in order to access the documents that are most similar to the query. The difficulty of indexing documents depends on the processed language. In other words, those languages with sophisticated syntactic and morphological systems, such as Arabic, require highly complicated logarithms [13].

The automatic indexing of Arabic texts enjoys the lion's share of the papers in the field of Arabic text retrieval. This type of indexing is divided into pre-indexing processing, stem-finding-based indexing, stem-making-based indexing, indexing based on stem-making and language rules, dictionary-based indexing, taksir plural indexing, and weighing indexing words.

- Stem-Making-Based Indexing

In stem-making-based indexing, the prefixes and suffixes are extracted from words and the stems are used to index documents. The new words tend to have the same meaning because these affixes are often used to indicate definition, number, sex, coordination, or preposition, which removal will not affect the meaning. Previous studies [15, 18, 19, 20] show that stem-making-based indexing outperforms original-word-based indexing, stem-finding-based indexing, and stem-making-context-related-based indexing in terms of precession and recall levels. Such high levels are attributed to the very derivative nature of Arabic, which makes the language highly sensitive to stem making [18].

- Indexing Based on Stem-Making and Language Rules

The indexing based on stem-making and language rules is similar to the stem-making-based method, but employs linguistic rules to obtain better results for the stem-making process. A recent study [13,21] shows that this method outperforms the others in terms of stem-making accuracy. Nevertheless, no experiment has been conducted to merge this type of indexing with an Arabic text retrieval system to measure its efficiency.

- Dictionary-Based Indexing

In dictionary-based indexing, each word in the document is indexed by using synonyms [13]. A study on the retrieval of Qur'anic verses shows that this method has a higher retrieval accuracy than the stem-finding-based technique. Another study [18] shows that this method increases the competence of an information retrieval system for Arabic texts by 18%.

- Pre-Indexing Processing (Normalization)

Pre-indexing process is an important stage to obtain the optimal results for the indexing process; this stage involves the removal of diacritics, letters, and stop words that do not have independent meanings [8] as well as the unification of the forms of letters. For instance, the varieties of the Arabic (ألف) letter (أ، ا، إ، آ) are all made [1]. The same applies to the (هاء) varieties (ة، ه), which are both made (ه), and the (ياء) varieties (ي، ئ، ى), which are made (ي) as in [16, 19]. Such shifts are proven successful in improving the retrieval of Arabic texts, which can be attributed to the fact that original texts do not consider the differences between these letters because of the weak Arabic writing language of those who enter such texts.

- Roots-Finding-Based Indexing

In roots-finding-based indexing, the roots are extracted from the document to be used as terms. All words with the same roots will be indexed under the same word even though they may not necessarily have the same meaning. This method has been investigated in many papers [19, 3] and its excellence over original-word-based indexing has also been proven. This technique has achieved high levels of precession and recall in those sets that contain limited or unchangeable numbers of documents, such as those of Qur'anic verses or Prophet Mohammad's traditions. Such high levels are attributed to the fact that this method retrieves all documents that contain any morphological form of the query words, thereby increasing the possibility of finding the required information. However, this technique is impractical in cases of huge and continuously renewed sets, such as those of the Internet. This technique also expands the search scope without providing the user with his/her target.

- Taksir Plural Indexing

Returning the taksir plurals to their original singulars presents a challenge to the Arabic language in general and to the retrieval of Arabic texts in particular. Unlike regular male and female plurals, taksir plurals are not immediately recognized from the text. Various infixes can also be used. Previous research [18] has attempted to address this problem by employing the n-gram technique, but this technique has been proven insufficient. Another study [25] has used a dictionary that lists the singular forms of the taksir plurals to recognize the words. Previous studies have proven that indexing techniques that bring back the taksir plurals to their original singulars outperforms the other indexing techniques.

- Weighing Indexing Words

In weighing indexing words, each term is given a weight that best fits the extent to which the word represents its origin document. Previous research [25,27] has investigated the effects of removing letters or stop words and using various types of weighing indexing words on the retrieval of Arabic texts. The OKAPI BM25 technique and the removal of the stop words can lead to better retrieval results than can the other weighing techniques, such as term frequency-inverse document frequency (tf-idf) and the relevance value of a document with respect to a query that measured by the Kullback-Leibler (KL) divergence between the query model and document model. In addition, when the text is not edited or when no words are removed, the prominent tf-idf method is considered the optimal technique.

Another study [21] explores 12 weighing techniques based on three factors, namely, the number of times the word is repeated in the document, the number of times the stems of such words are found, and the distribution of the word in the document. This technique has been proven efficient in terms of precession and recall.

B. *Automatic Query Reformulation*

Query reformulation is an information retrieval technique that is applied for adding and/or re-weighing query words to obtain the largest number of matching documents. Query reformulation can be conducted in three ways, namely, relevance feedback, automatic local analysis (inductive query by example), and automatic global analysis [2]. The automatic reformulation of Arabic queries has been investigated in many studies over the past decade.

- Relevance *Feedback Query Reformulation*

In relevance feedback, the user is requested to determine whether the retrieved documents are relevant to his/her query. Accordingly, the query is reformulated by adding words that are mentioned in relevant documents, by removing words that are found in irrelevant documents, or by re-weighing the terms. The new query is entered in the information retrieval system to retrieve another set of documents that may be more relevant. This method is sometimes repeated until the user is satisfied with the results.

In a related study, the user is asked to classify the retrieved documents as relevant or irrelevant. The user is also requested to choose synonyms to the appropriate terms from a dictionary and then include these synonyms in his/her new query. If the added synonyms are highly relevant to the original terms, such an interactive method for investigating the meanings of words and expanding the query can lead to satisfactory results in terms of precession and recall. However, such results cannot be obtained if the synonyms have a general nature.

Furthermore, an experiment-based study [18] shows that expanding the query by such an interactive way (relevance feedback by the user) outperforms the automatic method (automatic local analysis) in terms of retrieval efficiency. Using either of these methods is better than any using other techniques for reformulating and expanding the query.

- *Automatic Local Context Analysis Query Reformulation*

Automatic local context analysis query reformulation, also called inductive query by example, provides the user an information retrieval system with a set of documents that are either relevant or irrelevant to his/her query. The system then deduces the main words from the relevant documents and sometimes excludes irrelevant words from a query in order to access other relevant documents [14]. However, this method is only employed with frequent queries instead of single-time queries [9].

Authenticity [11] is a major Arabic text retrieval system that is based on the Prophet's traditions. This system identifies the roots of the words that are used in the query and matches them with a roots-finding-based index to produce an initial list of documents. Afterward, automatic local context analysis is used to reformulate the query. After application to one of the queries, the method has yielded 0.66 and 0.80 precession and recall scores, respectively. The success depends on the set of documents to which the method is applied. This method is more appropriate for a highly limited and unchangeable set because the search results can somehow be limited. By contrast, this method is less efficients for larger sets. Specifically, the precession and recall levels are lowered as the scope of the search is significantly expanded.

- Automatic *Global Analysis Query Reformulation*

Unlike the previous two methods, automatic global analysis query reformulation establishes a relation among all terms for all documents in the set and not only between the relevant and irrelevant documents. Most of the techniques attempt to build a dictionary of similarity to determine the relation between terms according to the concept that they represent and not only their simultaneous existence in the same document [2].

Many studies have investigated the application of this method to Arabic text retrieval. For example, the Arab search engine Barq [17] depends on the automatic or manual addition of new query words on three concept dictionaries and on the unification of forms of letters as mentioned above in automatic indexing. This method has increased the precession measure to 75%. Mustafa et al. [28] propose a method for expanding the query by finding synonyms to terms and their derivations. The Neuro-Fuzzy logic has been adopted to obtain the closest derivations to the meaning of the original terms, thereby providing the user with options to expand the query. Researchers have conducted further experiments to prove the efficiency of the method in text retrieval. Another study [7] attempts to expand the query to retrieve information from an Arabic text with or without diacritics. The same method has been applied to the Noble Qur'an by using four types of indexes, namely, index for words with diacritics, index for words without diacritics, root-finding-based index, and synonym-set-based index. He then compares the stem-finding-based index with the query-expansion-based index and finds that the latter outperforms the former in terms of average accuracy.

Another study [28] proposes a modification to the concept-based query expansion—introduced in [30]— to remove the irregular values that are generated by the presence of a very similar word that outshines the less similar ones. This method has improved the retrieval system efficiency by 3.3%.

C. Matching Function Adjustment

In matching function adjustment, the entered query is matched with the index to retrieve documents that are identical to the query. Such documents are called relevant documents that arranged in a descending manner according to their relation to the subject. When designing the matching function, which matches the query with the index, the following must be considered: (1) how to decide whether the provided document is relevant, and (2) how to arrange the relevant documents according to their relevance or ranking [6]. The matching function efficiency depends on several external factors, such as the size of the document set, subject of the document, and culture of the user that has formulated the query [6]. Therefore, unless used in all the information retrieval systems, a particular matching function cannot be proven as successful.

Only few studies have investigated the matching of Arabic texts with the measures of similarity to be used in the field of Arabic information retrieval. One of these studies [28] have explored the efficiency of the n-gram technique in matching and retrieving Arabic text. They have successfully applied such technique with other languages, such as English, because of the highly derivative nature of Arabic, which words also contain infixes. In another study, the n-gram technique is modified to suit the Arabic language. Specifically, the non-consecutive letters of a word are selected and matched them with the letters of other words. In addition to taking the prefixes and suffixes from the stem, the modified technique yields better results than the classical technique. The same technique has been modified by other scholars [8,23] to fit the Arabic language searching in specific locations of the target word. Such modifications aim is to increase the possibility of finding a significant degree of similarity between two words that do not hold the same concept. The modifications outperform the classical methods in terms of precession and recall. These modifications also help find high degrees of similarity among different derivations of the word.

In a recent study [24], researchers build an information retrieval system in Arabic according to the Fuzzy model, believing that this system suits the nature of the Arabic language and can discover the similarity between various synonyms and different sentence structures. This system is based on two dictionaries, namely, one with a matrix that indicates the relation among all words (correlation) and one for synonyms. To determine the similarity between two sentences, the correlation is calculated between each word and each sentence in which the word is found. Afterward, the similarity between the two sentences is calculated. This system outperforms those information retrieval systems that are based on the Boolean model in terms of precession and recall, thereby proving that the former can detect similarities between similar documents yet requires costly and complicated calculations.

D. Automatic Documents Classification

In the field of information retrieval, if the documents of the same set react similarly to a query [1], then they are classified accordingly. In other words, if one document in a certain set is relevant to a certain query, the rest of the documents in that same set tend to be classified as relevant. Based on the sets of documents to be classified and the aspects of information retrieval to be improved, several applications for automatic classification can be divided into two types. In the first type, the search results are classified in a particular point or in the entire set of documents. In the second type, the classification is performed to improve the interface or experience of the user as well as the efficiency of the search system [1].

Only few studies have classified Arabic documents for the purpose of information retrieval. One of these studies [25] perform a classification based on the Naive Bayes logarithm to create an index of subjects that can facilitate the search process. The documents are divided into five main subjects, namely, sport, business, culture and arts, science, and health. Before the classification process, the diacritics are removed and the stems are identified. The classification accuracy reaches 68.78%. Other scholars [19] propose a logarithm for the automatic classification of Arabic documents by finding those words that cover the main concept of each document subject. Each word is weighed based on the number of times it is repeated and to its locations in the documents. The above classification logarithm enhances the efficiency of the information retrieval system.

In [13] and [28], the efficiency of two logarithms in splitting the text is measured, and these logarithms have been proven successful in both English and Arabic. TextTilling and C99 have excellent application in Arabic, with the former outperforming the latter.

E. Web Page Automatic Search

Crawlers are programs that track hyperlinks on the web, gather pages, and make these pages available to search engines for indexing. These programs are often given URLs or keywords, track the hyperlinks on these webpages, and then move to other pages [6, 25]. Searching in webpages represents a significant challenge because of their large number, which increases on a momentary basis. In addition, given that their contents continue to change, the webpages that are visited earlier must be found and stored to be re-visited and indexed later. The changes in a webpage are unstable and vary according to the type of websites. Webpages can be stored in the following ways [13]:

- *Uniform Policy: All previously indexed webpages are updated whether their contents have been changed.*
- *Proportional Policy: The webpages are updated according to their average change.*
- *Optimal Policy: Only those webpages with trackable changes can be updated.*
- *Curve Fitting Policy: The calculation covers the changes between two consecutive images of the webpage and the number of changes as reflected in the change date.*

The Arabic context remains in its early stage. According to [12,13], Arabic webpages only account for 0.1% of the total webpages, which explains the lack of research on the Arabic language. Another study [5] modifies the curve fitting policy to suit the Arabic language by omitting pronouns, relative pronouns, and prepositions from the content without changing the meaning. They also take the various derivations of the same word with the same meaning. Such modification has reduced time and space, which are important factors in searching for webpages. In another study [13], to search for webpages in Arabic and other languages, a program is distributed to more than one server to enhance speed and efficiency. The speed can reach 160 webpages per second.

III. Conclusion

Information retrieval in Arabic has witnessed tangible progress over the last decade. Specifically, the Arabic document set has provided researchers with a huge number of data. This research has used two sets, first set is published by Saad [29] contains queries and documents that was collected from CNN Arabic website, and second set is BBC Arabic corpus, which has been collected from BBC Arabic website. However, these documents set has several flaws, such as limited syntactic structures, forms of nouns, and verbs as well as many misspelled names of people and non-Arabic places.

Furthermore, given the importance of stem-making for Arabic information retrieval systems, researchers must build an efficient, accurate tool for the stem-making of Arabic words that pay special attention to taksir plurals. Those texts with diacritics must also be reconsidered, and the presence of diacritics must be utilized in disambiguating the meanings of words before starting the indexing process. Differentiation must also be performed between limited, near-constant texts, such as the Noble Qur’an and Prophet Mohammad’s Hadith traditions, and huge, continually changing texts, such as webpages. The future work of this research coud be provided and investaged syntactic structures, and forms of nouns of Arabic languge by utilizing disambiguating of words meanings before starting the indexing process.

Acknowledgment

This research is funded by the Deanship of Research and Graduate Studies in Zarqa University /Jordan

Reference

[1] A. Abdelali, J. Cowie, H. Soliman, "Arabic information retrieval perspectives", In Proceedings of JEP-TALN 2004 Arabic Language Processing, 2004.

[2] A.Alhroob, H. Khafajeh , N. Innab, 2013. Evaluation of different query expansion techniques for Arabic text retrieval system. Am. J. Applied Sci., 10: 1018-1024.

[3] M. AL-Kabi, H.Wahsheh, I.Alsmadi, (2014). A Topical Classification of Hadith Arabic Text, IMAN 2014: 2nd International Conference on

Islamic Applications in Computer Science and Technologies, 12th – 13th October 2014, Amman, Jordan, pp. 1-8.

[4] D. Kraft, F. Petry, B. Buckles, T. Sadasivan, “The use of genetic programming to build queries for information retrieval,” Evolutionary Computation, 1994 search. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, 1994, pp. 468-473 vol.1.

[5] D. Ezzat, M. Abdeen, M.F. Tolba, "A Memory Efficient Approach for Crawling Language Specific Web: The Arabic Web as a Case Study," icime, pp.584-587, 2009 International Conference on Information Management and Engineering, 2009.

[6] D. Manning, P. Raghavan, and, H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[7] E. Hanandeh, “SIMILAR THESAURUS BASED ON ARABIC DOCUMENT: AN OVERVIEW AND COMPARISON,” International Journal of Computer Science, Engineering and Applications (IJCSEA), Vol.3, No.2, April 2013

[8] E. Hanandeh, K. Mabreh. "EFFECTIVE INFORMATION RETRIEVAL METHOD BASED ON MATCHING ADAPTIVE GENETIC ALGORITHM "Journal of Theoretical and Applied Information Technology, 30 th November 2015 –Vol. 81. No. 3 - 2015

[9] F. Ahmed , A. Nürnberger, “N-grams Conflation Approach for Arabic”, ACM SIGIR Conference, 2007.

[10] F. Ataa Allah, S. Boulaknadel, A. El qadi, D. Aboutajdine, “Arabic Information Retrieval System Based on Noun Phrases,” Information and Communication Technologies, 2006. ICTTA '06. 2nd, 2006, pp. 1720-1725.

[11] F. Harrag, A. Hamdi-Cherif, E. El-Qawasmeh, “Vector space model for Arabic information retrieval — application to “Hadith” indexing,” Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the, 2008, pp. 107-112B.

[12] G. Kanaan, R. Al-Shalabi, M. Sawalha, "Improving Arabic Information Retrieval Systems Using Part of Speech Tagging", Information Technology Journal, vol.4, 2005, pp.32-37.

[13] G. Kanaan, R. Al-Shalabi, M. Ababneh, A. Al-Nobani, “Building an effective rule-based light stemmer for Arabic language to improve search effectiveness,” 2008 International Conference on Innovations in Information Technology, Al Ain, United Arab Emirates: 2008, pp. 312-316.

[14] H. Khafajeh, A. Abu-Errub, A. Odeh, N. Youse, (2012) NOVEL AUTOMATIC QUERY BUILDING ALGORITHM USING SIMILARITY THESSAURUS, American Journal of Applied Sciences 9 (9): 1373-1377, ISSN 1546-923.

[15] H. Khafajeh, N. Yousef, (2013) Evaluation of Different Query Expansion Techniques by using Different Similarity Measures in Arabic Documents, International Journal of Computer Science Issues, Vol 10, Issue 4, No 1, July 201, (p.p. 160-166) .

[16] I. El Emary, J. Atwan, "Designing and building an automatic information retrieval system for handling the Arabic data", American Journal of Applied Sciences, 2005.

[17] J. Mayfield, P. McNamee, C. Costello, C. Piatko, A. Banerjee, “JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video and Web retrieval”, InTREC 2001 Proceedings, 2001.

[18] J. Xu, A. Fraser, R. Weischedel, “Empirical studies in strategies for Arabic retrieval,” Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland: ACM, 2002, pp. 269-274.

[19] L. Larkey, L. Ballesteros, M. Connell, “Light Stemming for Arabic Information Retrieval,” Arabic Computational Morphology, 2007, pp. 221-243.

[20] M. Aljlayl, O. Frieder, "On arabic search: improving the retrieval effectiveness via a light stemming approach", CIKM 2002, pp.340-347.

[21] N. Mansour, R.A. Haraty, W. Daher, M. Houri, “An Auto-indexing Method for Arabic Text,” Information Processing and Management: an International Journal., vol. 44, 2008, pp. 1538-1545.

[22] N. Yousef, A. Abu-Errub, A. Odeh, H. Khafajeh, AN IMPROVED ARABIC WORD’S ROOTS EXTRACTION METHOD USING N-GRAM TECHNIQUE, Journal of Computer Science 10 (4): 716-719, 2014, ISSN: 1549-3636, © 2014 Science Publications, doi:10.3844/jcssp.2014.716.719 Published Online 10 (4) 2014

[23] P. Pathak, M. Gordon, Weiguo Fan, “Effective information retrieval using genetic algorithms based matching functions adaptation,” System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, 2000, p. 8 pp. vol.1.

[24] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, Wokingham, UK, 1999.

[25] S. Alzahrani, N. Salim, "On the Use of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents", Proceedings of the 5th Postgraduate Annual Research Seminar, UTM, pp.256-260, 2009.

[26] S. Boulaknadel, B. Daille, A. Driss, "Multi-word term indexing for Arabic document retrieval", Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008), Marrakech, Morocco 2008.

[27] V. Shkapenyuk, T.Suel, (2002), 'Design and implementation of a high-performance distributed web crawler', In Proc. of the Int. Conf. on Data Engineering.

[28] S. Mustafa, Q. Al-Radaideh, "Using N-Grams for Arabic Text Searching", Journal Of The American Society For Information Science And Technology, vol.55, pp.1002–1007, 2004.

[29] www.sourceforge.net/projects/ar-text-mining

[30] Y. Qiu, H. Frei, "Concept based query expansion". In proceedings of the 16th International ACM SIGIR Conference on R & D in Information Retrieval, ACM Press, New York, 1993, pp. 160-169.

Preliminary Study of Software Performance Models

Issamjebreen
Faculty of Information Technology
Zarqa University, Zarqa, Jordan

Mohammed Awad
Faculty of Applied Engineering
University of Palestine

***Abstract*—Context: Software performance models can be obtained by applying for specific roles, skills and techniques in software life cycle, and it depends on formulating the software problem as well as gathering the performance requirements. This paper presents a preliminary review of the software performance models. This constitutes a reference for the IT companies and personnel that help them select the suitable model for their projects. Also, the study helps researchers find out further research areas in this field. A preliminary review according to a predefined strategy is used to conduct previous approaches of software performance models integrated with software development cycle in early software cycle. A review has been done for exploring and comparing the software performance models that are published previously. This study results in a comprehensive review for the existing software performance models. This review composes a clear reference for highlighting the weak and strength points of these models.**

Keywords*—*Performance Models; Measurement Model; Performance Prediction; Performance evaluation

I. Introduction

Developing software requires ensuring that software performance requirements are considered and achieved. Software performance is a process to predict and evaluate if the system meets business goals. Performance Predictive models (Model-Based) require detail descriptions at run-time behavior of a system, in order to estimate the execution time and other performance issues i.e. cache misses. It is used by architects to avoid the performance problems at system implementation time, and to estimate designs, and to explore a new optimization by compiler writers. In addition, developers can adjust their programs. Conversely, Evaluation models (measurement models) attempts to measure the system performance activities when the system has been implemented. In order to defined performance problems and bottlenecks.

This paper presents a preliminary review of the software performance models. The main goal is to presents explorations of research of performance models as well as clarifying the variance of elements used for each model. In order to help researchers to find out further research areas and to constitute a reference for the IT companies and personnel that helps them select the suitable model for their projects.

The rest of this paper is organized as follows: Sec2: Research Description and Presentation; Sec3: Literature Review, Sec4: Preliminary Results and Discussion; Sec5: Conclusion and Future Work.

II. The Review Process

This section declares the review process, research framework and the objectives of the preliminary review. This paper aims to present a preliminary review between “Performance models”, in order to clarify the elements used to generate these models. The framework of this review considers studies of software performance based on simulation, Analytical methods, and component based approach in order to explore the elements used to generate the performance models.

A. *Research Planning*

The research Strings was established by academic as following: (“Software Performance Engineering, Modeling Techniques, UML, Performance Models.”), the Research question (RQ) is: RQ. What are the elements used to generate the performance models? Our research resources namely: IEEE and ACM Digital Library.

B. *Conducting the Research*

The criteria for determining whether a study should be included as a related study (named “Primary Study”) or not, was by first analyzing research titles, abstracts, keywords and introductions from the studies retrieved through search.

C. *Selection of the Primary Studies*

The inclusion criteria for the selection of primary studies are listed below:

- Studies that proposed Performance models”.
- Studies that describe their methods in details.

The exclusion criteria for the selection of primary studies are listed below:

- Studies that don’t answer the research question.
- Studies that don’t present Models OR Meta-Models of software performance.

III. Primary Studies

Smith & Williams [1], who have defined SPE (Software Performance Engineering) information requirements, have proposed integrated software development cycle with performance models; They defined information requirements for Early Life-Cycle performance analysis, the performance analysis according to authors were: performance objective, performance scenario that includes software plans + workload, execution environment, resource requirement and processing overhead.

They starting building the Meta model SPE from the performance scenario, using the workload to describe the ratio of different types of requests; SW plans defined the execution path for each workload, also used class diagrams to define the objects and the relationships between them, and thus modeling the performance scenarios using a form Execution Graph EG. That enables transferring information between CASE Tools and performance model. Also Smith et al, [2] updated the SPE meta-model that is proposed by Smith & Williams by adding subclass to processing node and adding project, facility node; then applying XML formats to Software Performance Model Interchange Format (S-PMIF meta-model) and export UML Diagrams when they are ready into the S-PMIF.

Additionally, Henia, Rafik, et al. [3] proposed an approach named SymTA/S, which is considered as system-level performance as well as timing analysis, and based on formal scheduling analysis techniques, in order to support diverse Architectures & task dependencies & collect optimization algorithms with analysis of rapid design. Support performance issues such as bus, processor utilization, and worst case scheduling scenarios. Moreover, D'Ambrogio, Andrea [4] presented a framework that aims to transfer source UML of software models into performance prediction models layered queuing networks LQNs, which required understanding the syntax & semantics for the source and P models. That enhances software designer’s productivity as well as software quality.

Smith, Connie U., & Lloyd G. Williams [5] said that most performance anti pattern problems result through the architecture/design stages, unfortunately these problems don’t appear during the implementation stage. The solutions need software changes opposed to system tuning changes. Smith, Connie U., & Lloyd G. Williams [5] presented three new performance anti patterns and gave examples to illustrate them. These anti patterns help developers and performance engineers avoid common performance problems.

Woodside, Murray, et al. [6] analyzed the exchange information provided from a performance model and the process of creating a performance model. They proposed PUMA transformations that define Performance evolutions from annotated UML Profile for Modeling and Analysis of Real-Time Embedded System MARTE. This approach enables to obtain performance measures such as throughout and response time throughout software life-cycle. Moreover, Sim, Jaewoong, et al. [7] proposed a framework in order to analyze the performance, which supports shed light of bottlenecks of GPGPU applications. In addition, this framework helps GPGPU Profile tools and supports programmers in measurements as well as metrics during run time.

Bammi, Jwahar R., et al. [8] proposed two approaches for handling issues of performance evaluation as well as SW cost for embedded system design. The first approach is called Source-Based approach which employs the integration of a virtual instruction set in order to evaluate the performance. The second approach called object-based approach which translates the assembler created by the target compiler (named assembler-level).

Lindemann, Christoph, et al. [9] proposed a framework for performance estimation which enables designers to predict performance during variance stages at design phase. They presented an algorithm that supports the state space creation resulted from State & Activity diagrams. In order to enable a quantitative analysis for Stochastic Process and Generalized Semi-Markov Process GSMP is used. Additionally, Denaro, Giovanni, Andrea Polini, and Wolfgang Emmerich [10] proposed an approach for performance testing in particular for distributed systems during early life cycle phases. They created test cases to examine these systems starting from architectural designs. They observed that middleware functionality e.g. transactions & remote communication primitives control these systems.

Bertolino, Antonia, and Raffaela [11] presented an approach named CB-SPE for component-based SW performance, which adopted CB (Component-Based) framework to model the standard RT-UML Profile restructure depending on CB Role. CB-SPE approach applied on both component layer (parametric performance estimation) and application layer (predictive performance for assembled components).

Tribastone, Mirco, and Stephen Gilmore [12] proposed a procedure to systematically map activity diagrams into stochastic process algebra model referred as PEPA Models. PEPA performance model clarifies a Markov [9] in semantics to enable the computation of performance issues i.e. workload, response time and the throughput. They are concerned about tools produced in Eclipse platform; to enable transfer from MARTE annotated UML activity diagrams into PEPA Models

Gu, Gordon P., and Dorina C. Petriu [13] present a method that enables transfer between the results annotated from UML with performance models, which is generated at a higher level of abstraction. They use a lower level XML trees manipulations i.e. XML algebra. They use also LQN to apply their method, which can be designed to other performance model formalisms. Moreover, Zheng, Gengbin, et al, [14] proposed a performance predictive model for big weigh computers (i.e. blue gene machine), that include a parallel simulator, bigsim, bignetsim. The simulator can deal with advanced features of modeling, also supports performance predictions for huge machines. In addition, Kähkipuro, Pekka [15] proposed a framework, in order to introduce performance modeling; at first they have explained an overview of the proposed framework and clarified the major components for this framework. After that they have clarified relationships between these components.

Finally, Zolfaghari, Rahmat [16] presented a method for transforming UML of SW architecture to QNM (Queuing Networks Model). In order to support performance as well as quality of the models that employ UML in designing software. They have used the deployment diagram in SW components with hardware resources. The activity diagrams extract the system behaviors and the use case diagrams extract workloads.

IV. Results and Discussion

Table 1 shows the Advantages & Limitations of Performance Models. Regarding Data Extraction, this research has predefined Database that contains Authors, Titles, Published Years, descriptions and summaries of this comprehensive review. For the Evidence Synthesis, most approaches employ the results annotated from UML, SPT profiles to integrate software models with performance models through high abstract level information. The Use Case diagrams are used to describe workload density, and behaviors, while the Activity and Sequence diagrams are used to extract computations of a system performs service requests to the devices resources (the dynamic behavior). Deployment Diagram provides hardware recourses such as passive and active resource modeling. Processing resources extract from Active resources (devices), while operating system processes extract from Passive resources.

TABLE I. ADVANTAGES & LIMITATIONS OF PERFORMANCE MODELS

Ref.	Advantages	Limitations
[1]	Provides an interchange format that enables CASE & Performance Tools to exchange information	Considers Class & ER Diagrams only.
[2]	Provides interchange formats that support flexibility in when & who performance specifications are provided	Needs SPT Profile to export the resource requirements Considers Class Diagrams only
[3]	Supports diverse Architectures & task dependencies Supports performance issues such as bus, processor utilization, and worst case scheduling scenarios.	Provides only a system-level performance
[4]	Transfers source UML of software models into performance prediction models layered queuing networks LQNs Enhances software designer's productivity as well as software quality	which required understanding the syntax & semantics for the source and Pmodels
[5]	Provides three new performance anti patterns. These anti patterns help developers and performance engineers avoid common performance problems	The solution need software changes opposed to system tuning changes
[6]	Provides PUMA transformations that define performance evolutions from annotated UML Profile for Modeling and Analysis of Real-Time Embedded System MARTE. Enables to obtain performance measures such as throughout and response time throughout software life-cycle	Focuses only on Real-Time Embedded System MARTE
[7]	Supports shed light of bottlenecks of GPGPU applications. Supports programmers in measurements as well as metrics during run time	it assumes that a memory instruction is always followed by consecutive dependent instructions; hence, MLP is always one. it assumes that there is enough instruction-level parallelism. Thus, it is difficult to predict the effect of prefetching or other optimizations that increase instruction/memory-level parallelism.
[8]	Employs the integration of a virtual instruction set in order to evaluate the performance. Translates the assembler created by the target compiler (named assembler-level).	it is quite difficult to account for potential compiler optimizations that do not fall into any of the Virtual Instruction categories
[9]	Enables designers to predict performance during variance stages at design phase. Supports the state space creation resulted from State & Activity diagrams.	Limited to Time-enhanced UML Diagrams
[10]	Enables test cases to examine systems starting from architectural designs	Cannot identify performance problems that are due to the specific implementation of late-available components. For example, if the final application is going to have a bottleneck in a business component that is under development, the approach has no chance to discover the bottleneck that would not be exhibited by a stub of the component. Performance analysis models remain the primary reference to pursue evaluation of performance.
[11]	Applied on both component layer (parametric performance estimation) and application layer (predictive performance for assembled components)	it leads to sound results only for a specific platform
[12]	Enable the computation of performance issues such as workload, response time and the throughput	Restricted to final node activities
[13]	Enables transfer between the results annotated from UML with performance models, which is generated at a higher level of abstraction. Uses a lower level XML trees manipulations such as XML algebra. Uses also LQN to apply their method, which can be designed to other performance model formalisms	Cannot build the complete behavior for every component
[14]	Provides a performance predictive model for big weigh computers (i.e. blue gene machine), that includes a parallel simulator, bigsim, bignetsim.	Large meshes must be generated, which is difficult with today's tools. The meshes must be partitioned for parallel execution
[15]	Introduces performance modeling Clarifies the major components and relationships between these components	the approach limits the available scheduling disciplines, service time distributions, and arrival rate distributions
[16]	Supports performance as well as quality of the models that employ UML in designing software. Uses the deployment diagram in SW components with hardware resources.	Must have both software and hardware components to follow it. It is not perfect for pure software solutions

V. Conclusion and Future Work

The results of this review show that most approaches widely used UML Diagrams and SPT Profiles to support generation of Performance Models. There are several performance models introduced to provide analytical assessment. These models aim to help designers and architects to predict performance measurements at different steps.

This review has revealed that data pre-processing has received considerable attention in the Software Engineering research community. The same cannot be said regarding data collection procedures and the identification of data quality issues, which can compose future research topics.

Acknowledgment

This research is funded by the Deanship of Research and Graduate Studies in Zarqa University /Jordan

References

[1] Williams, L., Smith, C., "Information Requirements for Software Performance Engineering", the National Science Foundation, 1995

[2] Smith, C., Lladó. C., Cortellessa, V., Marco, A., Williams, L., " From UML models to software performance results: An SPE process based on XML interchange formats", WOSP, 2005

[3] Henia, Rafik, et al. "System level performance analysis–the SymTA/S approach." IEE Proceedings-Computers and Digital Techniques 152.2 (2005): 148-166.

[4] D'Ambrogio, Andrea. "A model transformation framework for the automated building of performance models from UML models." Proceedings of the 5th international workshop on Software and performance. ACM, 2005.

[5] Smith, Connie U., and Lloyd G. Williams. "More new software performance antipatterns: Even more ways to shoot yourself in the foot." Computer Measurement Group Conference. 2003.

[6] Woodside, Murray, et al. "Transformation challenges: from software models to performance models." Software & Systems Modeling 13.4 (2014): 1529-1552.

[7] Sim, Jaewoong, et al. "A performance analysis framework for identifying potential benefits in GPGPU applications." ACM SIGPLAN Notices. Vol. 47. No. 8. ACM, 2012.

[8] Bammi, Jwahar R., et al. "Software performance estimation strategies in a system-level design tool." Proceedings of the eighth international workshop on Hardware/software codesign. ACM, 2000.

[9] Lindemann, Christoph, et al. "Performance analysis of time-enhanced UML diagrams based on stochastic processes." Proceedings of the 3rd international workshop on Software and performance. ACM, 2002.

[10] Denaro, Giovanni, Andrea Polini, and Wolfgang Emmerich. "Early performance testing of distributed software applications." ACM SIGSOFT Software Engineering Notes. Vol. 29. No. 1. ACM, 2004.

[11] Bertolino, Antonia, and Raffaela Mirandola. "Towards component-based software performance engineering." Proceedings of the 6th ICSE Workshop on Component-Based Software Engineering. 2003.

[12] Tribastone, Mirco, and Stephen Gilmore. "Automatic extraction of PEPA performance models from UML activity diagrams annotated with the MARTE profile." Proceedings of the 7th international workshop on Software and performance. ACM, 2008.

[13] Gu, Gordon P., and Dorina C. Petriu. "From UML to LQN by XML algebra-based model transformations." Proceedings of the 5th international workshop on Software and performance. ACM, 2005.

[14] Zheng, Gengbin, et al. "Simulation-based performance prediction for large parallel machines." International Journal of Parallel Programming 33.2-3 (2005): 183-207.

[15] Kähkipuro, Pekka. "UML based performance modeling framework for object-oriented distributed systems." «UML»'99—The Unified Modeling Language. Springer Berlin Heidelberg, 1999. 356-371.

[16] Zolfaghari, Rahmat. "Software Performance Evaluation with Converting UML Description of Software Architecture to QNM." Int. J. Emerg. Sci 3.3 (2013): 268-27

Implementing Project Management Category Process Areas of CMMI Version 1.3 Using Scrum Practices, and Assets

Ahmed Bahaa Farid
Faculty of Computers and Information
Helwan University
Cairo, Egypt

A. S. Abd Elghany
Faculty of Business Administration
Higher Technological institute
10th Ramdan, Egypt

Yehia Mostafa Helmy
Faculty of Commerce & Business Administration
Helwan University
Cairo, Egypt

Abstract—**Software development organizations that rely on Capability Maturity Model Integration (CMMI) to assess and improve their processes have realized that agile approaches can provide improvements as well. CMMI and agile methods can work well together and exploit synergies that have the potential to improve dramatically business performance. The major question is: How to realize the integration of these two seemingly different approaches? In an earlier work, we have conducted a field study within six companies. These companies worked with agile methods for years and the Egyptian Software Engineering Competence Center (SECC), which is the regional CMMI appraisal center, assessed them. This study was mainly conducted to enhance the empirical understanding in this research field. Additionally, it showed that companies usually don't use agile in a good way that helps in covering the CMMI specific practices. In this paper, we present a new approach for mapping between CMMI and Scrum method. This mapping has been analyzed, enhanced, and then applied to the same companies. Putting in considerations that other previous efforts have worked in the same context but for an older version of CMMI, our research is using the latest CMMI version, which is 1.3. The research shows that our mapping approach has resulted in 37% satisfaction and achieved 17% partial satisfaction for CMMI specific practices. This resembles 19.4% enhancement in the satisfaction, and 6.2% improvement in the partial satisfaction against the previous related research effort that was already not targeting the latest CMMI version.**

Keywords—Software Engineering; Scrum Software development; Process Improvement; CMMI; Scrum; Scrum CMMI Mapping; Project Management CMMI Process Areas; CMMI-Dev version 1,3; CMMI Project Management Category

I. Introduction

Quality management systems (QMS) like ISO 19011, CMMI or SPICE, have been a quite popular in the software industry since many years ago. This popularity comes from the fact that the improvement of development processes is empirically linked to the improvement of software quality. However, the documentation and formalization-overhead of those frameworks are massive. Many companies suffer from this overhead and desire more efficiency in development and project management while maintaining the high quality of their software products [1].

Lately, a new approach to the software development has gained a foothold in the software industry, which is called agile development. The agile manifesto [2], which was written in 2001, derived the term agile development. Agile processes have been proposed to overcome the flexibility issues of traditional procedures. Agile development methodology is an umbrella term that describes several agile methods such as Scrum, XP, ASD, Crystal, FDD, DSDM, and Lean. Most of the agile methods promote development iterations, teamwork, collaboration, and process adaptability throughout the life-cycle of the project [3]. Agile practices have been criticized for a lack of discipline and argued of being suitable only for some particular types of projects [4].

On one hand, companies that assess and improve their processes based on CMMI are now realized that agile approaches can provide simplification, improvements as well. On the other hand, there are increasing numbers of agile champion companies, which are looking for more structured processes [5]. For many software engineering schoolers, Agile methods and CMMI best practices are often perceived to be at odds with each other at first glance. This wide gap between two options is not the accurate picture. In 2008, the Software Engineering Institute (SEI) published a technical report that clarifies two reasons behind this inaccurate picture in some engineers' minds. The first reason is the early adopters of both approaches came from different software development paradigms. The second reason lies in the misuse of both perspectives that resulted in misperceptions in both camps about the another [6]. Therefore, the question has to be raised if these two seemingly different approaches could be combined with each other and if the combination brings more benefits than either one alone.

In [7] we have explained a conducted empirical study to check the value of simplifying CMMI version 1.3 through using some agile practices while enhancing it with non-agile ones. The Study primary objective was to increase the understanding of CMMI and agile integration and to explore the reconciliation of these two approaches. The previously mentioned research showed that Scrum practices (as one of the Agile methods) with non-scrum improvements could satisfy the CMMI process areas' specific practices. This means that an organization that uses Scrum without any further improvements would not be able to achieve capability level 2 on Project Planning (PP), Project Management and Control (PMC), Requirements Management (REQM), Risk Management (RSKM), and Integrated Projected Management (IPM). However, Scrum could be used, and enhanced with other agile techniques. To make this possible, some objectives have to be achieved. First, Scrum practices have to be mapped to the CMMI specific practices; one-by-one to have a better understanding of the relation between CMMI and Scrum. To achieve the latter goal, it is important to study the scrum coverage to the Project Management (PM) category process areas, and to identify the earliest scrum methods that have a significant contribution in satisfying the CMMI specific practices [8], [9], [10].

According to the Forrester state of agile development report 65% of mid-sized enterprises (with less than 1000 employees) has reported that 100% of their teams uses agile for their software development. 85% of the agile adaptors embraces scrum as their applied agile method [11]. Having a systematic way that could enhance satisfying the CMMI certification using the commonly used scrum practices and assets could help more than 55% (85% of the 65%) of the mid-sized enterprises to get, or keep their CMMI certification. Some previous efforts tried to provide some ways to do this for CMMI Version 1.2 in some CMMI process areas. Some other process areas (i.e the process areas of Project Management) have not touched in previous research efforts. This paper shows how the scrum practices and assets could cover the Project-Management category's process areas of CMMI 1.3 with the use of normal scrum practices and assets without any extra enhancements.

The paper is organized as follows: Section 2 presents the background overview of CMMI and Scrum. Section 3 focuses on describing the problem- solving approach that is used to do the research. Section 4 presents the conducted research activities. Section 5 discusses the paper results against previous work that has been carried out by another previous research. The last section concludes the paper with final remarks.

II. Background Overview

A. Project Management Category of CMMI

The focus of this paper is on the project management category that covers the management activities that are related to planning, monitoring, and controlling the project. Table 1 shows the project management-related process areas grouped by its maturity levels [12], [13], [14].

TABLE I. CMMI Project Management Process Areas

Level	Process Area
2	Project Monitoring and Control (PMC) Project Planning (PP) Requirements Management (REQM) Supplier Agreement Management (SAM)
3	Integrated Project Management (IPM) Risk Management (RSKM)
4	Quantitative Project Management (QPM)

Project Planning (PP)

The purpose of project planning (PP) is to establish and maintain plans that define project activities [15], [16], [17], [18], [19], [20], [21]. PP contains three specific goals (SG): SG 1 "establish estimates", SG 2 "develop a project plan", and SG 3 "obtain commitment to the plan", enclosing 14 specific practices (SPs).

Project Monitoring and Control (PMC)

The purpose of project monitoring and control (PMC) is to provide an understanding of the project's progress so that appropriate corrective action is taken when the project's performance deviates significantly from the plan [22]. PMC contains two specific goals: SG 1 "monitor project against plans" and SG 2 "manage corrective action to closure", enclosing ten specific practices [23], [24], [25], [26], [27].

Requirements Management (REQM)

The purpose of requirements management (REQM) is to manage requirements of the project's products and product components and to ensure alignment between those requirements and the project's plans and work products [28]. REQM has one specific goal: SG 1 "manage requirements", enclosing five specific practices.

Integrated Project Management (IPM)

The purpose of Integrated Project Management (IPM) is to establish and manage the project and the involvement of relevant stakeholders according to an integrated and defined process that is tailored from the organization's set of standard processes [29]. The specific goals of the IPM are SG 1 "use the project's defined process" and SG 2 "coordinate and collaborate with relevant stakeholders".

Risk Management (RSKM)

The purpose of risk management (RSKM) is to identify potential problems before they occur so that risk handling activities can be planned and invoked as needed across the life of the product or project to mitigate adverse impacts on achieving objectives [30]. The specific goals of the RSKM are SG 1 "prepare for risk management", SG 2 "identify and analyze risks", and SG 3 "mitigate risks".

Supplier Agreement Management (SAM)

The purpose of supplier agreement management (SAM) is to manage the acquisition of products from suppliers [31]. The specific goals of the RSKM are SG 1 "establish supplier agreements" and SG 2 "satisfy supplier agreements".

Quantitative Project Management (QPM)

The purpose of quantitative project management (QPM) is to quantitatively manage the project to achieve the project's established quality and process performance objectives [32]. The specific goals of the QPM are SG 1 "prepare for quantitative management" and SG 2 "quantitatively manage the project". Scrum does not mention practices to address this process area. Therefore, all of its practices are *unsatisfied*.

B. Related Work

CMMI and agile methods have been compared in several studies [33] and mappings between agile and CMMI practices have been proposed [34]. For example, Fritzsche and Keil analyzed [22] in their study which CMMI process areas can be covered by Scrum and XP. Unfortunately, most of the findings were not clearly derived and they did not provide empirical evidences. On the other hand, Pikkarainen and Mäntyniemi proposed an approach for assessing agile software development by using CMMI. However, only two process areas were covered (PP and REQM) and only from a CMMI goal (not specific practice). Marcal et al. [32], in turn, presented a more detail mapping between CMMI project management process areas and Scrum practices, but they did not provide empirical evidences.

Finally, Diaz et al. [35], [36], [37] presented a mapping between Scrum practices and three process areas (PP, PMC, and REQM). In addition, they reported empirical results that provide evidences that those process areas were largely covered. However, the mapping presented was high level and the use of the now outdated CMMI-DEV Version 1.2 limited the results. Unlike these researches, our work tries to increase the detail of the previous mapping between CMMI and Scrum, and to cover the process areas related to the higher maturity levels (IPM, RSKM, and QPM) in CMMI-DEV Version 1.3 which is the latest version up till now. In addition, we consider in our study not only the CMMI specific goals, but also the specific practices. Furthermore, a complete view of the CMMI project management process areas covered by Scrum practices will be established.

III. Problem Solving Approach

A. Research Questions

The study that has been conducted in [7] was the base for a further research to discover the answers of some questions. First, which of the CMMI process areas could be satisfied by Scrum? Second. Which scrum assets could cover this satisfaction? Third, By which percentage this satisfaction coverage happens? Fourth, What are the most prior scrum assets to be considered when trying to cover CMMI using scrum? Fifth, what are the major gaps that are still there? Finally, which process areas are in conflict?

B. Defining Problem Solving Approach

In order to answer research questions; the study has been designed to be accomplished on Five steps that are all limited by the Project Management category's Process Areas (PAs) of CMMI-Dev Ver. 1.3. First, reviewing the available scrum assets that have to be studied in order to discover what are the possible ones that could be used to be mapped to the available specific practices (SPs) of CMMI-Dev Ver 1.3. Second, based on the reached mapping between scrum assets and CMMI practices, for the sake of a better understanding, a matrix should be developed to show all possible mappings between Scrum practices and CMMI-Dev SPs within the PM category. Third, according to the developed mapping matrix, the scrum coverage for CMMI SPs should be calculated according to a set of well-defined functions. This will help in finding out what are the prior scrum practices to take care of, when it is needed to streamline CMMI implementation using scrum. Discovering the prior scrum practices is the fourth designed step in this study. Finally, The results should be discussed, and compared to other related work to define the percentage of enhancement (if any) this study may achieve. The whole study has been applied on 6 companies out of the CMMI certified companies in Egypt and that are appraised by the regional CMMI center in Cairo that is called Software Engineering Competency Center (SECC).

IV. Study Activities

Step 1: Mapping Scrum assets to the corresponding CMMI SPs:

During the study, we analyzed each CMMI project management process area and all of its specific goals and practices in details and compared them with known Scrum practices. Table 2 depicts how we could satisfy the specific practices of the PP process area through using Scrum Practices. Each mapping set (single raw in that table) shows the satisfaction rating at the last column. In case the mapped Scrum Practices could satisfy the corresponding CMMI's specific practice fully (Rating :S), Partially (Rating: PS). If the set of Scrum assets could not satisfy, the corresponding SP at all it is considered as unsatisfied and tagged in the table with Rating U. Table 3 shows the same idea of scrum assets mapping to the SPs of PMC process area. Table 6 shows this for the REQM process area while Table 7 shows this for the context of IPM process Area.

TABLE II. Mapping Between Pp Process Area Specific Practices and Scrum Practices

CMMI Practice		Scrum Assets	Rating
SP 1.1	Estimate the scope of the project	• The product backlog is a WBS for product releases. • The sprint backlog is a WBS for sprint.	S
SP 1.2	Establish estimations of work product and task attributes	• User stories are estimated relatively in story points during release planning. • Tasks are estimated in hours during sprint planning. • Story point's estimates are based on complexity, uncertainty, and time. • Tasks time estimates are based on expert judgment.	S
SP 1.3	Define project lifecycle	• Scrum defines process which contains three phases: pre-game, development and post-game [35].	S
SP 1.4	Estimate Effort and Cost	• In Scrum, estimations occur twice. ○ The first estimation occurs during pre-game phase. ○ The second estimation occurs at the beginning of each sprint. • Story points, velocity and sprint buffer are rational for effort estimation. • The Scrum framework does not explicitly address calculation of costs.	PS
SP 2.1	Establish the budget and schedule	• The schedule is composed of the set of all predefined sprints in the release plan and the sprints plans [36]. • Project's budget can be derived from the estimated effort. • Scrum does not explicitly mention orientations about establishing budget.	PS

TABLE III. (CONTD.): MAPPING BETWEEN PP PROCESS AREA SPECIFIC PRACTICES AND SCRUM PRACTICES

SP 2.2	Identify project risks	• Risk identification is continually done through release planning, sprint planning, daily stand-ups, and retrospective and review meetings. • Risks are captured as impediments. They are registered on white-boards, flip charts, or impediments list. • Risk assessment, categorization and prioritization occur in an informal manner [37].	PS
SP 2.3	Plan for data management	• Scrum stores data in public folders or white-boards which are available for everyone that has a physical access. • Scrum does not address a formal data management plan or procedure. • Data privacy is another weakness [1].	U
SP 2.4	Plan for project resources	• During the pre-game phase, the project team, tools and other resources are defined. • During the project, scrum master ensures that the necessary resources are available.	S
SP 2.5	Plan for needed knowledge and skills	• At the start of a Scrum project, the knowledge and skills needed to perform the project are defined. • Knowledge and skills which are not found in the organization are considered as impediments and are resolved during the daily and retrospective meetings [33].	S
SP 2.6	Plan stakeholder involvement	• Scrum defines roles, responsibilities, and involvement of the stakeholders during the project's execution. • The scrum master is responsible for ensuring that all stakeholders involved in the project follow scrum rules and practices.	S
SP 2.7	Establish the project plan	• The vision document, the product backlog, the release plan, and the sprint plan are considered as a high-level plan for the project.	S
SP 3.1	Review plans that affect the project	• Reviews are carried out during planning and retrospectives meetings. • The CMMI model does not mention clearly what the plans should be reviewed [32].	S
SP 3.2	Reconcile work and resource levels	• Work reconciliation occurs during the sprint planning meeting where the team along with the product owner define the work to be developed in the sprint.	S
SP 3.3	Obtain plan commitment	• Commitment to the plan is obtained iteratively at the beginning of each sprint. • In the sprint planning meeting, the team selects as much user stories as it believes it can complete by the end of the sprint.	S

TABLE IV.
MAPPING BETWEEN PMC PROCESS AREA'S SPECIFIC PRACTICES AND SCRUM PRACTICES

CMMI Practice		Scrum Assets	Rating
SP 1.1	Monitor project planning parameters	• This occurs in the daily and retrospective meetings and by the use of burndown charts. • On release level, velocity, completed points, and total scope are monitored. • On sprint level, task effort and remaining effort are monitored.	PS
SP 1.2	Monitor commitments	• Commitments to the plan are established during sprint planning meeting and monitored through the use of burndown charts and daily and review meetings.	S
SP 1.3	Monitor project risks	• Risks are captured as impediments and are monitored through daily standup meeting and retrospectives meeting. • Scrum does not mention practices to define sources, parameters or categories to control the risk management effort [37].	PS
SP 1.4	Monitor data management	• The Scrum framework does not include any practices for planning and tracking the project data.	U
SP 1.5	Monitor stakeholder involvement	• Stakeholder involvement is monitored by the scrum master during project meetings. • Evidence of updated impediment backlog, product backlog and sprint backlog support the fulfilling of this practice [32].	S
SP 1.5	Conduct progress reviews	• Reviewing project progress is done on many levels: ○ On release level, it is done through sprint reviews, retrospectives, and release plan. ○ On sprint level, it is done through daily standup meetings, task boards, and sprint burndown chart.	S

TABLE V. (CONTD.): MAPPING BETWEEN PMC PROCESS AREA'S SPECIFIC PRACTICES AND SCRUM PRACTICES

SP 1.7	Conduct milestone reviews	• In Scrum, there are two main milestones: ○ Sprint milestone occurs at the end of the sprint. ○ Release milestone occurs when the team has completed the sprints in the release.	S
SP 2.1	Analyze issues	• Collecting and analyzing project issues are done either during daily standup meetings or through retrospectives. • Issues are registered on a white board, flip chart or impediment list.	S
SP 2.2	Take corrective action	• As mentioned before, issues are collected and analyzed during the daily and retrospective meetings. • The team could decide whether they take the corrective actions immediately or they fix it in an	PS

		upcoming sprint. - Scrum does not track how these actions are planned, monitored and implemented [32].	
SP 2.3	Manage corrective action	- Issues are registered on a white board, flip chart or impediment list. - All corrective actions are monitored until closing. - The results are not analyzed to determine their efficacy.	PS

TABLE VI. MAPPING BETWEEN REQM PROCESS AREA'S SPECIFIC PRACTICES AND SCRUM PRACTICES

CMMI Practice		Scrum Assets	Rating
SP 1.1	Understand Requirements	- The intense involvement of the stakeholders guarantees a common understanding of the requirements to all people involved [1]. - User stories can facilitate the common understanding of the requirements between the team and the customer.	S
SP 1.2	Obtain Commitment to Requirements	- Commitments to requirements are made collectively by the whole team through release planning and sprint planning. - The obtaining of commitment is a task of the Scrum master who can take necessary actions to gain commitment [1].	S
SP 1.3	Manage Requirements Changes	- The product owner frequently changes the user stories in the product backlog and makes it ready for the next sprint. - If a user that was already implemented changed, a new story is created and linked to its old story. - The product owner and the team discuss the changes to the user stories in the sprint planning meeting.	S
SP 1.4	Maintain Bidirectional Traceability of Requirements	- The hierarchy of Scrum requirements formats - themes, epics, user stories, and tasks - supports the traceability among the requirements. - If a user story dependent on another, the first one is prioritized higher than the second one. - Dependencies between user stories are discussed in the daily meeting or in the sprint planning meeting.	S
SP 1.5	Ensure that plans and work products remain aligned with the requirements	- Scrum backlogs help to ensure constancy between plans and requirements. - The sprint backlog helps to ensure that only the work that has been committed will be implemented. - No activities are implemented that do not belong to a user story. - The definition of done (DOD) supports constancy between work products and plans.	S

It is important to note that IPM process area, Scrum does not define a set of organizational standard processes, but it just establishes a set of practices and rules defined for the project. In other words, the project's defined process is not derived from a set of organizational processes. Therefore, all of the specific practices related to the Specific Goal 1 (SG1) are unsatisfied, except SP 1.6 "Establish Teams". Other specific practices of SG2 are satisfied, or partially satisfied as depicted in the table.

Concerning Risk Management (RSKM) process area, risks are captured as impediments and registered on white-boards, flip charts, or impediments list. However, scrum has no practices to define sources, parameters, or categories to analyze and control the risk management effort. Thus, risk assessment, categorization, and prioritization occur in an informal manner [37]. Therefore, all of the specific practices of RSKM are *unsatisfied*, except SP 2.1 (identify risks), because it is *partially satisfied* for the same reasons presented for PP SP 2.2.

For Supplier Agreement Management (SAM) Scrum does not mention practices to address the acquisition of products from suppliers. So, all of its specific practices are *unsatisfied*.

For Quantitative Project Management (QPM) Scrum does not mention practices to address this process area. Therefore, all of its practices are *unsatisfied*.

TABLE VII. MAPPING BETWEEN IPM PROCESS AREA'S SPECIFIC PRACTICES AND SCRUM PRACTICES

CMMI Practice		Scrum Assets	Rating
SP 1.6	Establish Teams	- In scrum, the well-defined responsibilities of scrum roles support team's establishment. - There are only three Scrum roles: the product owner, the team, and the Scrum master. - All management responsibilities in a project are divided among these three roles.	S
SP 2.1	Manage Stakeholder Involvement	- Scrum practices and rules implicitly define how stakeholders will be involved in the project. - This involvement is monitored by the Scrum master	S

TABLE VIII. MAPPING BETWEEN IPM PROCESS AREA'S SPECIFIC PRACTICES AND SCRUM PRACTICES (CONTD.)

SP 2.2	Manage dependencies	- Dependencies can be identified through Scrum daily meetings. - The Scrum master is responsible for resolving any identified problem as soon as possible. - Scrum does not handle coordination outside the Scrum team.	PS
SP 2.3	Resolve coordination issues	- This practice is partially satisfied, for the same reasons presented for IPM SP 2.2.	PS
		-	

Step 2: Mapping scrum practices to the appropriate CMMI Specific Practices

Our study showed that a satisfaction matrix could be built to map some scrum practices to the CMMI specific practices of the Project Management (PM) related process areas. Figure 1 and figure 2 depict a resulting CMMI-Scrum Mapping satisfaction matrix that shows how the available scrum practices could be applied to the available CMMI practices.

CMMI Practices	Scrum Practices																		
	Pre-Game	Vision Document	Product Backlog	Sprint Backlog	Release Planning	Sprint Planning	Daily meeting	Sprint Review	Sprint Retrospective	Release Review	User Stories	Story Points	Scrum Process	Velocity	Sprint Buffer	Burndown Charts	Impediment Backlog	Task boards	Definitions of Done
PP																			
SP 1.1			×	×															
SP 1.2					×	×					×	×							
SP 1.3													×						
SP 1.4												×		×	×				
SP 2.1					×	×													
SP 2.2					×	×	×	×	×										
SP 2.3																			
SP 2.4	×																		
SP 2.5	×								×										
SP 2.6													×						
SP 2.7		×	×		×	×													
SP 3.1						×			×										
SP 3.2						×													
SP 3.3						×													

Fig. 1. Mapping matrix of suitable scrum practices to the PP specific practice of CMMI-Dev

CMMI Practices	Scrum Practices																		
	Pre-Game	Vision Document	Product Backlog	Sprint Backlog	Release Planning	Sprint Planning	Daily meeting	Sprint Review	Sprint Retrospective	Release Review	User Stories	Story Points	Scrum Process	Velocity	Sprint Buffer	Burndown Charts	Impediment Backlog	Task boards	Definitions of Done
PMC																			
SP 1.1							×		×							×			
SP 1.2						×	×	×								×			
SP 1.3							×		×										
SP 1.4																			
SP 1.5																	×		
SP 1.6					×		×	×	×							×		×	
SP 1.7								×		×									
SP 2.1							×		×								×	×	
SP 2.2							×		×										
SP 2.3																	×	×	
REQM																			
SP 1.1			×	×	×	×	×	×	×	×	×								
SP 1.2				×	×														
SP 1.3			×			×													
SP 1.4						×	×												
SP 1.5			×	×															×
IPM																			
SP 1.6													×						
SP 2.1													×						
SP 2.2							×												
SP 2.3							×												
RSKM																			
SP 2.1					×	×	×	×	×										

Fig. 2. Mapping matrix of suitable scrum practices to the PMC, REQM, IPM, and RSKM specific practice of CMMI-Dev

Step 3: Discovering the prior Scrum practice in the mapping

Table 6 shows the number of CMMI specific practices that is supported by each scrum practice. The third row in the table shows a Mapping Importance Score (MIP), which is an index score that has been designed in the study in order to discover what are the prior scrum practices when implementing CMMI Project Management category using scrum practices. It is clear from the table that the top five prior scrum practices are; Sprint Planning, daily meeting, sprint retrospective, release planning, and product backlog. The index assigns 3 points for S satisfaction rating whiling assigning only 1 point for PS rating.

TABLE IX. MAPPING IMPORTANCE SCOPE AND THE RESULTING RANK

	Scrum Practices																		
	Pre-Game	Vision Document	Product Backlog	Sprint Backlog	Release Planning	Sprint Planning	Daily meeting	Sprint Review	Sprint Retrospective	Release Review	User Stories	Story Points	Scrum Process	Velocity	Sprint Buffer	Burndown Charts	Impediment Backlog	Task boards	Definitions of Done
CMMI Specific Practices Covered	2	1	5	4	8	12	12	6	10	2	2	2	4	1	1	3	3	3	1
Fully Satisfied	2	1	5	4	5	9	5	4	5	2	2	1	4	0	0	2	2	2	1
Partially Satisfied	0	0	0	0	3	3	7	2	5	0	0	1	0	1	1	1	1	1	0
Mapping Importance Score(MIP)	6	3	15	12	18	30	22	14	20	6	6	4	12	1	1	7	7	7	3
Index Rank	9	10	5	7	4	1	2	6	3	9*	9*	10	7*	11	11	8*	8*	8**	10

Step 4: Calculating the degree of scrum practices to CMMI SP

Each practice was rated according to the ratings in Table 7 [32]. For a certain rating i (i=1…n) and a process area j (j=1…m), we have number of practices $X_{i,j}$. So, the present for ratings can be calculated as follows:

$$P_i = \frac{x_{i,j}}{\sum_{i=1}^{n} \mathrm{x}_{i,j}} \times 100\% \qquad (1)$$

Also, the total percent for all the process areas is given by

$$P_j = \frac{x_{i,j}}{\sum_{j=1}^{\mathrm{m}} \mathrm{x}_{i,j}} \times 100\% \qquad (2)$$

TABLE X. SHOWS AN INDEX OF PRACTICES RATING

Rating	Criteria
Satisfied (S)	Means that CMMI practice is fully addressed by Scrum practices.
Partially Satisfied (PS)	Means that some parts of CMMI practice is addressed by Scrum practices.
Unsatisfied (U)	Means that CMMI practice is not addressed by Scrum practices.

The following gives a summary on the covered specific practices of CMMI Project Management process areas by Scrum practices. Tables 8 shows the results of mapping scrum practices to the CMMI specific practices of the project management related specific areas (i.e. PP, PMC, REQM, SAM, RSKM, IPM and QPM). Table 9 explains this through depicting the coverage percentage per process area. A complete view of the CMMI Project Management process areas covered by Scrum practices is shown in Figure 3. This result shows that 37% of specific practices of CMMI project management process areas are satisfied, 17% are partially satisfied and 46% are unsatisfied. In other words, CMMI project management process areas are not fully covered with Scrum practices mainly related to SAM, RSKM, IPM and QPM process areas.

TABLE XI. COVERAGE OF CMMI PROCESS AREAS BY APPLYING THE RECOMMENDED SCRUM PRACTICES

	PP	PMC	REQM	SAM	IPM	RSKM	QPM	Total
Satisfied	10	5	5	0	2	0	0	22
Partially Satisfied	3	4	0	0	2	1	0	10
Unsatisfied	1	1	0	6	6	6	7	27

TABLE XII. COVERAGE PERCENTAGE OF CMMI PROCESS AREAS BY APPLYING THE RECOMMENDED SCRUM PRACTICES

	PP	PMC	REQM	SAM	IPM	RSKM	QPM	Total
Satisfied	72%	50%	100%	0%	20%	0%	0%	37%
Partially Satisfied	21%	40%	0%	0%	20%	14%	0%	17%
Unsatisfied	7%	10%	0%	100%	60%	86%	100%	46%

Considering each CMMI project management process areas according to its maturity level, another analysis can be made as shown in Figure 4. Table 10 and Table 11 show the results of the mapping for CMMI project management process areas, grouped by maturity level. In this way, process areas related to maturity level 2 (PP, PMC, RQM and SAM) have 57% of its specific practices satisfied by Scrum, 20% are partially satisfied and 23% are unsatisfied. If SAM are not considered, Scrum becomes more compliant with level 2 without major adaptations (69% satisfied, 24% partially satisfied and 7% unsatisfied).

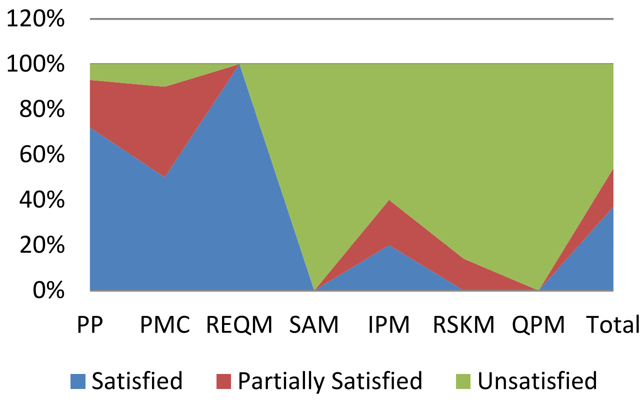


Fig. 3. General Result from the Mapping

TABLE XIII. CMMI PROJECT MANAGEMENT PROCESS AREAS COVERED BY APPLYING THE RECOMMENDED SCRUM PRACTICES, GROUPED BY MATURITY LEVEL

	PM Category Level 2	PM Category Level 3	PM Category Level 4
Satisfied	20	2	0
Partially Satisfied	7	3	0
Unsatisfied	8	12	7

TABLE XIV. COVERAGE PERCENTAGE OF CMMI PROCESS AREAS BY APPLYING THE RECOMMENDED SCRUM PRACTICES, GROUPED BY MATURITY LEVEL

	PM Category Level 2	PM Category Level 3	PM Category Level 4
Satisfied	57%	12%	0%
Partially Satisfied	20%	18%	0%
Unsatisfied	23%	70%	100%

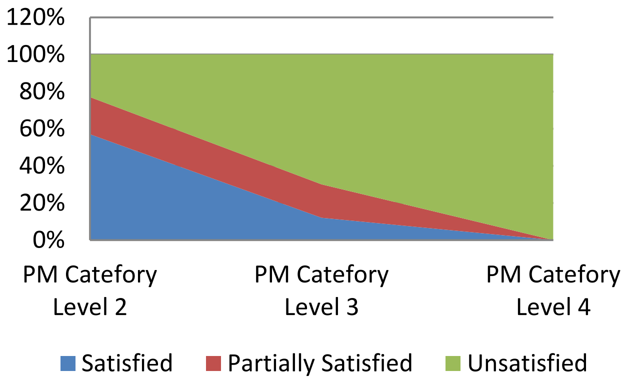


Fig. 4. General Result from the Mapping on PM Category, Grouped by Maturity Levels

With regard to process areas related to maturity level 3 (IPM and RSKM), these process areas have 12% of its specific practices satisfied by Scrum, 18% are partially satisfied and 70% are unsatisfied due to the lack of practices for managing risks and the absence of a defined process derived from a set of organizational processes. Finally, process areas related to maturity level 4 (QPM) are unsatisfied because Scrum does not mention practices to address these process areas.

The major gaps between Scrum and PP, PMC, REQM, SAM, RSKM, IPM and QPM process areas are presented below:

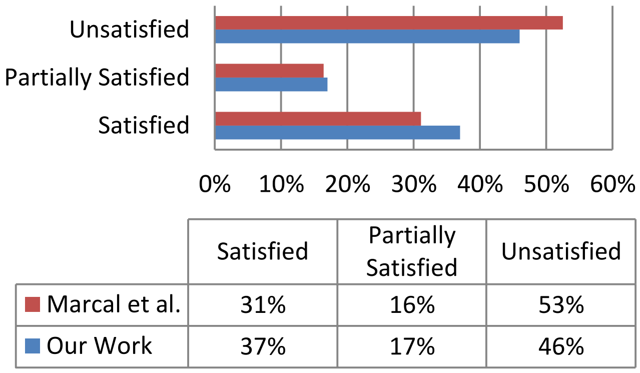
- Scrum framework does not provide explicit description for planning and controlling of project's budget, affecting PP and PMC.
- Scrum framework does not explicitly address calculation of project's costs, affecting PP and PMC.
- Lack of practices for managing risks, affecting RSKM, PP and PMC practices.
- Scrum framework does not include any practices for planning and tracking data management, affecting PP and PMC practices.
- Scrum does not define a set of organizational standard processes, but it just establishes a set of practices and rules defined for the project, affecting IPM practices.
- Scrum does not mention practices to address the acquisition of products from suppliers. So, all SAM specific practices are *unsatisfied*.
- Scrum has no practices to address the QPM process area, so all its practices are *unsatisfied.*

V. RESULTS DISCUSSION, AND FUTURE WORK

A. *Results Discussion*

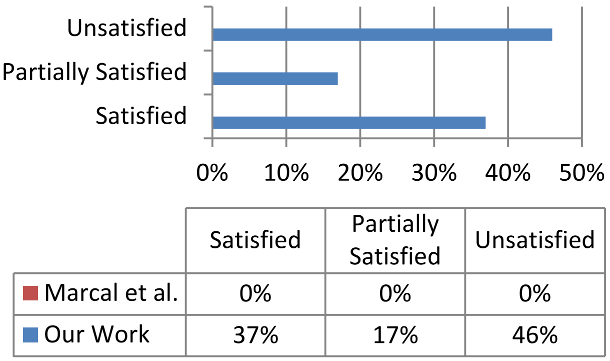
based on the above, Figure 5 and Figure 6 shows a comparison between our overall results of the mapping and those of Marcal et al [32].

It should be noted that the mapping between the specific practices of the CMMI process areas and Scrum practices according to Marcal et al. considers the staged representation of CMMI-DEV version 1.2, but our work considers the CMMI-DEV version 1.3 which is the latest version till now. Working with older versions of CMMI-DEV will not be of a much help to the new companies that are looking for a CMMI certification.



	Satisfied	Partially Satisfied	Unsatisfied
🟥 Marcal et al.	31%	16%	53%
🟦 Our Work	37%	17%	46%

Fig. 5. A Comparison between Our Overall Results of the Mapping for PM category and Those of Marcal et al., Considering CMMI Version 1.2



	Satisfied	Partially Satisfied	Unsatisfied
🟥 Marcal et al.	0%	0%	0%
🟦 Our Work	37%	17%	46%

Fig. 6. A Comparison between Our Overall Results of the Mapping and Those of Marcal et al., Considering CMMI Version 1.3

B. *Future Work*

More work will be done to deliver a scrum coverage for other process areas that have not been covered on this paper (i.e SAM, RSKM, and QPM). Since SAM is highly required nowadays due to the increasing size of the outsourcing engagements in the software development industry [38], automating SAM within Scrum process could be a good start for the future work since it is unclear area in the integration between CMMI and Scrum.

VI. Conclusion

In our previous research work [7], it was shown that working with agile could cover a portion of the CMMI specific practices with applying some non-scrum practices. This study proposed a new way to use the simpler scrum practices and assets in order to satisfy the more complicated CMMI-Dev version 1.3’s specific practices. This leads to satisfying the CMMI generic practices thus, satisfying the CMMI process areas. The coverage that this work has achieved haven’t been achieved for any previous similar work. Additionally, other older efforts haven’t worked on the currently used CMMI version (1.3). This research has been conducted to delve in more details concerning the important scrum practices that could be improved using existing scrum assets to streamline CMMI specific practices implementation in the Project Management category, and what could be the effect of that on the overall scrum coverage of the CMMI specific practices. This research has designed a new score index that could measure the importance of specific Scrum practices in the CMMI practices coverage (MIP Score Index). This score index showed that the top five practices that should be always improved to enhance the CMMI coverage were in order; Sprint Planning, Daily meeting, sprint retrospective, release planning, and product backlog. After applying improvements on these practices in 6 different CMMI appraised companies in Egypt some positive results have been concluded compared to other previous research efforts that have been done by Marcal et al. The work that has been done by Marcal et al. has been done in the same context but with two shortcomings. The first is that it was not updated for CMMI version 1.3. This shortcoming means that it couldn’t help teams that will be appraised based on this version. The second shortcoming is that it did not provide enough coverage of Scrum practices to CMMI version 1.3 specific practices. Our research has resulted in delivering a better satisfaction with additional 6% coverage, which is 19.4% development. Moreover, this research increased the level of scrum partial satisfaction coverage of the CMMI practices by extra %1 which resembles 6.2% development percentage. In addition to that, previous work did not consider CMMI version 1.3. That is why, if compared to the version 1.3, the comparison percentages will be against 0% enhancements from Marcal’s previous work.

The conclusion is that CMMI and Scrum method can work well together, and both approaches can bring more benefits than either one alone. So, Scrum method provides good practices for streamlining CMMI project management process areas related to maturity level 2, and 3. On process areas related to level 4, Scrum method covers only a little part of these areas. Other alternative practices are necessary to make these two approaches more compliant. This alternative is useful for companies that are adopting agile methods while searching for a CMMI certification.

References

[1] A. Preis, "Integration Evaluation of Scrum and CMMI," Organizational Science, 2012.

[2] "Manifesto for Agile Software Development," 29 November 2013. [Online]. Available: http://agilemanifesto.org.

[3] A. Garg, "Agile software development," DRDO Science Spectrum, pp. 55-59, March 2009.

[4] M. Pikkarainen and A. Mäntyniemi, "An Approach for Using CMMI in Agile Software Development Assessments: Experiences from Three Case Studies," in SPICE, 2006.

[5] H. Glazer, "Agile CMMI: Why Isn’t This Conversation Dead Yet?," Cutter IT Journal, vol. 25, no. 11, November 2012.

[6] H. Glazer et al, "CMMI or Agile?: Why Not Embrace Both?!," Carnegie Mellon University, Software Engineering Institute, Nov, 2008.

[7] "SIMPLIFYING CMMI VERSION 1.3 IMPLEMENTATION BY USING AGILE," International Journal of Intelligent Computing and Information Sciences (IJICIS), vol. 14, no. 4, pp. 31-45, October 2015.

[8] Irrazabal, E., et al, "Applying ISO/IEC 12207:2008 with SCRUM and Agile Methods," in 11th International Conference, SPICE, 2011.

[9] "CMMI Institute," [Online]. Available: http://cmmiinstitute.com/. [Accessed 29 April 2013].

[10] M. B. Chrissis et al.,, CMMI for Development, Addison-Wesley, 2011.

[11] D. L. Giudice, "The 2015 State Of Agile Development," Forrester Research, August 3, 2105.

[12] CMMI Product Team, "CMMI for Development, Version 1.3," Software Engineering Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2010.

[13] SCAMPI Upgrade Team, Standard CMMI Appraisal Method for Process Improvement (SCAMPI) A, Version 1.3: Method Definition Document, Pittsburgh, Pennsylvania: Software Engineering Institute, Carnegie Mellon University, 2011.

[14] M. Staples and M. Niazi, "Two case studies on small enterprise motivation and readiness for CMMI," in 11th International Conference on Product Focused Software Development and Process Improvement, 2010.

[15] K. M. Calo et al, "A Quantitative framework for the evaluation of agile methodologies," JCS&T, vol. 10, no. 2, 2 June 2010.

[16] "Agile Alliance," 10 January 2013. [Online]. Available: http://www.agilealliance.org/.

[17] k. Pathak and A. Saha, "Review of Agile Software Development Methodologies," International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), vol. 2, 2013.

[18] B. N. Nathan-Regis and V. Balaji, "Evaluation of the most used agile methods (XP, Lean, Scrum)," International Journal of Engineering Science and Technology (IJEST), vol. 4, no. 1, January 2012.

[19] M. Hneif and S. Hock, "Review of agile methodologies in software development," International Journal of Research and Reviews in Applied Sciences, vol. 1, October 2009.

[20] M.Qasaimeh et al, "Comparing Agile Software Processes Based on the Software Development Project Requirements," IEEE Computer Society, p. 2008.

[21] "Scrum Alliance," 2 March 2013. [Online]. Available: http://www.scrumalliance.org/.

[22] CMMI Product Team, "CMMI for Development Quick Reference," November 2010. [Online]. Available: http://cmmiinstitute.com/assets/CMMI-DEV_Quick_Ref.pdf.

[23] J. Sutherland, Scrum Handbook, Scrum Training Institute, July, 2010.

[24] N. Potter and M. Sakry, "Scrum - Lessons From The Trenches," The Process Group-Post Newsletter, vol. 19, no. 1, 2012.

[25] K. Schwaber, "Agile Project Management with Scrum," Microsoft, 2004.

[26] Microsoft Team, "MSF for Agile Software Development v5.0," Microsoft, May 2010. [Online]. Available: http://msdn.microsoft.com/en-us/library/dd380647(v=vs.100).aspx .

[27] K. Schwaber and J. Sutherland, "The Scrum Guide.," October 2011. [Online]. Available: http://www.scrum.org/Portals/0/Documents/Scrum%20Guides/Scrum_Guide.pdf#zoom=100.

[28] M. Foegen, "Scrum and CMMI – Does it fit together?," wibas GmbH, November 2010. [Online]. Available: http://www.wibas.com/publications/scrum/scrum_and_cmmi/index_en.html.

[29] C. R. Jakobsen and J. Sutherland, "Scrum and CMMI Going from Good to Great," in Agile Conference, 2009.

[30] H. Glazer, "Love and marriage: CMMI and Agile Need Each Other," CROSSTALK: The Journal of Defense Software Engineering, 2010.

[31] M. Fritzsche and P. Keil, "Agile Methods and CMMI: Compatibility or Conflict," e-Informatica Software Engineering Journal, vol. 1, 2007.

[32] A. S. C. Marcal et al, "Mapping CMMI project Management Process Areas to Scrum Practices," in 31st IEEE Software Engineering Workshop (SEW 2007), 2007.

[33] J. Diaz et al., "Mapping CMMI Level 2 to Scrum Practices: An Experience Report," EuroSPI, 2009.

[34] N. Potter and M. Sakry, "Implementing Scrum (Agile) And CMMI® Together," The Process Group-Post Newsletter, vol. 16, no. 2, 2009.

[35] P. Abrahamsson et al, "Agile Software Development Methods," 2002. [Online].

[36] M. Foegen and D. Croome, "How Scrum helps with CMMI," wibas GmbH, Febreuary 2011. [Online]. Available: http://www.wibas.com/publications/cmmi/cmmi_and_scrum/index_en.html.

[37] B. Reddaiah et al., "Risk Management Board for Effective Risk Management in Scrum," International Journal of Computer Applications, vol. 65, no. 12, p. 0975 – 8887, March 2013.

[38] S. Islam and M. D. Ahmed, "Business process improvement of credit card department: case study of a multinational bank," Business Process Management Journal, vol. 18, pp. 284 - 303, 2012.

Development of System Architecture for E-Government Cloud Platforms

Margulan Aubakirov
JSC National Information Technologies
Astana, Kazakhstan

Evgeny Nikulchev
Moscow Technological Institute
Moscow, Russia

***Abstract*—Requirements and criteria for selection of cloud platform and platform visualization are stated by which optimal cloud products will be chosen for the Republic of Kazakhstan e-Government considering quality-price ratio, and also the framework of information and communication architecture will be introduced.**

Keywords—cloud technologies; outsourcing; Kazakhstan; cloud platform; e-Government

I. INTRODUCTION

Cloud technologies is an effective model for reduction of aggregated value of information systems ownership due to resources incorporation to shared pool (for example, computer capacities, data storage systems, channel capacity and memory) from which resources can be immediately allocated and deployed in accordance with changes made to requirements.

The cloud services provide an opportunity for users to keep data (for example, pictures or e-mail messages), use software (for example, social media, video and audio files, games, etc.). For companies, including government agencies, cloud services can be applied as substitutes to internal data centers and departments in charge of ICT [1]. Companies not having equity contributions to the creation of information infrastructure can suggest services and work to their future customers. In general, cloud technologies embody a further industrialization (standardization, scaling-up, major distribution) just like provision of electrical power by electric supply stations to end users. Due to a standardized interface, the absence of necessity to solve issues of data centers creation, launch and safety provision and servicing of a variety of users it is possible to achieve an effect from the scale. In this regard, cloud technologies distribution on the national level enables achieving an optimal expenditure level [2]. Economic surveys results acknowledge the importance of cloud technologies application and forecast their distribution throughout the world [3].

Application of cloud technologies allow decreasing capital and current operating expenses and increase the level of equipment usage, which currently amounts to 10% in public sector. It means that 90% of actually acquired equipment in public sector is on downtime, which in its turn means its inefficient use [4].

Thus, the common objective of a new government agencies' informatization service model development is reduction of expenses for IT-resources management; reduction of costs for IT-personnel, optimization of budget costs on procurement of IT-equipment, implementation of single pricing policy, increase of servicing quality; enhancement of government agencies' IT infrastructure and information safety, reduction of risks regarding data loss and corruption [4]. Above-mentioned conditions can boost the percent of Kazakhstani participation in services procured by government agencies and provided by cloud platform operator.

In this regard, cloud computing is designated as crucial technology according to the result of technological forecast executed as part of Framework for Innovative Development of the Republic of Kazakhstan until 2020 in the field of information and communication technologies.

One of the mechanisms for efficiency enhancement of information technologies application in government agencies is the implementation of a new informatization model based on migration to the use of cloud computing, ICT-outsourcing and orders consolidation.

The results of the technology implementation are as follows: budget consolidation and saving, government agencies' business-processes efficiency.

As part of G-Cloud realization, on the 1st phase of the project, it is planned to deliver IaaS (the service for the provision of the virtual machine and service for allocation of virtual space for backup and storage) and SaaS (the service of e-workflow, e-postal service).

II. ANALYSIS OF CONDITIONS FOR CREATION OF G-CLOUD PUBLIC PLATFORM IN THE REPUBLIC OF KAZAKHSTAN

One of the mechanisms for efficiency enhancement of information technologies application in government agencies is implementation of a new informatization model based on migration to application of cloud computing. Proposed informatization (service) model of government agencies implies the consolidation of government agencies' IT-infrastructure, provision of a number of IT-services in accordance with principles of "cloud computing", including such services as e-mails, e-workflow and others.

The project implementation will address the following issues.

Today, development of information technologies in government agencies is characterized by the following issues:

- Procurement and maintenance of IT-equipment is not government agencies specialty.

- Unequal level of required IT-infrastructure provision and maintenance.
- Partial non-compliance with requirements of life sustenance systems backup.
- Impossibility to track the level of equipment load and maximum effective use.

Annually, government agencies provide budget for maintenance and development of existing information systems, infrastructure required for systems operation and procurement of servers, network and cross connect equipment, etc.

The first point is that costs on server equipment maintenance continue to rise by two-three times. In case if some information system requires more resources, additional purchase of necessary equipment or server`s complete replacement will be made. The process of server replacement (beginning from procurement and finishing with the system setup) might last for more than a month. In other words, the problem of server farm modernization will only become worse and lead to the slowdown of government agencies' informatization process, which means more frequent equipment failures, unsatisfactory provision of services and citizens dissatisfaction. Development of public cloud platform will help to solve problems with inefficient use of equipment, provide a required safety level and information systems efficiency.

Secondly, each government agency procured or sold information systems for automation of certain functions in order to boost efficiency. Therefore, budget funds were consistently provided for automation of government agencies standard functions. Besides, government agencies provide budget for maintenance and development of existing information systems, infrastructure required for systems operation and procurement of servers, network and cross connect equipment, etc. on an annual basis. All these lead to irrational spending of budget funds as in practice funds are spent several times for the same purposes by different agencies.

Likewise, when retail software is acquired for functions automation, agencies face the problem when a product doesn`t meet all the requirements. In this case, agencies have to provide additional budget resources for upgrading/development of such solutions. Automation of such standard functions inside each agency are partial as agencies procure /sell various solutions for automation of certain activities, which in its turn, inevitably leads to a large number of information systems inside an agency, each of which automates only one function. In the long run, agencies will have to address the problem of various information systems management, which considerably increases costs.

Annually, government agencies spend considerable budget funds on procurement of computer equipment (including servers), salaries and administrative expenditures of information and communication units of government agencies.

Also, there is imbalance in government agencies information and communication assets (availability of excessive assets in government agencies executing minute volume of functions and providing narrow spectrum of public services and lack of computation capacities in government agencies executing a large volume of functions and providing a wide spectrum of public services). The most part of procured technical means are not used in full extent and rapidly become obsolete because of technical progress in the field of informatization.

Government agencies performance has general and similar components. Today, each government agency solves the issue of informatization on their own, acquiring IC-equipment and software, creating data centers and automating public services. This includes budgeting and its approval for inspection of government agencies' IC-infrastructure, consulting services, feasibility study, software development, equipment procurement, licensing, maintenance and also, execution of tenders and contracting.

In many government agencies and quasi-public sector unlicensed software is used, which is a violation of international liabilities of the Republic of Kazakhstan in the field of intellectual property protection. To solve these problems it is planned to migrate on cloud computing technology based on virtualization one. Having cloud computing technology, a user gets full-featured virtual server that equals physical one.

As the result, there would be no necessity for government agencies to acquire physical equipment, service it and bear expenditures on salaries for personnel.

III. Requirements for Cloud Platform

Information and communication platform (hereinafter – ICP or G-Cloud) is hardware and software package intended to provide services to government agencies in the field of informatization with application of “cloud” technologies. Information and communication service (IC service) is an assembly of services for rent and allocation of computing resources, provision of software and equipment and also, communication services via which the stated services operate.

As part of G-Cloud implementation, on the 1st phase of the project it is planned to deliver IaaS (the service for provision of virtual machine and service for allocation of virtual space for backup and storage) and SaaS (the service of e-workflow, e-postal service).

Fig. 1. demonstrates resources management by client and managing company for every type of main services.

The main requirement specified for G-cloud platform is that it should be based on leading solution in the field regarding development of virtual and cloud infrastructures.

Cloud platform should provide a quick transfer of existing government agencies data centers to computing cloud. The platform should offer such possibilities as high availability, data recovery, "hot" migration, "hot" resources inclusion, resilience, automated distribution of resources, virtual hub for serial ports and store API-interfaces for integration of sets and maintenance of alternative I/O ways.

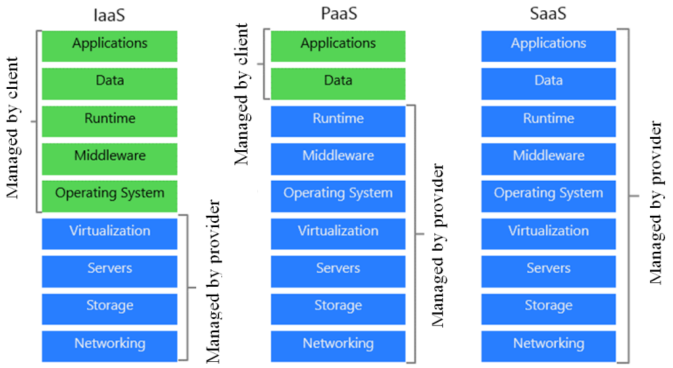


Fig. 1. Levels of resources management for various types of services

Application of cloud platform should provide a consolidation of existing infrastructure and optimization of government agencies' IT-equipment, ensure flexible scaling and reduce costs for solutions regarding provision of services operation continuity and emergency recovery.

The platform should ensure the possibility of functionality expansion through installation of additional modules and possibility to create cloud infrastructure (private or public computing cloud) with several renters due to resources incorporation in virtual data centers. Afterwards, the centers are offered to users via web-portals and software interfaces as fully automated and catalogued services.

The platform should support three-landscape approach for systems implementation:

Development environment is an environment which has a complex of programs with functionality required for systems development [6]. "Development environment" concept can be understood in various ways. For instance, a programmer`s main tool is integrated development environment, which has all the tools for creation of codes, compilation and so on. On the higher level, development environment is an adjusted environment together with development server, data, data base processor and other required tools needed for examination and testing of developing system`s components performance.

Test environment is an important link between development environment and real production environment. It consists of equipment (servers, operating computer (s), etc.) and components of logical level (server operating system, client operating system, database server, client user interface, web-browser (explorer) and other software [7]). The environment encompasses client components as well as server ones and uses the same versions of software that are located on clients, i.e. similar to production environment if possible. Otherwise, during the implementation process it might turn out that the system doesn`t work and needs readjustment. In test environment performance tests and efficiency tests are executed together with tests on systems upgrade and error control. Also, user acceptance tests can be performed. There is a thumb-rule according to which the whole test environment is separated from production environment and that all updates and adjustments are controlled in test environment and only afterwards are installed to production environment. Test environment is also suitable for clients training, which guarantees that during the process data, for instance, won`t be corrupted.

Production environment is an environment in which actual work is executed, i.e. activities executed by a company every day. Similar to test environment, production environment consists of complete software and hardware package.

Key requirements:

1) Support of open standards for storage and distribution of virtual machines (OVF);

2) Open software interface for integration with external systems;

3) Support of possibility to use cloning for deployment of virtual machines and applications;

4) Support of virtual machines bound copies constructed on the basis of "golden" patterns in order to save disk space of storage system and optimize applications deployment;

5) Built-in features of virtual infrastructure network security;

6) Support of infrastructure services catalogues with possibility of services publication in the catalogue by users;

7) Logical partitioning of resources pools provided by virtualization platform on virtual computing data centers with fixed service quality;

8) Support of operation with various organizations: isolation of virtual resources, independent LDAP authentication;

9) Self-service portal for users and administrators access;

*10)*Enhanced users capabilities on independent management of organizations infrastructure (computing resources, storage resources and local network resources management);

*11)*Support of virtual distributed switch;

*12)*Integration with solutions on provision of network safety;

*13)*Possibility to create protected network circuit for data exchange between various cloud infrastructures.

IV. Cloud Platform Functional Structure

In this section we`ll focus on cloud platform functional structure scheme for e-government agencies', its subsystems and their functionality. Fig. 2 demonstrates the scheme of infrastructural division.

"ICP for Internet" and "ICP for government agencies' Intranet" platforms are functionally identical and contain the following subsystems allocated by purpose and functionality. Pic. 3 demonstrates the scheme of ICP functional structure.

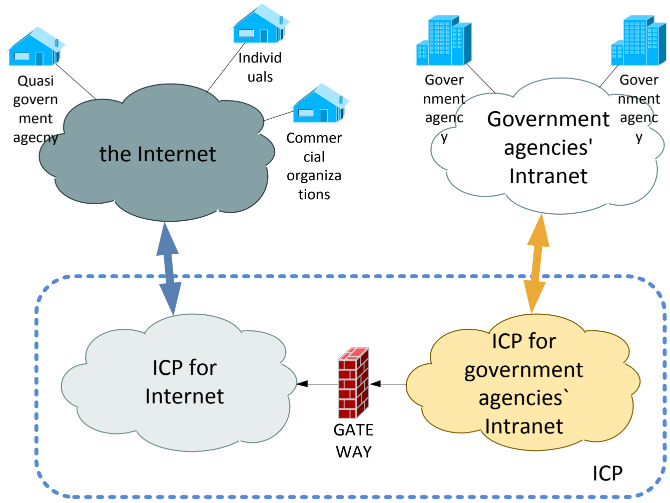


Fig. 2. Scheme for infrastrucutral division (G-cloud)

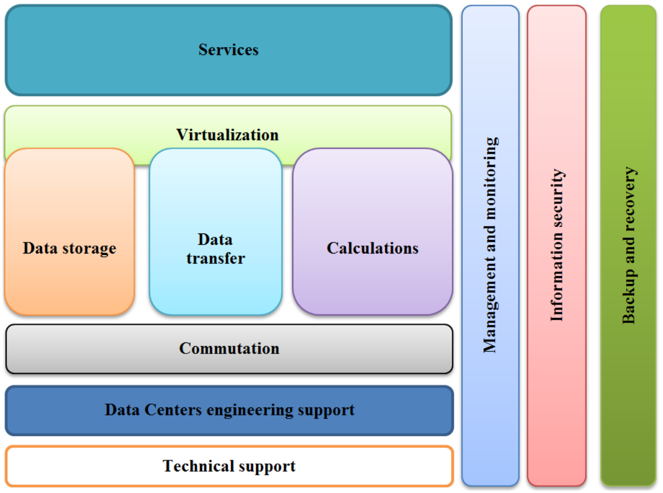


Fig. 3. Scheme of ICP functional structure

Subsystem of services provides key services (cloud services) to the system users.

Subsystem of resources virtualization implements software virtualization of physical computing resources.

Subsystem of computing platform provides the platform of unified physical servers.

Subsystem of data transfer provides traffic transfer in virtual environment.

Subsystem of data storage provides equipment for data storage and distribution. File and block access to data.

Subsystem of commutation provides physical computer network.

Subsystem of data centers engineering assistance provides data centers engineering subsystems.

Subsystem of management and monitoring provides subsystems management and monitoring functionality.

Subsystem of information safety provision provides the required information safety package.

Subsystem of backup and recovery provides the required functionality for data backup and recovery.

Subsystem of technical assistance provides organizational and technical support of the system operation.

V. Services Subsystem

Services subsystem provides users with software services, classification of which match below-mentioned criteria.

IaaS (infrastructure as a service). Rent of virtual capacities in cloud. Service model which provides users with virtualized technological infrastructure using which it is possible to deploy and execute software, including operating systems and server applications. Control and management over major physical and virtual cloud infrastructures, including networks, servers, types of operating systems in use, storage systems is executed by cloud provider [8].

PaaS (platform as a service). Ret of platforms for developers. Service model which provides users with environment for code deployment and execution, creation or acquisition of applications on cloud infrastructure with application of tools and programing languages supported by the platform with integrated service of e-Gov infrastructure. Management and control over major physical and virtual cloud infrastructure, including networks, servers and operating systems are executed by cloud provider except developed and installed applications and also, if possible control is provided f environment (platform) configuration settings [8].

SaaS (software (application) as a service). Service model providing users with access to various applications operating in cloud infrastructure. Applications can be of various types and available from all devices with different operating systems. Users' access to applications can be provided via dedicated software clients (including mobile platforms) or via web-browser. Management and control over major physical and virtual cloud infrastructure, including networks, servers and operating systems (except a restricted set of application configuration settings) are executed by cloud provider. The service model can also provide authorized users with access to general-purpose applications and to various specialized systems [8].

VDI is the creation of desktops in virtual environment. With the help of desktop virtualization technology, a user having any device with network access (smartphone, tablet, thin client) can receive access to personal desktop and corporate information resources.

Data remote backup is a service providing users with system for data backup and storage. Data remote backup systems are embedded into client program. This program capture, crunch, encipher and transfer data to ICP servers.

The fig. 4 below model of user`s interaction with cloud platform.

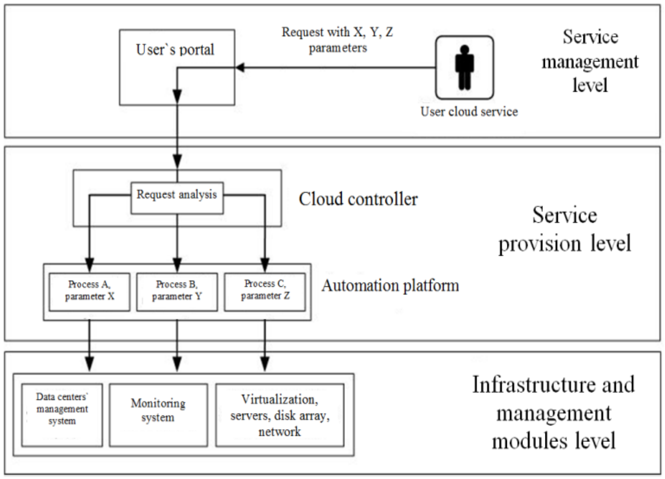


Fig. 4. User`s interaction with cloud platform

Cloud environment user makes a request for service management on service management level (creation, withdrawal, modification, etc.). The request contains the required action and parameters. For example, a user needs to add one more virtual server to computing resource. Such request will require server addition and its parameters will contain the following data: number of CPU, number of RAM, operating system type and so on, including computing cluster identifier to which it is required to add one more node.

The request is made in user-friendly graphics interface and transferred to service provision level to cloud controller. Cloud controller receives and analyses the request. According to analysis result it is required to check availability of required resources, resources reservation for certain user`s needs and afterwards, launch of required processes in underlying automation platform. Automated processes, which are initiated during the scenario implementation, receive required parameters transmitted from user`s request. During execution of the processes, the platform interacts with objects of infrastructure level and management modules. Scenario of virtual server inclusion to virtual cluster will include the following systems: virtualization system for creation of virtual machine with required parameters; servers management system for cluster software installation; platform`s connectors for configuration execution straight on a new node; monitoring system for automatic turn-on of a new node to general monitoring outline.

The fig. 5 demonstrates the logic of cloud platform's interaction with user.

User`s requirements are analyzed and transmitted to the set of automated procedures, which in its turn execute required technical operations. The system`s feedback is initiated by various subsystems of cloud environment on service provision level and infrastructure level. Feedback scenarios implement principles and tasks of computing capacities operation. Interaction between management and monitoring systems and cloud controller represents the main scenarios. Upon execution of some configuration task or receipt of system`s emergency message, the launch of corresponding processes on automation platform is initiated, which in their turn send required requests to cloud controller. Cloud controller associates the requests with services of certain user and publishes information on corresponding home page of service portal. Another example is publication of the results of computing tasks executed in cloud environment.

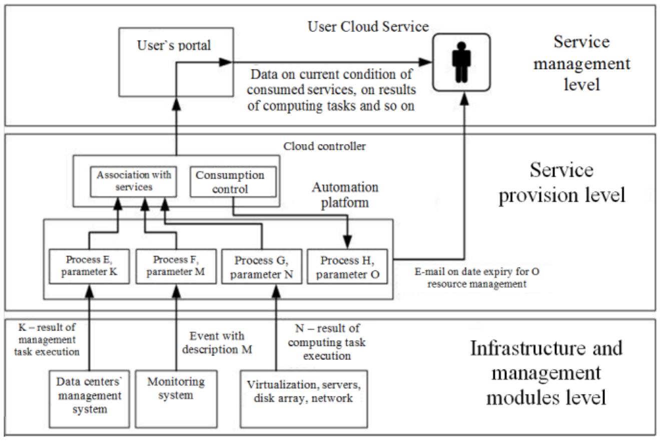

User Cloud Service
User`s portal
Data on current condition of consumed services, on results of computing tasks and so on
Service management level
Cloud controller
Association with services
Consumption control
Automation platform
Service provision level
Process E, parameter K
Process F, parameter M
Process G, parameter N
Process H, parameter O
E-mail on date expiry for O resource management
K – result of management task execution
Event with description M
N – result of computing task execution
Data centers` management system
Monitoring system
Virtualization, servers, disk array, network
Infrastructure and management modules level

Fig. 5. Cloud platform feedback to user

Another standard scenario is control over the use of computing center services and users timely notification on date expiry via portal or e-mail.

VI. Subsystem of Resources Virtualization

Subsystem of resources virtualization includes the following modules:

1) *Virtualization of file access*
2) *Virtualization of block access*
3) *Virtualization of data communication networks*
4) *Virtualization of computing resources*

Fig 6 shows a diagram for virtualization subsystem.

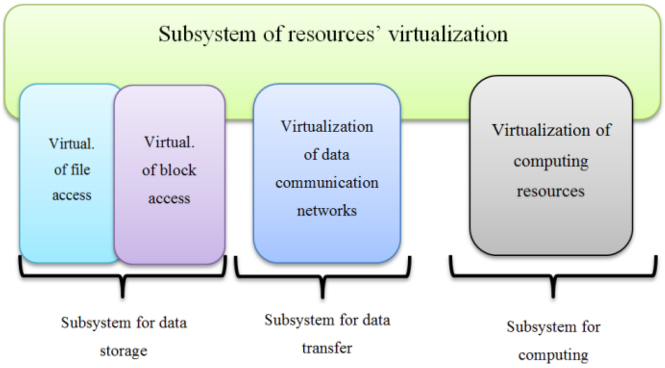

Subsystem of resources' virtualization
Virtual. of file access
Virtual. of block access
Virtualization of data communication networks
Virtualization of computing resources
Subsystem for data storage
Subsystem for data transfer
Subsystem for computing

Fig. 6. Block diagram for virtualization subsystem

Virtualization is creation of flexible substitute for physical resources with the same functions and external interface but with different attributes such as size and efficiency. Such substitute is called virtual resources and usually operating systems are not aware of substitution made.

Virtualization is applied to physical hardware via incorporation of several physical resources in one pool from which users can get virtual resources. With the help of virtualization it is possible to create several virtual resources from a physical one.

Virtual resources can have functions or peculiarities that are absent on physical ones. During virtualization, several virtual systems are created from one physical system. Virtual systems are independently operating environments, which use virtual resources.

Usually, system virtualization is executed with the help of hypervisor technology. Hypervisor (irrespective of type) is a multilayer application, which separates hardware from its guest systems. Each guest operating system sees virtual machine instead of physical equipment [9].

VII. Subsystem of Computing Platform

Subsystem of computing platform is divided on servers designated for solving of centralized management tasks and servers designated for execution of production tasks (service provisions). Combination of management servers is called “management cluster” and combination of managed servers is called “resource group”. Only one (resilient) copy of management cluster is installed whereas there might be several resource groups (depends on service levels or equipment set).

It includes the following servers:

- x86 architecture
- with 2 physical processor sockets
- with 4 physical processor sockets
- floor-standing version
- blade version

Main requirements to server equipment:

- availability of build-in technologies for provision of crucial components resiliency;
- duplication of power supply units;
- duplication of I/O interface;
- management of single errors in random access memory;
- selection of compatible processors scale preferably with the similar capacity for consistent operation of software hypervisor;
- for 2 socket servers the number of installed random access memory should be not less than 256 Gb;
- for 4 socket servers the number of installed random access memory should be not less than 512 Gb;
- local HD are not mandatory as servers should download OS from external data storage systems.

VIII. Conclusion

Application of cloud technologies enables reducing capital and current (operation) costs and increasing the level of equipment use, which amounts to 10% in government sector.

Cloud computing provides a more effective use of government agencies’ computing resources and at the same time, the resources are available for all government agencies and can be rationally distributed as workload changes. Balance of government agencies’ information and communication assets will be achieved.

Besides, efficiency of computing capacities per kilowatt-hour will increase, which in its turn will lead to rise in environmental friendliness of government agencies performance.

Upon transferring to functions on development of information systems and information assistance with mandatory support of domestic providers to IT-outsourcing, government agencies will be disengaged from non-profile purposes and assets and obtain a qualitative final result with the right for intellectual property use from software developers and service companies; competition in public sector in the field of informatization will grow together with Kazakhstani participation in procurement; budget saving on a nationwide scale will be achieved.

Thus, nowadays, there is a necessity to address the issues through changes in the Republic of Kazakhstan legislation in the field of informatization.

The basic practical relevance of the publication is the list of criteria for selection of cloud platform solution by which optimal cloud products will be chosen for the Republic of Kazakhstan e-Gov considering quality-price ratio, and also the framework of information and communication architecture will be introduced.

References

[1] K. K. Smitha, T. Thomas and K. Chitharanjan, “Cloud based e-governance system: A survey,” Procedia Engineering, vol. 38, pp. 3816-3823, 2012.

[2] F. Mohammed and O. Ibrahim, “Models of Adopting Cloud Computing in the E-Government Context: A Review,” Jurnal Teknologi, vol. 73, no. 2, 2015.

[3] N. C. Ferreira et al., “Challenges in the implementation of public electronic services: lessons from a regional-based study,” Journal of Business Economics and Management, vol. 16, no. 5, pp. 962-979, 2015.

[4] J. Liang, “Government cloud: enhancing efficiency of e-government and providing better public services,” In Service sciences (IJCSS), 2012 International Joint Conference on IEEE, pp. 261-265, 2012.

[5] A. Cordella and N. Tempini, “E-government and organizational change: Reappraising the role of ICT and bureaucracy in public service delivery,” Government Information Quarterly, vol. 32, no. 3, pp. 279-286, 2015.

[6] E. Nikulchev, E. Pluzhnik, D. Biryukov, O. Lukyanchikov and Simon Payain, “Experimental Study of the Cloud Architecture Selection for Effective Big Data Processing,” International Journal of Advanced Computer Science and Applications, vol. 6, no. 6, pp. 22-26, 2015.

[7] E. Nikulchev, O. Lukyanchikov, E. Pluzhnik and D. Biryukov, “Features Management and Middleware of Hybrid Cloud Infrastructures,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 1, pp. 30-36, 2016.

[8] The NIST Definition of Cloud Computing, Peter Mell, Timothy Grance, Computer Security Division Information Technology Laboratory National Institute of Standards and Technology Gaithersburg, MD 20899-8930, September 2011.

Comparative Study for Software Project Management Approaches and Change Management in the Project Monitoring & Controlling

Amira M. Gaber, Sherief Mazen, Ehab E. Hassanein
Department of Information System, Faculty of Computer and Information Sciences, Cairo University

***Abstract*—A software project encounters many changes during the software development life cycle. The key challenge is to control these changes and manage their impact on the project plan, budget, and implementation schedules. A well-developed change control process should assist the project manager and the responsible team in monitoring these changes.In this paper, we examine a number of approaches for project monitoring & control with different scenarios of project schedules. The comparison shows the effect of applying each approach on the cost and the time of the project. The evaluation illustrates that the *integrated software Project and Change Management (IPCM)* is a more efficient for providing more control in tracking change requests, and improve the performance monitoring process.**

Keywords—Software Engineering; Software Project Management; and Software Change Management

I. Introduction

One of the most important processes in developing any software project is project monitoring and control process. It controls the operation of the project according to the project plan. Also, it is one of the CMMI® process areas level 2 [1]. To measure the success of the software development project, the actual and the estimated project plans are compared and analyzed. When the performance of the project is deviated from the actual project plan, the corrective actions will be followed to achieve project success by using analysis result [2].

There are different monitoring approaches for the development of the software project; *the first approach* is the classic approach, which tracks different baselines on a regular basis of the project to estimate percentages of budget spent, work done and time elapsed [3]. *The second approach* is The Earned Value Analysis approach (EVA) which compares the planned amount of work with what is actual completed to determine if *Cost*, *Schedule* and *Work accomplished* are proceeding as planned or not. EVA measures the progress of the project by providing consistent numerical indicators, which can be used to evaluate, compare , forecast projects completion dates and final costs, and provide schedule and budget variances along the project [4].

Finally, the proposed Integrated Project and Change Management Framework (IPCM) tracking approach [4], which mainly integrates the activities of both software project management and software change management.

This integration useful for the following points:

1) When a change occurs, it will go through the whole change management process before execution as in Figure 1.

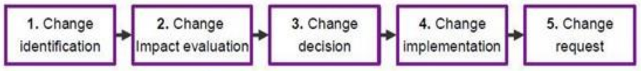


Fig. 1. the change management process [5]

2) Managerial reports will be prepared and delivered based on performance and improvements the activities of the project. When the Change Management Team (CMT) checks for the Change Request (CR), The CMT uses an Employee performance sheet for each CR responsible according to the completion date of the related CR (i.e. early CR means the responsible person will be rewarded; otherwise punished).

3) Providing an accurate Time Accumulator (TA) to record any CRs finished earlier than planned which will result in revising the project plan or any delayed CRs that will assist the next level of taking decisions to take corrective actions to the project plan.

4) Helping the higher level of management for better decision making in case of any budget deficit or surplus by using the Cost Accumulator (CA) to monitor the expenditures of the project plan.

5) Notifying the higher level of management in case of any possibility of early project completion, which might result in the revision of the operational processes that depend on the deadline of the project.

In the first two approaches, there is no integration and synchronization between software project management and change management. The change management must address in the project plan for:

- Including the procedure for handling Scope and variance changes.
- Creating forms to record and evaluate Change requests.
- Identifying the revision and approval Processes for changes.
- Adjusting the Change process to the Current plan.
- Avoiding the problem of unrecognized changes as team members may decide to add scope items to improve the final product, and don’t realize the added cost or time that will be incurred during execution. Therefore, ignoring to monitor the changed scope items

will definitely have an effect on the time and cost of the project.

So, the change management process should be used to allocate and evaluate changes before execution. When approved a change, the budget for the affected work package is modified [6].

This paper focuses on the project monitoring and control process area. This process area provides an understanding of the project's progress so that appropriate corrective actions can be taken in case of:

- The performance of the project deviates from the project plan.
- The project takes more time or cost or both than planned.
- The scope changes are not controlled which will affect the project schedule and budget. An effective change control process will help the project manager and team for managing the scope changes.

Also, we will compare the first two approaches with the Integrated Project and Change Management Framework (IPCM) approach during the project monitoring and controlling phase for the following purposes:

- Managing the whole software project and its change requests.
- Facilitating the distribution of tasks (change requests) and the availability of information about each change request
- Providing more control over the project by solving most of the technical issues instead of all of them reaches to management.Reporting the status of time and cost using accumulators to the management to take corrective actions.

The rest of paper we will evaluate each monitoring approach and shows its impact on the efficiency of the project with different schedule scenarios that might occur along the project life cycle.

II. Related Work

➔ *Monitoring Approaches*

A. *The Classic Approach*

After planning the project and building the project network, with each task assigned an estimated duration (start and end dates) and an estimated cost. The project manager starts with a baseline project schedule which remains fixed and never to be used during the project lifetime.

It is only used for comparison against other project schedules to check if tasks are executed according to the given schedule or not. The baseline project schedule can also be used to evaluate the project cost against the estimated budget. The project monitoring and controlling cycle is shown in Fig 2 [7].

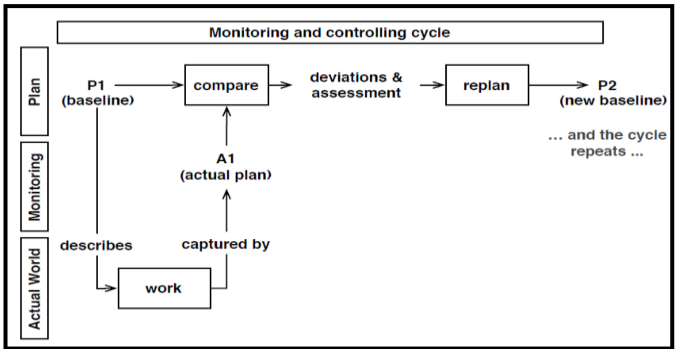


Fig. 2. Project Monitoring and Controlling Cycle [7]

B. *The Earned Value Analysis (EVA) Approach*

Earned Value Analysis (EVA) is a project management approach used to measure the project's progress in an objective manner. According to the Project Management Institute (PMI) [8] EVA provides an early warning of performance problems. Referring to EVA as EV (Earned Value) most often, the EV measures the project performance and the project progress by efficiently integrating the management of the three most important elements in a project; cost, schedule and scope. In fact, it calculates the cost and time performance indices of a project, estimates the completion cost and the completion time of a project, and measures the performance and the progress of a project by comparing the planned value and the actual costs of activities to their corresponding earned values [9].

EVA is used to

- Provide an integrated view of the project by measuring planned cost, earned value, and actual costs in terms of monetary values.
- Calculate schedule variance. A negative variance means that the project is behind schedule.
- Calculate cost variance. A negative variance means that the project is over budget.
- Calculate Cost Performance Index (CPI) that measures the amount of accomplished physical work against the money spent to accomplish that work.

○ CPI > 1, project is efficient
○ CPI < 1, project is inefficient

- Calculate Schedule Performance Index (SPI) that measures the amount of accomplished physical work against the amount of scheduled work.

○ SPI > 1, project is ahead of schedule
○ SPI < 1, project is behind schedule

- **Earned Value Computation**
- **50/50 Rule** (50% of planned value at start and 50% at end)
- **20/80 Rule** (20% at start and 80% at end)
- **0/100 Rule** (0% at start and 100% at end)[7]

Fig 3 shows the EVA Graph displaying cost cumulative curves: Actual, Planned, and Earned.

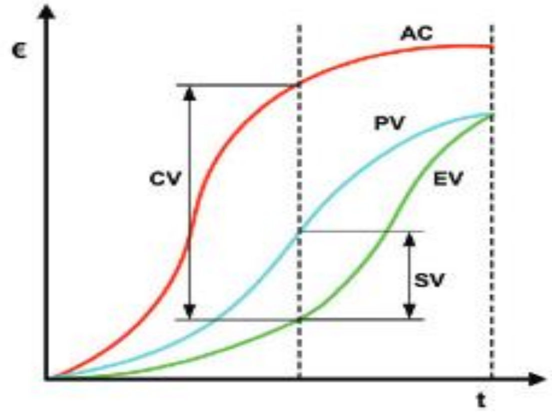


Fig. 3. Earned Value Analysis Graph [10]

C. The Integrated Project and Change Management (IPCM) Approach

The Integrated Project and Change Management (IPCM) approach integrates the activities of change management and project management as shown in Figure 4. It treats each task as a Change Request (CR). ***A Change Request (CR)*** is a formal proposal for an alteration associated with a task in the project plan. CR Types includes: Add Requirement, Update Requirement, Delete Requirement, New Version, and New Release [11]. IPCM provides an accumulation of resources in terms of time and cost using the Time Accumulator (TA), and the Cost Accumulator (CA). IPCM is used as an indicator to the status of the project plan in term of saved time and cost. Time Accumulator (TA) uses the following equation:

$TA_n=TA_{n-1}$ + *Estimated time (ET)* $_n$ – *Actual Time (AT)* $_n$

And Cost Accumulator (CA) uses:

$CA_n=CA_{n-1}$ + *Estimated cost (ES)* $_n$ – *Actual cost (AC)* $_n$

In addition, the IPCM Tracks change requests, and if there is any delayed or early finished tasks, then CMT will know who is responsible for that and put them in the project tracking performance sheet for either punishment or reward [4].

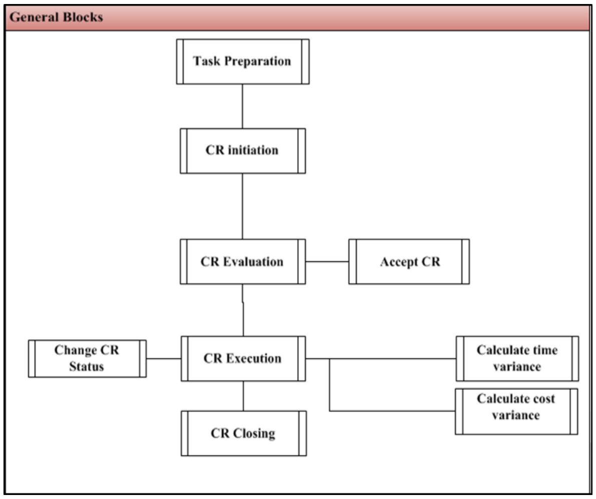


Fig. 4. The Overall Components of the Integrated Framework

III. IPCM Framework Evaluation

The IPCM Framework Evaluation is based on five possible scenarios that can occur along the project life cycle. A comparison and an evaluation are made between the different monitoring approaches in terms of both the cost and the time of the project for each of the possible scenarios.

Scenario 1: Some CRs on critical path finish earlier than planned

In this scenario, the project will be affected & finish earlier than the scheduled completion date.

Scenario 2: Some CRs on critical path finish earlier than planned and appear on another critical path

In this scenario, the project will be affected & finish earlier than the scheduled completion date. The affected CRs may appear on another critical path as a result of the early finish, thus the PM is responsible to revise the project plan and recalculate the critical path

Scenario 3: Some CRs on critical path finish later than planned, but within the contingency plan of the project

In this scenario, the PM determines the allowable delay time in the schedule that won't affect the project completion date. The allowable delay time should be defined in the contingency plan of the project. The PM checks if the delay is within the contingency plan or not.

Scenario 4: Some CRs on critical path finish later than planned, but not within the contingency plan of the project

In this scenario, the PM determines the allowable delay time in the schedule that won't affect the project completion date. The allowable delay time should be defined in the contingency plan of the project. The PM considers the delay in case of violation of the contingency plan.

Scenario 5: Some CRs finish later than planned, and don't affect the scheduled completion date.

Figure 4 and 5 show the network diagram and Gantt chart for a project consisting of 8 tasks. The project's start date is 03 December, 2014 and the completion date is 13 January, 2015 and the calculated critical path is A-B-D-E-G-H

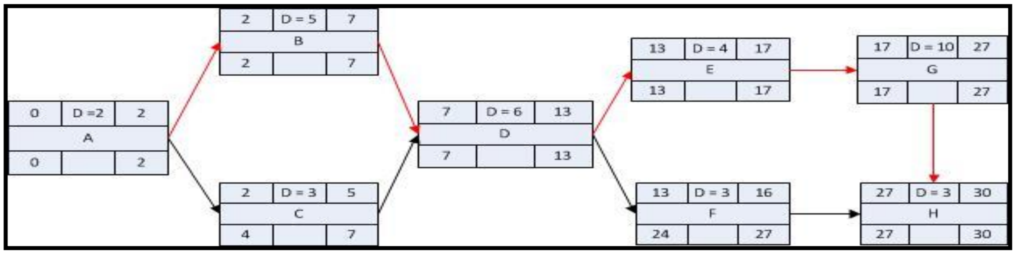
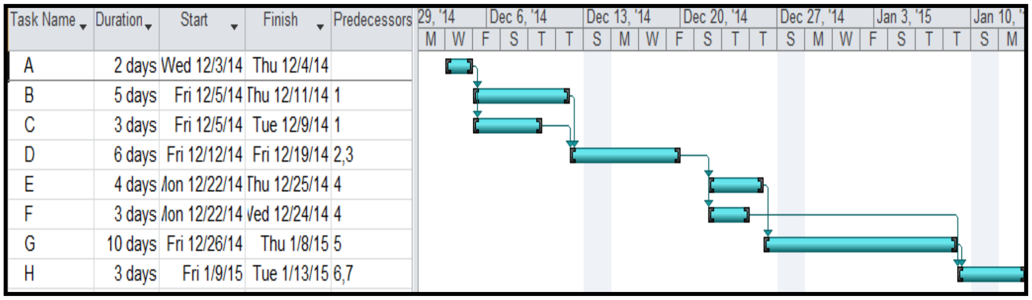


Fig. 5. Project Network Diagram



Task Name	Duration	Start	Finish	Predecessors
A	2 days	Wed 12/3/14	Thu 12/4/14	
B	5 days	Fri 12/5/14	Thu 12/11/14	1
C	3 days	Fri 12/5/14	Tue 12/9/14	1
D	6 days	Fri 12/12/14	Fri 12/19/14	2,3
E	4 days	/lon 12/22/14	Thu 12/25/14	4
F	3 days	/lon 12/22/14	Ved 12/24/14	4
G	10 days	Fri 12/26/14	Thu 1/8/15	5
H	3 days	Fri 1/9/15	Tue 1/13/15	6,7

Fig. 6. Project Gantt Chart

For these scenarios, this is the estimated and actual times for the tasks on the critical path and its Time accumulator for each scenario.

TABLE I. THE TIME ACCUMALATOR ,ESTIMATED TIME AND ACTUAL TIME FOR ALL TASKS FOR ALL SCENARIOS

		Scenario1		Scenario2		Scenario3		Scenario4		Scenario5	
Task (CRs)	**(ET)** *	**(AT)** **	**(TA)** ***	**(AT)** **	**(TA)** ***	**(AT)** **	**(TA)** ***	**(AT)** **	**(TA)** ***	**(AT)** **	**(TA)** ***
A	2	1	1	2	0	3	-1	1	1	3	-1
B	5	5	1	1	4	6	-2	7	-2	10	-4
D	6	3	4	6	4	7	-3	7	-3	4	-2
E	4	2	6	2	6	5	-4	8	-6	10	-4
G	10	8	8	5	11	11	-5	15	-11	4	2
H	3	3	8	3	11	4	-6	11	-19	3	2

* ***Estimated Time (ET)***
****** ***Actual Time (AT)***
****** ***Time Accumulator (TA)*** *TA_n=TA_{n-1} + (ET) – (AT)*

A. The Classic Approach

1) Scenario 1

As a result from the actual and the estimated time of this scenario, the PM has no clue about the tasks that finish earlier than planned. From the given table above, one day is saved for task A, 3 days for task D, 2 days for task E and 2 days for task G. The saved 8 days are not used to accelerate any tasks in the project, and the project ends up at the scheduled completion date.

2) Scenario 2

As a result from the actual and the estimated time of this scenario, the recalculation of the critical path is not regularly supported. Figure 6 shows the new recalculated critical path that needs to be monitored by the PM.

3) Scenario 3, 4:

As a result from the actual and the estimated time of this scenario, the PM checks the contingency plan and takes any necessary corrective actions.

4) Scenario 5:

As a result from the actual and the estimated time of this scenario, the PM monitors each task at the end of its delivery, and takes corrective actions to avoid any delay of the scheduled completion date.

In the classic approach, the PM has no information about the tasks that finish earlier than planned. Therefore, the resources of the project (saved time and cost) are not utilized efficiently, and can't be used to affect other tasks.

B. The IPCM Approach

In this approach, the Change Management Team (CMT) tracks the actual time against the planned time for each CR after execution. The decision making depends on the numeric results generated by the Time Accumulator for all the CRs [4].

1) Scenario 1:

According to Equation (1); tasks A, D, E and G finished earlier than their due dates: the time accumulator for task A is 1 day which means that the PM is alerted by the CMT to revise the project plan and start the next CR 1 day earlier than planned. The TA for tasks B and D are 1 and 4, then the PM starts the next CR 4 days earlier than planned. In the same way, tasks E, and G result in starting next CRs by 6 and 8 days earlier than the planned start dates. At the end, the project finishes 8 days earlier than planned.

2) Scenario 2

According to Equation (1); tasks B, E, G and H finished earlier than their due dates: the TA for task B is 4 days which means that the PM is alerted by the CMT to revise the project plan and start the next CR 4 days earlier than planned. In the same way, tasks E, and G result in starting the next CR 11 days earlier. The CMT detects that task F deviates from the estimated time by 6 days (3 – 9 = - 6) which leads to a change of the critical path. Therefore, The CMT asks the PM to revise the project plan and recalculate the critical path. At the end, the project finishes 11 days earlier than planned with a new critical path (A-B-D-F-H).

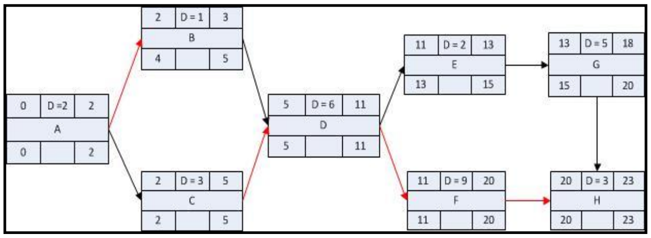


Fig. 7. Gantt Chart for the Recalculated Critical Path

3) Scenario 3

According to Equation (1); tasks A,B,D, E, G and H are delayed than planned: the TA for tasks A, B, D, E, G and H are -1,-2,-3,-4,-5,-6, and the CMT alerts the PM since the first delayed task (task A in the given example). In case of a delay, the PM checks the contingency plan to take corrective actions or raise the issue to a higher level of management. At the end, the project is delayed 6 days after the original plan with no violation to the contingency plan.

4) Scenario 4

According to Equation (1); tasks A, B, D, E, G and H are delayed than scheduled: the TA for tasks A, B, D, E, G and H are 1,-2,-3,-6,-11,-19, and the CMT alerts the PM since the first delayed task. In case of a delay, the PM checks the contingency plan if the delays are within range or not to take corrective actions or raise the issue to a higher level of management. In the given example, the project is delayed 19 days from the original plan which violates the contingency plan.

5) Scenario 5

According to Equation (1); tasks A,B, D, E, G and H are delayed after their due dates: the TA for tasks A, B, D, E, G and H are -1,-4,-2,-4,2,2. Using the TA for all tasks, the project is not delayed, and the delayed tasks can borrow some time from the predecessor tasks that finish earlier than their due dates. Therefore, tasks can be delayed from their planned due dates without affecting the whole project. In this scenario, the IPCM approach takes advantage of the accumulated time and cost to accelerate the delayed CRs. This advantage is unique to this approach that can help crashing the project.

In the IPCM approach, tracking the saved time after executing each CR leads to several advantages. One advantage is that accurate estimation of both the cost and time variances of the project. Another benefit is that tasks can finish earlier than planned. It is important to note that the CMT puts everyone responsible for accomplishing a task in the project tracking performance sheet for further punishment or reward.

C. The Earned Value Analysis Approach

In this approach, the PM uses a suitable rule to calculate "% complete" of each task [7]. There are three rules to do that: *first*; 50%-50% rule; the PM grants 50% of the time to the actual start date and the task remains marked as "50% completed" until an actual finish is recognized when the remaining 50% is granted and the task become "100% completed". *Second*; 20%-80% rule; the PM grants 20% of the time to the actual start date with the remaining 80% granted at the actual finish of the task. *Third*; 0%-100 % rule; grants nothing until the work package is 100% completed.

1) Scenario 1

The PM uses the suitable rule to be applied on the given case. Regardless of the applied rule, the EVA calculates the Schedule Performance Index (SPI) based on how far the project is ahead of the schedule. So, the EVA detects an early finish based on applying the suitable rule.

2) Scenario 2

The PM uses the suitable rule to be applied on the given case. Regardless of the applied rule, the EVA calculates the Schedule Performance Index (SPI) based on how far the project is ahead of the schedule. In this scenario, the PM is not aware of any emerging critical paths. Therefore, the EVA detects an early finish based on applying the suitable rule without recalculating of the critical path.

3) Scenario 3, 4

The PM uses the suitable rule to be applied on the given case. Regardless of the applied rule, the EVA calculates the Schedule Performance Index (SPI) based on how far the project is ahead of the schedule. The EVA approach may give better prediction than the IPCM approach in case if tasks will be delayed before completion, but this depends on the applied rule. In the IPCM approach, it predicts the delay of the tasks after completion. In addition, the IPCM approach calculates the actual delay more accurately than the EVA approach. The IPCM approach solves the delay problem by dividing the MCR into sub CRs that can be controlled and executed in parallel. It is important to note that no awareness of the contingency plan is available to be checked by the PM.

4) Scenario 5

The tasks can be delayed without affecting the whole project. The IPCM approach gives the advantage to accumulate the time and cost to give the delayed CRs extra time and cost from the accumulators to accomplish and recover the delay. This advantage can help the project in crashing. This scenario is not supported in both the classic and EVA approaches.

The EVA approach is better than the classic approach in tracking changes along the project. EVA can provide indices to improve the oversight of the project, and is better than the

IPCM approach in the early prediction of any possibility of an early completion of any task. EVA has the disadvantage of not using the change management process along the project [6] while the IPCM approach provides index variables to monitor the project, and integrates the change management and project activities.

IV. Conclusion

In this paper, we made a comparison between the classic approach, the IPCM approach and the EVA approach. The comparison proved that to improve the project monitoring and controlling phases, the change management and project management activities should be integrated. The IPCM Framework Evaluation of the comparison has different perspectives:

A. *From the tasks perspective*

The IPCM approach deals with each task as a Change Request (CR) and integrates the CR with the change management process. On the other hand, the classic and the EVA approaches are not well defined the CR procedure.

B. *From the quality perspective*

The IPCM approach only integrates the software project and change management activities to achieve the quality. However, the EVA and the classic approaches don't include any quality guidelines within the project.

C. *From the project resources perspective*

The IPCM approach effectively utilizes the project resources while the EVA approach does partial utilization and no utilization at all in the classic approach.

D. *From the time perspective*

The IPCM approach monitors all tasks or CRs along the project after completion dates and gives accurate values for any deviations from the project schedule. On the other hand, the classic approach monitors the project as baselines on regular basis. However, the EVA monitors the project based on rules: 0-100% or 20-80% or 50-50 %, which determine if the project is ahead or behind schedule.

E. *From the cost perspective*

The IPCM approach monitors all tasks or CRs along the project after completion dates and give accurate values for any cost deviations.

The classical approach monitors the project as baselines on regular basis with respect to the estimated and actual costs, but depends on the experience of the project while the EVA Approach controls the project on some basis rules which determine if the project needs to be recovered or not.

V. Futeure Work

There are some open problems are mentioned below to be considered in future:

- The CMT alerts the project manager to revise the project plan and recalculate the critical path when a change occurs. A simple enhancement can be done by avoiding the recalculation of the critical path from scratch because it is a waste of time. The recalculation should only include activities with a change in duration, which can be found by doing a simple time intersection of the project activities.

- The framework can be extended to take into consideration the people side of the project in the same way taken by the ADKAR Model in order to maximize the profit of the integration.

- The proposed framework works along the development phase only. An interesting point to consider is to extend the framework to study the maintenance phase activities after software delivery.

References

[1] CMU/SEI, “Capability Maturity Model Integration,CMMI for Development, Version 1.3”, CMU/SEI-2010 RT-033, Pittsburgh, Software Engineering Institute Carnegie Mellon University, 2010.

[2] Sunart Wanapaisan, Taratip Suwannasart, and Apinporn Methawachananont ," An Approach for Monitoring Software Development Using Timesheet and Project Plan"Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong

[3] Malte Foegen," Project Management 08 – Project Monitoring and Control", wibas IT Maturity Services GmbH © 2007

[4] Amira M. Gaber, Sherif Mazen, Ehab E.Hassanein,"Framework for Integrating Software Project Tasks and Change Requests ", International Journal of Computer Applications (0975 – 8887) Volume 125 – No.12, September 2015

[5] Henri Elemo," DEFINING SOFTWARE CONFIGURATION MANAGEMENT FOR PRODUCT DEVELOPMENT ", May 30, 2008

[6] Joseph A Lukas," Earned Value Analysis –Why it Doesn't Work", AACE INTERNATIONAL TRANSACTIONS, 2008

[7] Ursula Kuehn, EVP," Earned Value Management - Why Am I Being Forced to Do It ? ”, AACE International Transactions, 2007

[8] Anbari, F., "Earned value project management method and extensions", Project Management Journal, volume 34, issue (4), 2003, P.12–23.

[9] Leila Moslemi Naeni, Shahram Shadrokh, Amir Salehipour," A fuzzy approach for the earned value management ", International Journal of Project Management, Volume 32, Issue 4, May 2014, Pages 709-716

[10] Fernando Acebes, Javier Pajares, José Manuel Galán, Adolfo López-Paredes," Beyond Earned Value Management: A Graphical Framework for Integrated Cost, Schedule and Risk Monitoring ", Procedia - Social and Behavioral Sciences volume 74, (2013), P. 181 – 189

[11] A. R. M. Nordin, S. Suhailan, "Managing Software Change Request Process: Temporal Data Approach "International Journal of Computer Science and Security, (IJCSS) Volume (3), Issue (3), 2009

N-ary Relations of Association in Class Diagrams: Design Patterns

Sergievskiy Maxim
National Research Nuclear University MEPhI
Moscow Technological Institute
Moscow, Russia

***Abstract*—Most of the technology of object-oriented development relies on the use of UML diagrams, in particular, class diagrams. CASE tools, used for automation of object-oriented development, often do not support n-ary associations in the class diagrams, and their implementation in the form of program code in contrast to binary rather time-consuming. The article will show how in some cases it is possible to move from the n-ary association between classes to binary and how can reduce the number of objects. The rules to transform models, that contain n-ary association, will be presented in the form of design patterns. Proposed three new design patterns can be used in the process of developing software systems. These patterns describe transformations of n-ary (often ternary) associations occur between classes in binary and the introduction of additional classes and binary association with the aim of optimizing the model.**

Keywords—UML; class diagram; multiplicity; ternary association; n-ary association; class-association; design pattern; object

I. Introduction

As it is known, UML is the standard tool for modeling software systems [1], [2], [3], [4]. Most of the object-oriented technology developments use general capabilities of this language. The design stage primarily uses class diagrams from the UML. They describe the model of a software system reflecting the main parameters of the subject area. In class diagrams, the base relationship is an association relationship. This is complex structural relation, which describes links between the objects of different classes of software system. At the later stage software system model in the form of a class diagrams will be transformed into a logical database model and object-oriented application code. It is important that a substantial part of the code can be generated automatically with the help of CASE tools. The majority of CASE tools do not support n-ary (in particular, ternary) association relationships in the class diagrams [2], [5]. Also, n-ary association, unlike binary, is a time consuming (this does not apply to databases). The article will demonstrate how in some cases it is possible to move from the n-ary association between classes (often ternary) to binary, and how you can reduce the number of potential objects of class-associations. Guidelines for the conversion of models containing n-ary association will be shown in the form of design patterns [6].

II. Replacing Ternary Association on Binary and Class-Association

Let assume that in the class diagram there is a ternary association, i.e. the association, which involves three objects. For example, take certain objects belonging to three different classes: STUDENT, SUBJECT and LECTURER (see Fig. 1). We define multiplicities for the classes from this association: STUDENT (1..*) SUBJECT (1..*), LECTURER (1).

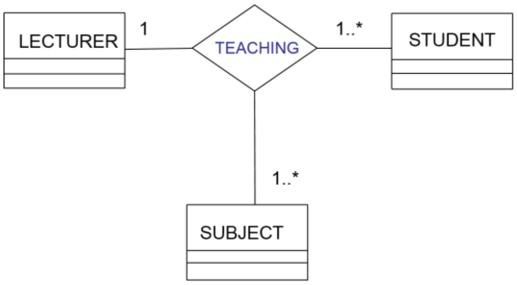


Fig. 1. The ternary association between the classes STUDENT, SUBJECT and LECTURER

The multiplicity of the association in relation to the class LECTURER is (1) because any fixed pair of objects the STUDENT and the SUBJECT corresponds to only one object class LECTURER. Each LECTURER may teach one subject with multiple students, so the multiplicity of the association in relation to the class of the STUDENT is equal to (1..*); the same lecturer can read several courses to any single student, so the multiplicity of the association in relation to the class SUBJECT equals (1..*).

Pattern_1. Assume that in the n-ary association there is a class with multiplicity (1). Then n-ary association can be replaced with a combination of (n-1)-ary association and class-association.

Proving the above is quite simple: show how it would work for a ternary association. Let's combine two classes with multiplicities, different from (1), with normal binary association. Then any two connected objects of these classes will correspond to exactly one object of the third class, which we can without loss of generality refer to the class-association. Thus, the class-association will replace the third class of ternary association. Moreover, in this class we can include attributes originally related to the ternary association.

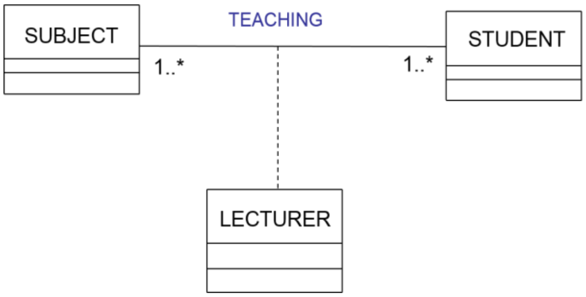


Fig. 2. Replacing ternary association on binary and class-association

Applying Pattern_1 will give us a new class diagram not containing ternary association (see Fig. 2). Such cases are fairly common. Here is another example of ternary association, which multiplicity to one of the classes (1): PLAYER (11..*), SEASON (*), CLUB (1).

The multiplicity associated with the class CLUB is equal to (1), because any player may change club only in the offseason.

The third example describes the case when all objects of the ternary associations belong to the same class - PEOPLE: this refers to the relation Father – Mother – Child (see Fig. 3). Here the multiplicities are as follows: PERSON (Father) - (1), PERSON (Mother) - (1), PERSON (Child) - (*).

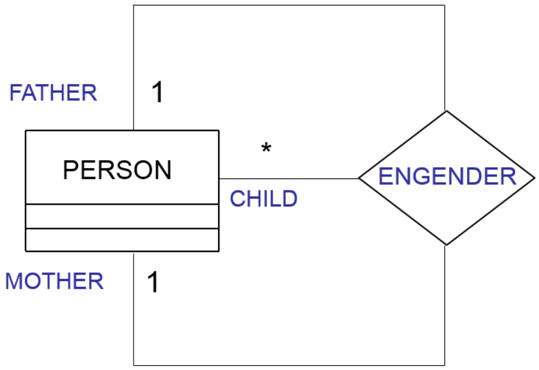


Fig. 3. Ternary association Engender (Father, Mother, Child)

III. Using Of Additional Classes

Let’s come back to the first example. For a ternary association, as for any other, there may be relevant attributes. In this case it can be starting and finishing time of the class session and the classroom number. Let’s try to strip out a subclass from a defined class of objects, which has commonalities in relation to these attributes. For example, all students are divided into different groups, for which similar classes are taking place at the same time. In this case there is a class - GROUP.

The ternary association between the classes STUDENT – SUBJECT – LECTURER is transformed into a ternary association GROUP – SUBJECT – LECTURER and the binary association STUDENT – GROUP. We define multiplicities for the classes involved in the new ternary association: GROUP (1..*), SUBJECT (1..*), LECTURER (1).

In this case, using Pattern_1, we can get the combination of simple binary association and class-association, then a class diagram will have a different appearance (see Fig. 4). The advantage of this chart is that a number of objects – instances of class-association LECTURER - will be reduced.

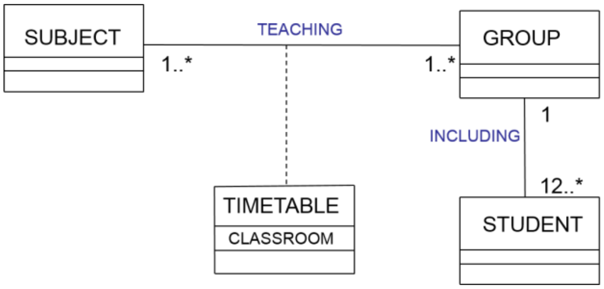


Fig. 4. The introduction of additional (to the class STUDENT) class GROUP

Based on the above, you have the following options to describe appropriate design pattern:

Pattern_2. Assume that for two or more classes there is a class-association with one or more attributes. If it is possible to split the objects of one of the classes into a subsets for which the attribute values of the class-association will be the same, then another class should be created with the association with multiplicities (1) and (*) to the first class.

But still redundancy in the form of repeated instances of a class-association LECTURER remains (in case the lecturer will teach the same subject to multiple streams). Let’s introduce another class-association – TIMETABLE. This is a class, not an attribute, because it can include already specified attributes: the lecture starting time, finishing time, classroom number. This new class will be connected to a normal class-association LECTURER. New association will have the following multiplicities: (1) – for class LECTURER and (*) – for the class TIMETABLE (see Fig. 5). This solution will help to reduce the number of instances of a class-association LECTURER. Here we use an operation, similar to the operation of standard normalization from the database theory [7].

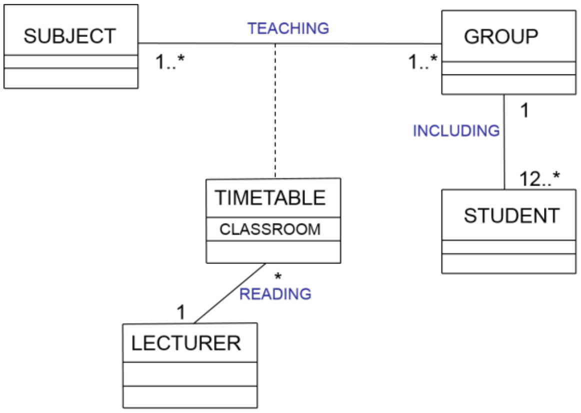


Fig. 5. Introduction of the class-association TIMETABLE

IV. Replacement of the N-Ary Association with Binary Ones

Let's describe another commonly occurring type of n-ary association that can exist between classes, or rather between objects of classes. Assume that on one side there is a single object of one class against involved in the association and on the other side there is a random number of objects of the second class. That means that the relation is defined as a set of tuples of variable length.

Let's give examples of such relations from the real subject domains [8], [9]. There are two classes: the DISEASE and GENOTYPIC_TRAIT. Relations between objects of these classes can be described as follows: object class DISEASE may be associated with any number of objects GENOTYPIC_TRAIT, and tuples of different lengths from 1 to N

(D1, GT1, ... , GTN),

characterized by an additional attribute - the probability of disease given the presence of these genotypic traits. It turns out that the tuples of the relations are of the form of:

(D1, GT1), (D1, GT1, GT2), (D2, GT2), (D2, GT3), (D2, GT1, GT2, GT3)

In case we are to show relationship between objects graphically, the result is that one object class the DISEASE may be associated with one object

GENOTYPIC_TRAIT more than once. That means the object diagram for the described relationship may be the following (see Fig. 6).

Let us go through another example. When a number of participants in a certain project are defined, the following problem often arises. The project may involve staff in different combinations. For example, performing Project_1 can attract Smit and Jones, participation could be limited to Smit only, or you can even add Clark. Resources spent in each of the above cases (time, finances, etc.) will vary, and can be added as additional attributes. Then between classes PROJECT and EMPLOYEE also encounter the recently described type associative relationship.

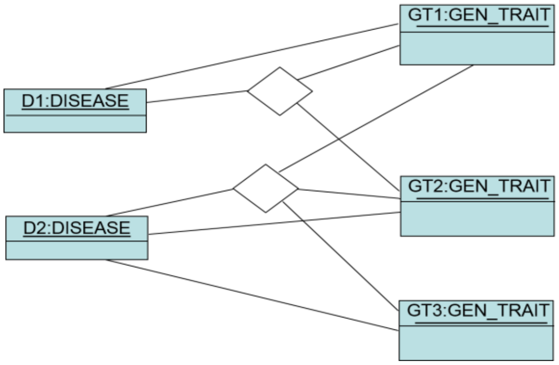


Fig. 6. Object diagram

It is obvious that means to describe and specify this relationship in UML is not. But we can solve this problem by entering additional class GROUP_GENES. In this case

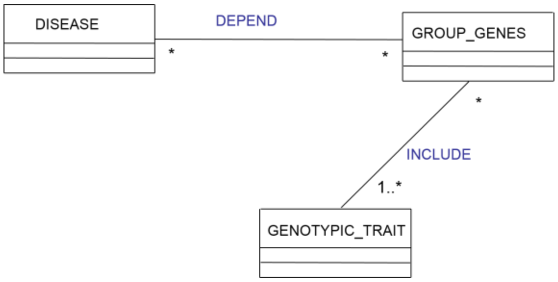


Fig. 7. Class diagram for the model genetic diseases

there is a class diagram that includes classes in addition to DISEASE and GENOTYPIC_TRAIT: the class

GROUP_GENES (see Fig. 7) and, if need, class-association between DISEASE and GROUP_GENES classes to store additional attributes.

Pattern_3. Assume that the n-ary association involves one object of the first class and random number of objects of the second class. In this case a third class should be entered to group the objects of the second class and associate it semantically different binary relations of the association with the first and second classes.

V. N-Ary Associations (N>3)

Since substantial part of this article is devoted to the n-ary association relationships, let's give examples of n-ary associations with n>3. Obviously, in real domain areas such associations are often met.

Take the domain area associated with the deliveries to the warehouses of different goods produced by different companies. We can distinguish four classes:

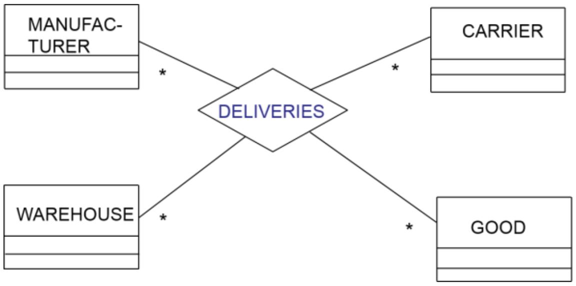


Fig. 8. Tetrary association

MANUFACTURER, WAREHOUSE, GOOD, CARRIER. Objects of these classes will be linked by relationship of association, which may have additional attributes such as delivery time, quantity, invoice number, etc. All the multiplicities in this case will be equal to (*) (see Fig. 8).

VI. Conclusion

The article describes three new design patterns which could be applied in developing software systems. These templates show transformations of n-ary (often ternary) associations, which occur between classes, into binary, as well as the introduction of additional classes and binary relations with the aim of optimizing the model. This task is very relevant given that in real domains n-ary associations are very common, and system analysts often confront with these facts

[10]. Thus, applying new templates already at the design stage make it possible: firstly to get rid of the complexity associated with modeling and realization of n-ary associations, and secondly to minimize the number objects, arising in the software system operating process.

References

[1] G.Booch, J.Rumbaugh, I. Jacobson, ”Unified Modeling Language”, Addison-Wesley, 2004

[2] L.Maciaszek, ”Requirements Analysis and System Design”, Addison-Wesley, 2004

[3] C.Larman, ”Applying UML and patterns”, Prentice Hall, 2005 G. Booch,”Object-Oriented Analysis and Design with Applications”, Addison-Wesley, 2007

[4] Methodical Materials IBM, https://www14.software.ibm.com

[5] E.Gamma, R.Johnson, Helm R., J.Vlissides, ”Design Patterns. Elements of Reusable Object-Oriented Software”, Addison-Wesley, 2001

[6] C.J. Date, ”An Introduction to Database Systems”, Addison-Wesley, 2004

[7] A. Konkin, M. Sergievskiy,”Integrating Bayesian Networks and Decision Trees for Calculating Probabilistic Rate of Complex Diseases”, Biology and Medicine 7(3): BM-119-15-4 pages, 2015

[8] O. Lukyanchikov, E. Pluzhnik, D. Biryukov, E. Nikulchev, ” Features Management and Middleware of Hybrid Cloud Infrastructures”, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, pp. 30-36, 2016

[9] G. Grenova, J. Llorens, P. Martrınez, ”The meaning of multiplicity of n-ary associations in UML”, Software System Modeling, No 1, pp. 86-97, 2002

Towards Domain Ontology Creation Based on a Taxonomy Structure in Computer Vision

Mansouri fatimaezzahra
Computer Science Engineering Laboratory
Department of Computer science
Marrakesh, Morocco

Elfazziki abdelaziz
Computer Science Engineering Laboratory
Department of Computer science
Marrakesh, Morocco

Sadgal mohamed
Computer Science Engineering Laboratory
Department of Computer science
Marrakesh, Morocco

Benchikhi loubna
Computer Science Engineering Laboratory
Department of Computer science
Marrakesh, Morocco

***Abstract*—In computer vision to create a knowledge base usable by information systems, we need a data structure facilitating the information access. Artificial intelligence community uses the ontologies to structure and represent the domain knowledge. This information structure can be used as a database of many geographic information systems (GIS) or information systems treating real objects for example road scenes, besides it can be utilized by other systems. For this, we provide a process to create a taxonomy structure based on new hierarchical image clustering method. The hierarchical relation is based on visual object features and contributes to build domain ontology.**

***Keywords*—Domain Ontology; Categorization; Taxonomy; Road scenes; Computer vision**

I. Introduction

In this paper, we treat modeling problems and representation of road scenes content using knowledge engineering methods. We seek to define and organize the knowledge of a field of study (road scenes our field of study), through ontologies, which will allow us to define the domain concepts and relations between them. The creation of a domain ontology will facilitate the task of object recognition, similarities search and it will facilitate the decision-making task.

The domain ontology is a common vocabulary for researchers who need to share information on a subject as concepts and the relations between them. It allows a knowledge formal representation of a specific scientific field.

The knowledge organization by classes minimizes the information complexity and improves the efficiency of information processing. This process also allows new elements classification, besides the information subsequent use in decision-making, evaluative judgments, selection and generation of new knowledge.

We adopt a structuring method through image classification using a measuring function based on the cue-validity of attributes [1], which evaluates each partition and preserves the best according to certain criteria. Our function allows us to generate a tree structure (taxonomic tree) with several different levels from each other by their level of abstraction and precision. This classification as a taxonomic tree will be the basis of our ontology of hierarchy.

The objective of this research is to propose a methodology for automatic generation of a taxonomic tree as a basis for visual objects ontology building. This generation uses an assessment that can select each level of the tree in accordance with the criteria of the categories accuracy and the recognition time. It is based on the characterization of objects by the visual attributes and organizing them hierarchically by techniques of non-supervised learning.

This work is about the object recognition in cognitive vision. In the knowledge acquisition phase of the domain of study using a class hierarchy of objects and subclasses.

Each class will be described in terms of visual concepts (shape, color, texture) provided by an ontology. Each visual concept of this ontology is associated with descriptors, the semantic gap is reduced to the expert who will intervene to add relations between concepts and place the objects in their membership classes.

All this process will facilitate the decision making in practice. For road scenes for example it will help driving by detecting the obstacles in real time.

Section 2 of this article presents the state of art we present related works, followed by section 3, which evoke the ontology building methods, section 4 details our approach to create an automatic domain ontology based on a taxonomy structure. We present an example in section 5, then a conclusion to close the article.

II. State of the Art

Ontology is the study of the knowledge about the world. We define ontology also as the study of the organization and the nature of the world, regardless of their perception [2]. Sowa suggests that the subject of ontology is the study of things categories that exist or may exist in a certain area [3]. With the emergence of the knowledge engineering, the community introduces ontology in artificial Intelligence as a response to the problems of knowledge representation and manipulation within computer systems. The most common definition of ontologies is the Gruber definition [4]; he defines ontology as explicit formal specifications of the terms of a domain and relations between them.

In image processing as in other areas, we use ontologies for knowledge structuring. Some studies use ontologies from the stage of images segmentation. In [5] ontologies include parameters for the segmentation algorithm and the potential label areas. In visual ontology, the description of concepts is mainly based on the geometric characteristics. After initial segmentation, we adjust the segments to get closer to their description in the ontology.

In [6] Hindle presents an early work on automated taxonomy building in which the names are grouped into classes. Hearst seminal work on the use of linguistic models also aimed to discover the taxonomic relations [7]. Recently, Reinberger and Spyns [8] present an application of clustering techniques in the biomedical field. In [9] we find a view of all clustering approaches for ontological learning structures.

Vision systems based on knowledge have proven to be effective for complex object recognition and scene interpretation. They offer the possibility of reuse and extensibility. Furthermore, in knowledge-based systems, we separate domain knowledge from the image processing knowledge. This implies better traceability of the different sub-problems. In literature, we find a variety of statistical approaches based on machine learning for annotating automatically image regions: SVM, decision trees, artificial neural networks, Bayesian networks. These approaches [10] learn matching functions between the characteristics and the regions classes. Although describing well the visual image's content, statistical methods do not adequately represent the picture's meaning as perceived by humans because semantic is limited to the learning results of the function linking low-level features to high-level concepts. These performances also depend on the number of classes learned.

Besides the statistical methods, some works [11] propose to use the domain concepts to annotate images: free annotation where no vocabulary is predefined, annotation by key-words in a set of words (or concepts) is proposed to the user and annotation by ontology where a set of words and the relations between them are provided to the user. Using ontology aims to different goals: unified description of image characteristics, visual description of the relations between characteristics (lines, region...), use of contextual information and finally the reconciliation between visual and semantic level. In the purpose of knowledge formalization, Neumann [12] proposed to model the scenes using a logical description. The main contribution of logical descriptions is to avoid mistakes when modeling knowledge and inference are intuitively constructed. Clouard [13] proposed an image processing ontology example. It contains 279 concepts, 42 roles and restrictions 192. However, the domain knowledge is not present, and image context and the user's knowledge are not taken into account. These approaches and ontologies have limitations, especially concerning their reuse.

For ontology construction, only a few automatic methods are proposed [14, 15, 16, 17]. Elliman [16] propose a method for ontologies construction to represent a set of web pages on a specified website. We use the map organization to build the hierarchy. In our case, we automatically modify the tree and the label organization in the hierarchy nodes. Bodner proposes a construction method based on a statistical hierarchy [14]. In [15] Hoothe offer various clustering techniques to illustrate the text using ontologies. All hierarchies will be constructed for multiple viewing only not in the ontology construction purposes. In addition, all these ontology construction methods are used in the text field; however, we address this problem in the *image domain.*

Regarding image processing, Latifur Khan in [17] proposes a method of ontology automatic construction from the automatic classification algorithm with a similarity based on color and shape. The results lead to a precision measurement on 6 categories known in advance.

III. Ontology Structuring and Building

A. Definitions

An ontology is an explicit formal description of concepts (also called Classes) in a given field, properties of each concept describing attributes and the attribute restrictions. An ontology and all the class's individual instances constitute a knowledge base. There is actually a fine line between the end of an ontology and the beginning of a knowledge base. How to construct an ontology is still subject to much discussion in the community. Our understanding of the different contributions made so far is that there is a three distinguish construction options:

- *Bottom-up approach*: The ontology is constructed by generalization starting from the low taxonomic concepts layers. This approach encourages the creation of specific and adapted ontologies.
- *Down approach*: The ontology is built by starting with specialization in high taxonomic concepts layers. This approach encourages the reuse of ontologies.
- *Centrifugal approach*: Priority is given to the identification of the central concepts in the application that will be generalized and specialized to complete ontology. This approach encourages the emergence of thematic domains in the ontology and promotes modularity.

To formally present an ontology, we should give concepts a lexicon and explicit the relations between them.

According to Gruber, M. Uschold [18]:

- An ontology involves or includes a certain view of the world for a given domain. This view is conceived as a *set of concepts*, their *definitions* and their *inter-relations*. This is called a *conceptualization*.
- An ontology can take different forms, but it necessarily includes a term vocabulary and specification of their meaning.

- An ontology is a specification, making partial account of a conceptualization.

B. *Formal structure of an ontology*

The quintuplet O= {C, R, H^C, rel, A^O} presented by Steffen Staab in [19] is the ontology structure:

- C and R are disjoint sets of concepts and relations
- H^C Concepts hierarchy (taxonomy): $H^C \subseteq C \times C$, $H^C(C_1, C_2)$ means that C_1 is a sub-concept of C_2 (oriented relation)
- Rel: the relation rel: R → C x C (define semantic relation) with 2 associated functions:

1. dom: $R \rightarrow C$ with dom(R):= $\prod 1$(rel(R))
2. range: $R \rightarrow C$ with range(R):= $\prod 2$(rel(R)) co-domain
3. rel(R) = (C_1,C_2) is written like $R(C_1,C_2)$

Often it defines abstract ontology.

We can add a lexicon to the ontology O:= {C, R, H^C, rel, A^O} which is a quadruplet L:= {L^C, L^R, F, G}

- L^C and L^R: disjoint sets of lexical concepts and relations
- F, G: 2 relations called references F (L^C(for the concepts), G (L^R x R (for the relations), fork L $\in L^C$: F (L) = {C (C / (L, C) (F}, F^{-1} (L) = {L (L / (L, C) (F}
- Idem For G and G^{-1}

Then we obtain the concrete ontology: couple (O, L)

In ontologies, inter-relations connect all the concepts. If there is an inter-relations R, between concepts Ci and Cj, then there is also an inter-relations R′ between concepts Cj and Ci. In Figure 1, we represent inter-relations by labeled arcs/links. We used three kinds of inter-relations to create the ontology: Is-a, Instance-of, and Part-of. These correspond to key abstraction primitives in object-based and semantic data models [20].

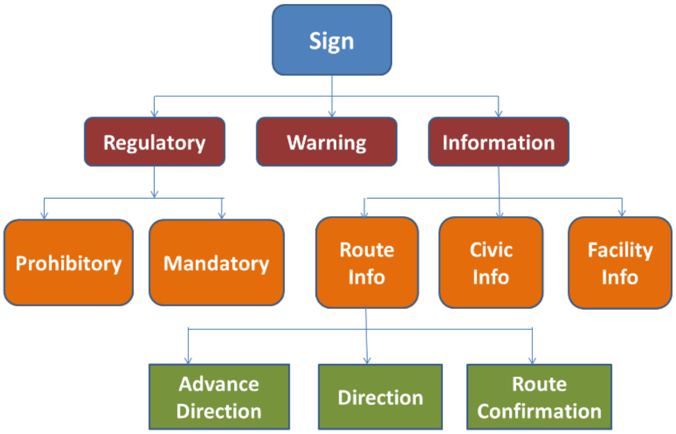


Fig. 1. Example of a road panels' tree

Figure 1 illustrates an example of an ontology for the road scenes domain. We obtain this ontology from generic terminology or from experts. The directed acyclic graph (DAG) describes it. Each node in the DAG represents a concept. The concepts in the ontology contain a label name and feature vector. A feature vector is a set of features and their weights. Each function can represent an object in an image.

Our contribution is to work on the concepts hierarchy, thus the automatic generation of H^c from clustering techniques on the objects aspects.

C. *General ontology construction scheme*

In a taxonomy, we organize the controlled vocabulary in a simple hierarchical format. This hierarchy often corresponds to a specialization. There is therefore a defined link between a term's vocabulary and its children. This link gives an extra meaning. From a controlled vocabulary, we can go to an organized vocabulary. In fact, even a lexicon or a taxonomy are forms of unformulated grammar ontology. When establishing a category and a hierarchy of this categorization, we establish dependencies between these terms. These hierarchies have meaning outside the vocabulary itself. An ontology corresponds to a controlled and organized vocabulary and the explicit formalization of relations created between the various terms of vocabulary. We can consider that the taxonomy is a semi-formal representation of the world while ontology gives a formal representation, the same representation will give us an operational model.

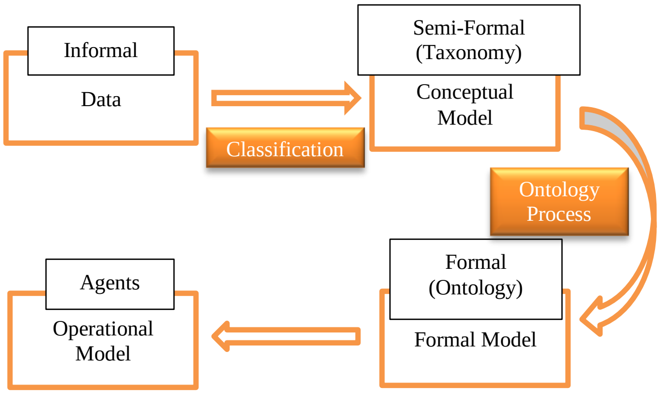


Fig. 2. From informal to formal model

IV. Proposed Approach

A. *General architecture of the proposed system*

Images number and resolution are increasing. We have then a tremendous increasing amount of information available but not exploitable. It is therefore necessary to develop semi-automatic processes, facilitating the organization of visual objects to minimize the recognition time while keeping the accuracy. In our ontology construction process, we adopt a down approach going from the top of the taxonomic tree.

Our goal is to present, formally, the concepts of the road scene domain using the ontology concepts to present knowledge. For this, we will be using the result of partitioning objects (taxonomic tree) while keeping the most consistent partitions giving the best recognition.

Our approach progress as the following figure:

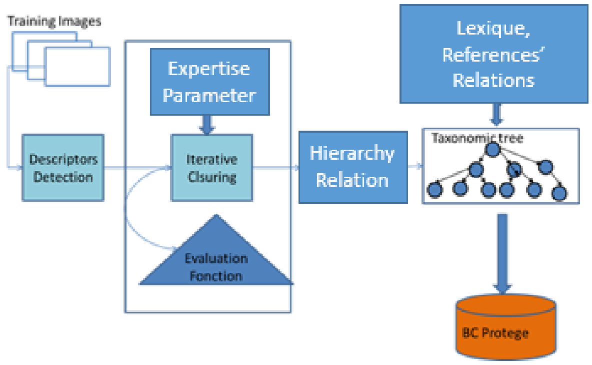


Fig. 3. General system architecture

It begins first with a visual features detection step of all images, then we go to the clustering step using an iterative clustering algorithm while evaluating each partition obtained, which will allow us to create a hierarchy that produces a taxonomic tree. After that, we create an ontology by adding a lexicon and relations between the taxonomy nodes. This ontology will be our knowledge base of the domain.

The following sections IV-B, C, D, E, F details all the steps.

B. Descriptors' selection

In this step, we choose descriptors that reflect the color, shape and texture of objects. We extract descriptors and subsequently store them in the database.

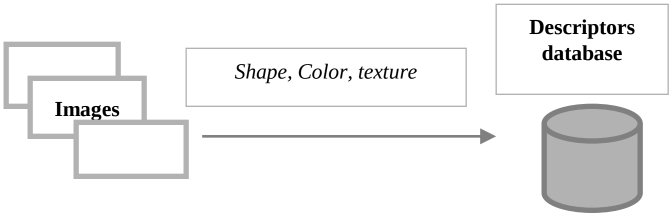


Fig. 4. Visual descriptor extraction

A Descriptor (feature) is a metric or any quantifiable value used to describe an image at a high-level perspective. Features related to color, texture, shapes, color blobs, corners are contained in an image. Color is a basic feature for image representation, and is invariant with respect to scaling, translation and rotation of an image [21]. It can be computed as Histograms (distribution of RGB, Hue,…), as SIFT descriptors or as moments of order p+q on RGB triplet or as a SIFT descriptor [22].

The haralick texture features are used for image classification. They capture information that emerges in patterns of texture. These kind of features is calculated by using co-occurrence matrix, which is computationally expensive [23].

In imaging applications, the shape of image objects provides a useful hint for similarity matching. For image retrieval the shape descriptor wants to be invariant to scaling, rotation and translation [24]. The descriptors are based on contour and region extraction.

In practice, we can combine multiple descriptors and choose those that are most significant (e.g. In some cases, we can neglect the texture if the objects have the same texture).

C. The Clustering method adopted

We use an unsupervised approach known as clustering. The main difference with the supervised classification is that the data set, from which we learn decision rules, does not include the information of an observation belonging to classes.

The problem of an automatic classification is to produce the labels {z1, ..., zn} of observations{x1, ..., xn} ∈ R*p* only on the knowledge of the values taken by the *p* variables. Unlike discriminant analysis, automatic classification does not have a learning phase to learn the classes' characteristics. An additional difficulty in automatic classification is that we do not necessarily know the number k of groups.

As in discriminant analysis, we divide clustering methods into two categories generative and discriminative methods. Generative clustering methods are quasi-exclusive, based on the mixture model and the EM estimation algorithm. Discriminative methods, in turn, all use a hierarchical classification structure.

We adopted a partitioning clustering method including K-means algorithm. By getting more partitions with K-means we produce a taxonomic tree.

D. Expert parameter and evaluation function

Our method of creating partitions does not aim to provide a hierarchical clustering algorithm, but to obtain a dynamic taxonomy. It is a structure for storing information in a memorable and operative way.

We construct the tree in an incremental way. At each level, we establish a partition of objects, meeting the criteria of accuracy and easy recognition. We need to have a measure to evaluate the partitions and retain the one with the mentioned criteria.

As the objects regrouping or separation depends on their looks or dissimilarity. We use the entropy measure [25] to establish partitions evaluation function. The process involves subsequently producing partitions and evaluating them. The difficulty in production methods is their efficiency to produce the most relevant partitions given the appearance NP-complexity.

It will add additional criteria to make combinations. Similarly, the user's interventions grow the system to introduce unnecessary hierarchy levels. Sometimes we get degraded situations where a node has only one son. There is therefore two simplification access to explore for a more refined structure.

We base the evaluation of a categorization on:

- The Cue-validity, which is the probability p(Cj | ai) of a category Cj having the descriptor ai.

- Category $p(ai \mid Cj)$ the probability of the existence of an attribute ai in a category Cj.

We have proposed in [1] a measure noted $Ep(\Gamma)$ for a partition Γ in the taxonomic tree. It is based on cue-validity, which expresses the non-uniformity of properties in categories. Minimizing $Ep(\Gamma)$ consist on building the highest possible homogeneous categories. This measure decreases every time we go down in the taxonomic tree.

On the other hand, recognition must be fast. It is therefore necessary to climb up the taxonomic tree before the matching process. Higher you go, the more the uncertainty of finding a category for an object decreases the top of the tree contains all objects. We have noted $Ec(\Gamma)$ the measure that expresses the uncertainty of categories.

For that we can say, that to allow the operation of recognition to succeed in the majority of cases on a quick and robust way, we must find a partition that achieves the minimum of $Ep(\Gamma)$ and $Ec(\Gamma)$.

Then we expressed this combination in a linear way:

$E(\Gamma) = \alpha.Ep(\Gamma) + \beta.Ec(\Gamma)$ (1) [1].

Ec and *Ep* plays antagonist roles. The tree root has only one category, the effort is focused on attributes, in this case, we keep only Ep. Otherwise, the level of the sheets (each object is one category), The effort is focused on the categories: we retain only Ec. For this we choose $\beta = 1-\alpha$. $0\leq\alpha\leq1$**,** and we have: $E(\Gamma) = \alpha.Ep(\Gamma) + (1-\alpha)Ec(\Gamma)$.

By adjusting the parameter α, we change the partition level in the taxonomic tree that minimize $E(\Gamma)$. In fact, this parameter expresses the degree of expertise in the object domain. The more knowledge we have, the closer we get to the leaves of the tree (α value approximates 1).

We propose many categorization procedures in an incremental way through minimizing the overall uncertainty $E(\Gamma)$ in each step. Our method idea is to find the optimal partition in the sense of minimizing the total uncertainty E, by progressive construction of the partition. We start from a small town of a few objects made necessary for the calculation of uncertainty. We proceed by adding objects to an existing agglomeration and decide whether to merge or divide them. The problem may arise at the dividing level of possible combinations based on the number of content objects. The category search of the object is done by minimizing the overall uncertainty $E(\Gamma)$.

The α variation permits the construction of several levels, including the root level ($\alpha = 0$) and the leaf level ($\alpha = 1$). We will then have a taxonomic tree.

To build taxonomic tree allowing us to generate the domain ontology, we must seek to have a unique connection between the upper and the lower partition, in the sense that a lower class has an inclusion link to a single partition larger and not 2 or more.

E. Taxonomic tree creation by hierarchical relations

1) Partitioning

An ontology should:

- Enable the reuse of a domain knowledge.
- Explain what is considered implicit in a domain underlying an implementation, which makes the modification of specifications possible in the case of the evolution of the domain knowledge.
- Distinguish domain knowledge of operational knowledge.
- Analyze knowledge in a domain since formal analysis of terms is extremely valuable as well when you want to use existing ontologies, as when we want to extend those.

There is not a single methodology to create an ontology. On this paper, we will focus on the automatic generation of domain ontologies hierarchically using a top-down method.

In a previous article, we proposed a new method of categorization of visual objects. The proposed function is based on entropy (as we have explained above). It allows us to create a taxonomic tree of objects from different partitions. That is to say by changing the value of a variable α we get to generate partitions with a number of classes. The gathering of these partitions creates a taxonomic tree with several levels, the highest level brings all classes while the lowest level is composed of several classes with a higher accuracy.

Choosing a partition is done by maximizing homogeneity and promoting the recognition process. We explicit our approach as follows, in the first step we categorize objects using a clustering algorithm. We use our evaluation function to create a taxonomic tree and this by changing the α value that varies in the range [0,1].

Getting close to 1 gives us more classes (clusters) thus better accuracy during the step of recognition but recognition time becomes expensive.

The principle of the procedure is as follows:

```
For $\alpha$=0 to 1 by step of st do
  For nbclasse= nbClassMin To nbClassMax by a step of 1 Do
      -Apply the clustering Algorithm
      -Evaluate the score produced by our evaluation function (cost)
  End For
      -Retain the partition minimizing function for $\alpha$
End For
```

Note that the classes with the minimum cost increases with α. The algorithm then divides the current partition (more classes) when α increases.

α Value increasing

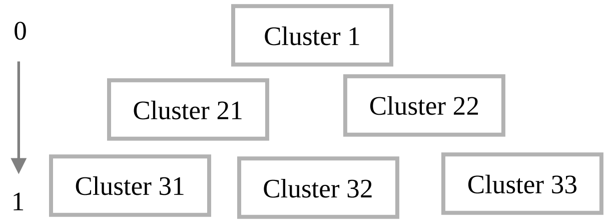


Fig. 5. Example of objects partitioning

2) Hierarchy process

As we mentioned α varies in the range [0,1], when it gets closer to 1 we get more classes per partition so we get a better accuracy.

The selected partition for each α value is the one with the minimum uncertainty E.

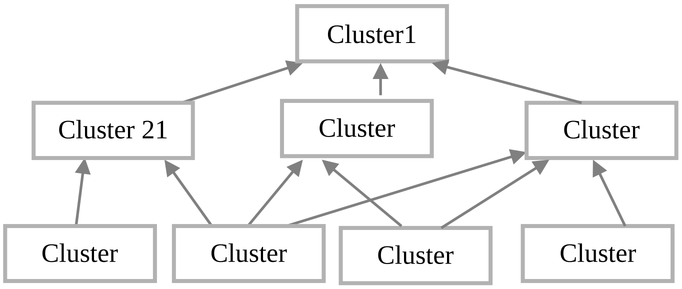


Fig. 6. Objects partitioning (decision step)

Cluster 1 is the first partition Cluster 21, 22 and 23 the second partition and the third partition is the cluster 31, 32, 33 and 34.

We can see that Cluster 32 objects belong to 3 clusters Cluster 21, Cluster 22 and Cluster 23.

The hierarchy property is an agglomeration of clusters in a low-level to form large clusters in a high-level. In order to decide which clusters should be combined (for agglomerative), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric and a linkage criterion, which specifies the dissimilarity of sets as a function of the pairwise distances of the observations in the sets.

By using *d* as the chosen metric, some commonly used linkage criteria between two sets of observations *A* and *B* are:

-Maximum or complete-linkage clustering:

$$\max\{\, d(a,b) : a \in A, b \in B \,\}$$

-Minimum or single-linkage clustering

$$\min\{\, d(a,b) : a\in A, b \in B \,\}$$

-Mean or average linkage clustering, or UPGMA (the Unweighted Pair Group Method with Arithmetic Mean).

$$\frac{1}{|A||B|}\sum_{a \in A}\sum_{b \in B} d(a,b)$$

-Centroid linkage clustering, or UPGMC (Unweighted centroid clustering)

-Centroid linkage clustering, or UPGMC (Unweighted centroid clustering)

$||C_s - C_t||$ where Cs and C_t are the centroids of clusters *s* and *t*, respectively

-Minimum energy clustering

$$\frac{2}{nm}\sum_{i,j=1}^{n,m}||ai - bj||2 - \frac{1}{n^2}\sum_{i,j=1}^{n}||ai - aj||2 - \frac{1}{m^2}\sum_{i,j=1}^{m}||bi - bj||2$$

Other linkage criteria include:

- The sum of all intra-cluster variance.
- The decrease in variance for the cluster being merged (Ward's criterion).
- The probability that candidate clusters spawn from the same distribution function (V-linkage).
- The product of in-degree and out-degree on a k-nearest-neighbor graph (graph degree linkage).
- The increment of some cluster descriptor (i.e., a quantity defined for measuring the quality of a cluster) after merging two clusters.

Assume the chosen link is ξ , the hierarchical relation can be obtained by:

$H^C(C_j, C)$ with $C = \arg(\min_{Ci} \xi(Cj, Ci))$, C_i in high level

For example, in figure 6 the objects of cluster32 are distributed over cluster21, cluster22 and cluster23. One way to decide for the hierarchy is to consider the UPGMC (Centroid linkage clustering) link, then we obtain H^C(cluster32, C) where C realizing the minimum of ($||C_{32} - C_{21}||$, $||C_{32} - C_{22}||$ and $||C_{32} - C_{23}||$) with C_{ij} is the centroid of the corresponding cluster.

F. Towards concrete ontology

1) Automatic labeling

After creating the tree, each cluster gets automatically a label produced according to the level class and the position of the class at that level (according to a particular agreement). We define the labels L_{ij} $1<i<N$ with N number of the level and $1<j<n_i$ with n_i class number in the level i, classes are the concepts of ontology.

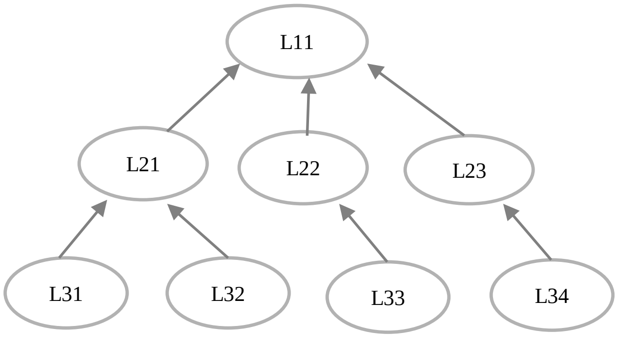
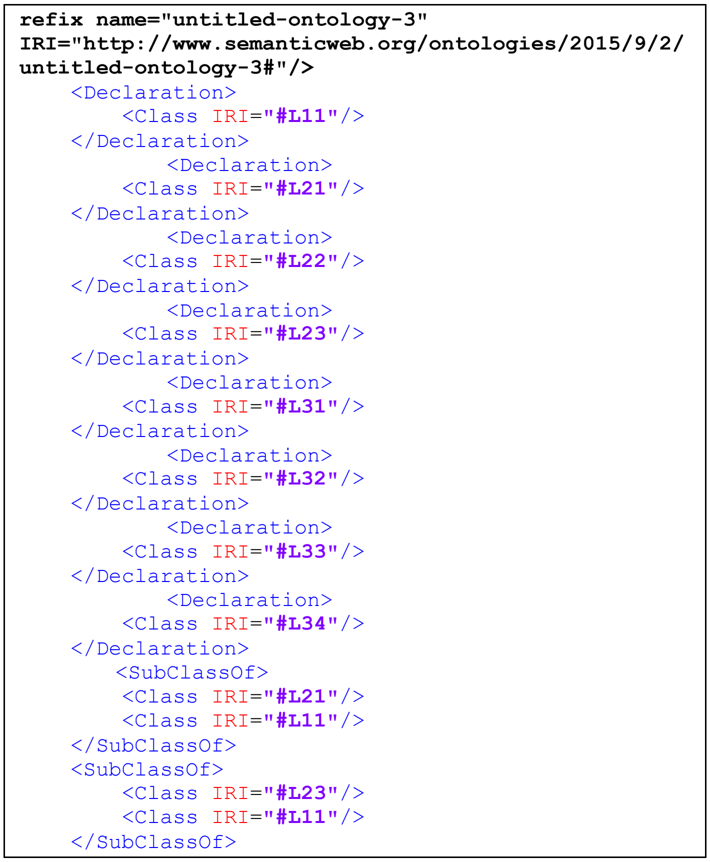


Fig. 7. Attributing labels to the ontology concepts

2) *Automatic generation of OWL code*

Then we will write our ontology in owl language and generate the graphic tree later using an ontology editor, respecting automatic labeling Ln.



```
refix name="untitled-ontology-3"
IRI="http://www.semanticweb.org/ontologies/2015/9/2/
untitled-ontology-3#"/>
    <Declaration>
        <Class IRI="#L11"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L21"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L22"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L23"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L31"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L32"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L33"/>
    </Declaration>
            <Declaration>
        <Class IRI="#L34"/>
    </Declaration>
        <SubClassOf>
        <Class IRI="#L21"/>
        <Class IRI="#L11"/>
    </SubClassOf>
    <SubClassOf>
        <Class IRI="#L23"/>
        <Class IRI="#L11"/>
    </SubClassOf>
```

Fig. 8. Writing ontology in OWL (Part of the code)

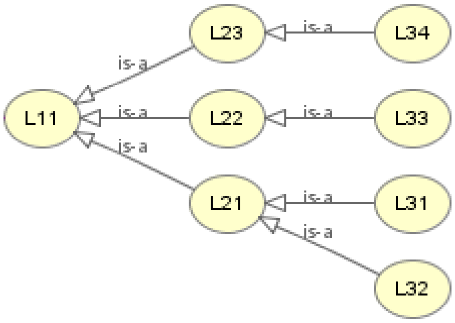
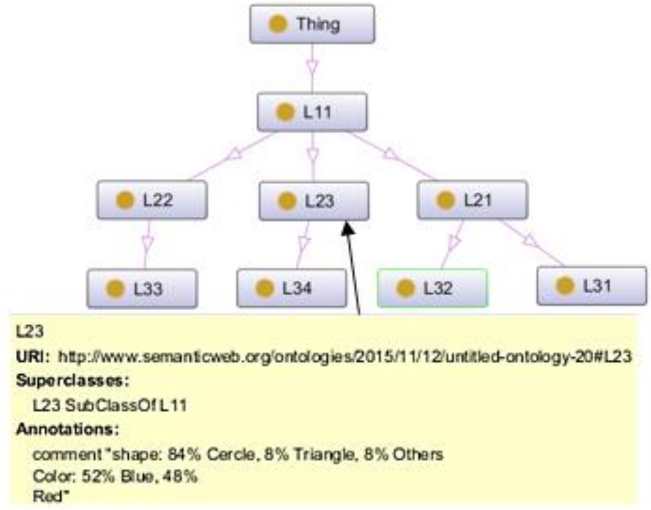


Fig. 9. Reading OWL file on Protégé and the generation of the tree

3) *The expert intervention*



L23
URI: http://www.semanticweb.org/ontologies/2015/11/12/untitled-ontology-20#L23
Superclasses:
L23 SubClassOf L11
Annotations:
comment "shape: 84% Cercle, 8% Triangle, 8% Others
Color: 52% Blue, 48%
Red"

Fig. 10. Generating the ontology graph from the OWL file with annotations

Now the domain expert can intervene to add other concepts to the ontology or to place new objects in their home concept by comparing their features with concept prototypes. From the taxonomic tree tagged with a detail on the homogeneity (shape, color, texture) of each class, a domain expert can intervene in this step to match the labels generated automatically by a lexicon belonging to the domain.

In the domain of the road scene, for example, and in particular for road signs, the expert can apply a decision tree similar to the figure:

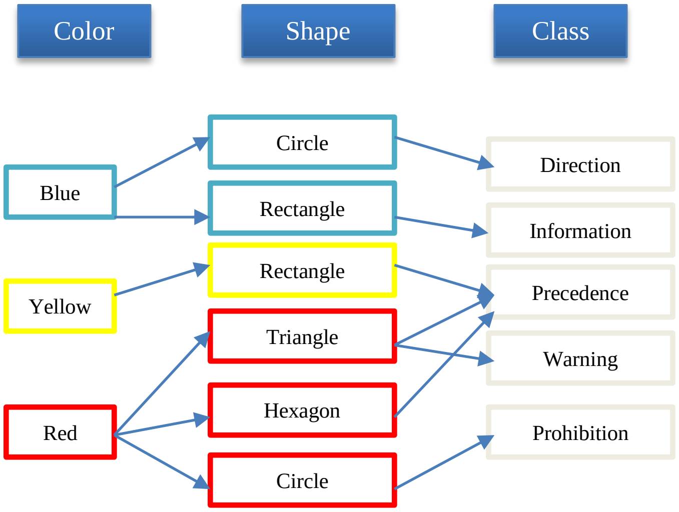


Fig. 11. Decision tree example for traffic signs

The expert can also adjust the tree with his domain knowledge (road scenes). The domain expert can intervene or not to keep the homogeneity of classes (concepts) and keep only the relevant tree.

V. Application Example

Goal: Creation of an ontology of hierarchy of road sign panels by implementing the different steps of our approach.

We considered 81 road signs images. The first step is to extract visual features (as shown above). We used color, shape and texture descriptors using 12 features.

TABLE I. Texture and Shape Descriptors

Descriptors	Definition
Entropy (F1)	Statistical measure of randomness. It characterizes the degree of organization
Contrast (F2)	Measures the local variation of the co-occurrence matrix of gray level
Correlation (F3)	Measures the joint probability of occurrence of specified pairs of pixels
Energy (F4)	Provides the squares sum of GLCM elements. Also known as uniformity or the angular second moment
Homogeneity (F5)	Measures the closeness of element distribution in the GLCM to the GLCM diagonal
Area (F6)	The number of pixels in the region
Perimeter (F7)	The perimeter of the distance between each adjacent pair of pixels around the region borders.
Convex area (F8)	The scalar indicating the number of pixels in Convex Image. This property is supported only for 2-D for input label matrices
Euler number (F9)	The scalar indicating the number of objects in the region, less the number of holes in objects.
Extent (F10)	The scalar that specifies the number of pixels in the pixel region of the total area delimitation. Calculated as divided by the area of the bounding box area. This property is supported only for 2-D input label matrices.

For color features we use the moments of second-degree features, calculating the energy derivatives, according to star and circle structure F11 F12.

TABLE II. A View of the Database of the 12 Features Used [1]

	F1	F2	F3	F4	F5	F6
Image 1	+3.154	+1.875	+0.398	+0.839	+0.879	+2148.000
Image 2	+3.275	+2.152	+0.344	+0.840	+0.866	+2157.000
Image 3	+2.910	+1.394	+0.446	+0.867	+0.899	+2166.000
Image 4	+3.367	+1.941	+0.337	+0.853	+0.864	+3250.000
Image 5	+3.163	+1.092	+0.397	+0.861	+0.895	+4.000
	F7	F8	F9	F10	F11	F12
Image 1	+403.647	+5104.000	-14.000	+0.244	+0.013	+0.005
Image 2	+313.421	+5109.000	-19.000	+0.245	+0.020	+0.006
Image 3	+313.421	+5110.000	-22.000	+0.246	+0.012	+0.006
Image 4	+313.421	+5104.000	-21.000	+0.369	+0.018	+0.009
Image 5	+6.000	+4.000	+1.000	+1.000	+0.010	+0.003

After extracting the features we classify the object using a clustering algorithm (kmeans) and use our evaluation function to keep the best partition for each iteration, we remind that the function is based on the entropy and by adjusting the α value we get different partitions (number of classes of each partition). In this application, we get a set of partition by varying α between 0,1 and 0,2 which gave us partitions with 2 to 8 classes as seen in table 3.

TABLE III. Partitioning Results Combining Color, Shape and Texture Descriptors Using Kmeans

	Color/Shape/Texture				
α value	0,1	0,12	0,15	0,17	0,2
Classes (number)	2	3	4	6	8

We consider this result and we create a hierarchy between those classes based on the centroid clustering or as we called it UPGMC (The Unweighted centroid clustering). UPGMC joins the objects or groups that have the highest similarity (or the closest distance), by replacing all the objects of the group produced by the centroid of the group. This centroid is considered as a single object at the next clustering step.

UPGMC, as well as WPGMC, can sometimes produce reversals in the dendrogram. This situation occurred in our example. This happens when:

- Two objects about to join (let us call them A and B) are closer to one another than each of them is with a third object C: $\|AB\|<\|BC\|$
- After the fusion of A and B, the centroid of the new group A-B is closer to C than A was to B before the fusion: $\|C_{AB}C\|<\|AB\|$

To make the decision of assigning object to the closest centroid, we need to measure the proximity. We followed the steps of proximity measurement explained in the hierarchy process section.

This gave us the following result (figure 12). For the first partition with two classes, the first one has mostly the prohibition and indication panels (with round shape), in the second class we get essentially danger panels (with triangular shape). Coming down on the tree we have partitions with more classes so we get more accuracy. In the last level of the 4 classes, we have the correct categorization. A class for prohibition panels, a class for intersection panels, a class for danger panels and a class for indication panels.

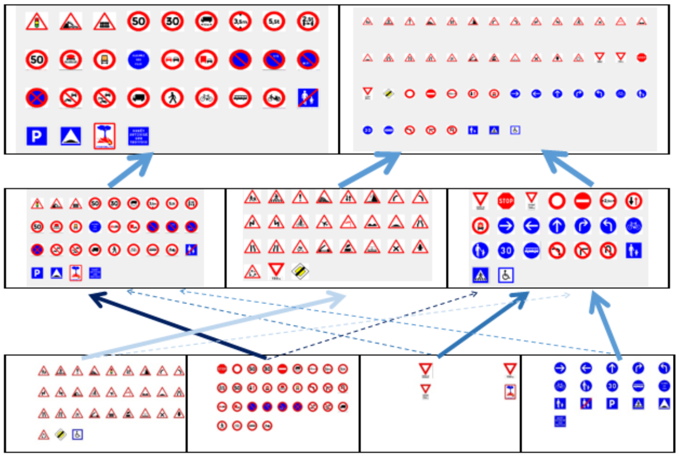


Fig. 12. Partitioning levels (Thick lines describe the final decision for the relation H^c with centroids distance, Thin lines describe short distances on centroids)

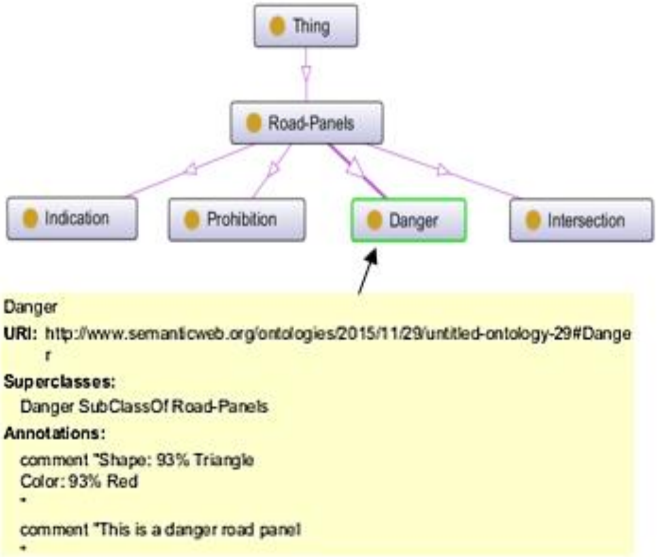
After having the classification results and the taxonomic tree now we generate our ontology of hierarchy by creating links between concepts. After that we write our OWL ontology, this file will be read later on an ontology editor where the expert can intervene.

```
<Declaration>
        <Class IRI="#Danger"/>
       <owl:AnnotationProperty
rdf:about="&dc;creator"/>
<owl:Class rdf:about="#Danger">
  <rdfs:label>this is the Class of Danger Road
Panels: Shape:93% Triangle, Color: 93%
Red</rdfs:label>
</owl:Class>
```

```
    </Declaration>
<Declaration>
        <Class IRI="#Indication"/>
    </Declaration>
    <Declaration>
        <Class IRI="#Prohibition"/>
    </Declaration>
<Declaration>
        <Class IRI="#Intersection"/>
    </Declaration>
<SubClassOf>
        <Class IRI="#Danger"/>
        <Class IRI="#Road-Panels"/>
    </SubClassOf>
<SubClassOf>
        <Class IRI="#Indication"/>
        <Class IRI="#Road-Panels"/>
    </SubClassOf>
    <SubClassOf>
        <Class IRI="#Prohibition"/>
        <Class IRI="#Road-Panels"/>
    </SubClassOf>
<SubClassOf>
        <Class IRI="#Intersection"/>
        <Class IRI="#Road-Panels"/>
    </SubClassOf>
```

Fig. 13. Writing the ontology in OWL (example panels) (part of the ontology)

This figure replaced the Lexicon by clusters' names to give them a meaning, but as mentioned above we give to each class a lexicon as Lij.



Danger
URI: http://www.semanticweb.org/ontologies/2015/11/29/untitled-ontology-29#Dange
r
Superclasses:
Danger SubClassOf Road-Panels
Annotations:
comment "Shape: 93% Triangle
Color: 93% Red
"
comment "This is a danger road panel
"

Fig. 14. Reading the OWL file in Protege (Danger Class with the annotations)

VI. Conclusion and Perspective

This paper presented an approach to object recognition and domain ontology creation from taxonomic tree. Our approach takes place in a set of steps, a knowledge acquisition step, where we describe a set of classes with visual concepts provided by a visual ontology concept.

We proposed in this article our current vision of domain ontologies we start from the image to identify the concepts and relations between concepts that emerge. In particular, we have shown that a methodology based on producing a consensus clustering (thus unsupervised) highlights semantically relevant concepts and relations.

We plan to develop this ontology as a basic vocabulary in a multi-agent system for road scene interpretation.

REFERENCES

[1] Mansouri F, Sadgal M, Elfazziki A, Benchikhi L, An evaluation of perceptual classification led by cognitive models in traffic scenes, Electronic Letters on Computer Vision and Image Analysis, vol 14(1) pp 38-60 2015

[2] Nicola Guarino, Formal ontology, conceptual analysis and knowledge representation, International Journal of Human-Computer Studies, Volume 43, Issues 5–6, November 1995, pp 625–640

[3] J F Sowa, Ontology, Metadata, and Semiotics, Chapter Conceptual Structures: Logical, Linguistic, and Computational Issues,Volume 1867 of the series Lecture Notes in Computer Science pp 55-81

[4] Gruber T.R. 1995. Toward principles for the design of ontologies used for knowledge sharing.Int J Hum, Comput Stud, vol 43, 1995, pp 907-928.

[5] Hassan S, Hétroy F, Palombi O, segmentation de maillage guidée par une ontologie, 22 ème journée de l'association francophone de d'informatique graphique, Arles, 2009

[6] Donald HindleAT&T Bell Laboratories, Murray Hill, NJ Noun classification from predicate-argument structures, Proceeding ACL '90 Proceedings of the 28th annual meeting on Association for Computational Linguistics, pp 268-275

[7] Marti A. HearstUniversity of California, Berkeley, CA, Automatic acquisition of hyponyms from large text corpora Proceeding COLING '92 Proceedings of the 14th conference on Computational linguistics - Vol 2, pp : 539-545

[8] Peter Spyns, Marie-Laure Reinberger, Lexically Evaluating Ontology Triples Generated Automatically from Texts,Chapter The Semantic Web: Research and Applications, Volume 3532 of the series Lecture Notes in Computer Science pp: 563-577

[9] Philipp Cimiano, Andreas Hotho, Steffen Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, Journal of Artificial Intelligence Research 24 (2005) 305–339.

[10] zhang G, Satya S, Samden D, From classification to epilepsy ontology and informatics, Epilepsia, Vol 53, Issue Supplement s2, pp: 28–32, July 2012

[11] Kompatsiaris Y., Hobson P. (ed.) 2008. Semantic Multimedia and Ontologies, vol. 1, Theory and applications springer

[12] Neumann B., Möller R., (2008). On scene interpretation with description logics. Image vision comput, vol 26, january, pp 82-101

[13] Clouard R., Renouf A., Marinette R. (2010). An Ontology-Based Model for representing image processing application objectives. International Journal of Pattern Recognition and Artificial Intelligence, vol 24, pp 1181-1208

[14] R. Bodner and F. Song, “Knowledge-based Approaches to Query Expansion in Information Retrieval,” in Proc. of Advances in Artificial Intelligence, pp. 146-158, New York, Springer.

[15] Dave Elliman, J. Rafael G. Pulido. “Automatic Derivation of On-line Document Ontology”, International Workshop on Mechanisms for Enterprise Integration: From Objects to Ontology (MERIT 2001) 15th European Conference on Object Oriented Programming, Budapest, Hungary, Jun 2001.

[16] A. Hotho, A. Mädche, A., S. Staab, “Ontology-based Text Clustering,” Workshop Text Learning: Beyond Supervision, 2001.

[17] Khan L, Wang L. Automatic ontology derivation using clustering for image classification. In: Eighth International Workshop on Multimedia Information Systems, Tempe, Arizona, November 2002

[18] MIKE USCHOLD, Knowledge level modelling: concepts and terminology The Knowledge Engineering Review The Knowledge Engineering Review / Volume 13 / Issue 01 / March 1998, pp 5-29

[19] Steffen Staab, Text Clustering Based on Good Aggregations Andreas Hotho, Alexander Maedche, Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA

[20] D. D. Dhobale, B. S. Patil, S. B. Patil, V. R. Ghorpade, Semantic understanding of Image content, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011,pp 191-195

[21] Automatic Image Annotation Using Group Sparsity. Shaoting Zhang, Junzhou Huang, Yuchi Huang, Yang Yu, Hongsheng Li, Dimitris N. Metaxas, Department of Computer Science, Rutgers,University, Piscataway, NJ, 08854 Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA

[22] Rekhil M Kumar, A Survey on Image Feature Descriptors, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7668-7673

[23] B. S. Manjunath, Member, IEEE, Jens-Rainer Ohm, Member, IEEE, Vinod V. Vasudevan, Member, IEEE, and Akio Yamada, Color and Texture Descriptors, ieee transactions on circuits and systems for video technology, vol. 11, no. 6, june 2001

[24] MPEG-7 Visual Shape Descriptors Miroslaw Bober ,ieee transactions on circuits and systems for video technology, vol. 11, no. 6, june 2001

[25] Shannon and Weaver. The mathematical theory of communication, 1949.

The Discovery of the Implemented Software Engineering Process Using Process Mining Techniques

Zayed, Mostafa Adel
Information Systems Department,
Faculty of Computers and Information,
Helwan University, Egypt

Ahmed Bahaa Farid
Information Systems Department,
Faculty of Computers and Information,
Helwan University, Egypt

***Abstract*—*Process* model guidance is an important feature by which the software process is orchestrated. Without complying with this guidance, the production lifecycle deviates from producing a reliable software with high-quality standards. Usually, teams break the process deliberately or impulsively. Application Lifecycle Management (ALM) tools log what teams do even if they break the process. The log file could be a key to discover the behavior of the undertaken process against the targeted process model. Since the date of its introduction, Process Mining techniques have been used in business process domains with no focus on the software engineering processes. This research brings the Process Mining techniques to the software engineering domain. The research shows a conclusive effort that used a Scrum adapted process model as an example of Agile adoption. This research has applied Process Mining discovery techniques to capture the actually implemented process by the Scrum team. This application clarifies the gap between the standard process guidance and the actually implemented one. The research's results showed that Process Mining techniques have the ability to discover and verify the deviation on both levels; the process itself as well as the work items state-machine workflows.**

Keywords—Process Mining; Process Models Discovery; Software Engineering; Agile; and Scrum

I. Introduction

Software engineering process has become an integral part of any software production lifecycle definition. Without a model that governs the state-machine of the process phases, software engineering process will not be clearly defined. Software factories choose one of the available process models according to a set of aspects e.g. product team preparation, customer type, project development period, etc. Correspondingly, contemporary ALM tools have been developed to provide sort of automation, collaboration, and process model guidance.

In late 1970s, computer scientists focused on efficient data retrieval and storing. Nevertheless, this was inadequate, because any business is not only composed of data but also a process. A good start was in the nineties, where processes' representation, monitoring, and visualization drew computer scientists' attention as an essential block for extending the capabilities of information systems to enforce business processes. This type of information systems is denoted by Business Process Management (BPM) systems [1]. Business process management systems examples are Staffware, MQSeries, Microsoft Work Flow, COSA, and FLOWer [2, 3, 4, 5]. Despite BPM systems do straighten many deviations, deviations do occur. Therefore, a methodological solution is needed to identify these deviations after they happened and logged, then analyze these deviations. Since information system records all that happened, an event log can be prepared –if it was not originally prepared, and use it to identify the actual process that the event log depicts [6]. This identification phase is denoted by Discovery. In addition, discovery can lead to process analysis. To check what rules are applied and what is neglected or overwritten that is denoted by Conformance Checking [7]. In order to check process conformance and analyze it, a discovery phase has to be applied to the recorded event log. Using both discovery and conformance checking the current process can be enhanced, considering the deviations are exceptions or unknown cases that are not handled by the process model. The result of enhancement is repairing or extending the current applied process model [8]. Discovery, Conformance Checking, and Enhancement are the Process Mining types. Fig. 1 depicts the need and positions the Process Mining from organizational process models' interaction with current systems, and rectification of the current process models.

II. Literature on Process Mining

A. Review of Process Mining algorithms and techniques

In [9, 10] α-algorithm and α^{+}-algorithm discuss the most common discovery algorithm with test datasets for testing the capabilities of the algorithm at the edge. Then, in [11] a HeuristicsMiner algorithm is introduced to discover the process and identify the noise in the process. The used dataset was hospital information system of patients. Then, in [12, 13] a Fuzzy Mining algorithm was built upon the HeuristicsMiner algorithm introduced in [11]. This algorithm is built to discover the unstructured logged processes. The datasets used were machinery test and usage logs, development process logs, hospital patient treatment logs, and logs of case handling systems and web servers. Then, a more mature framework was introduced in [14], which introduced a sound and complete solution for discovering structured processes that has compliance with the four quality dimensions introduced in [15].

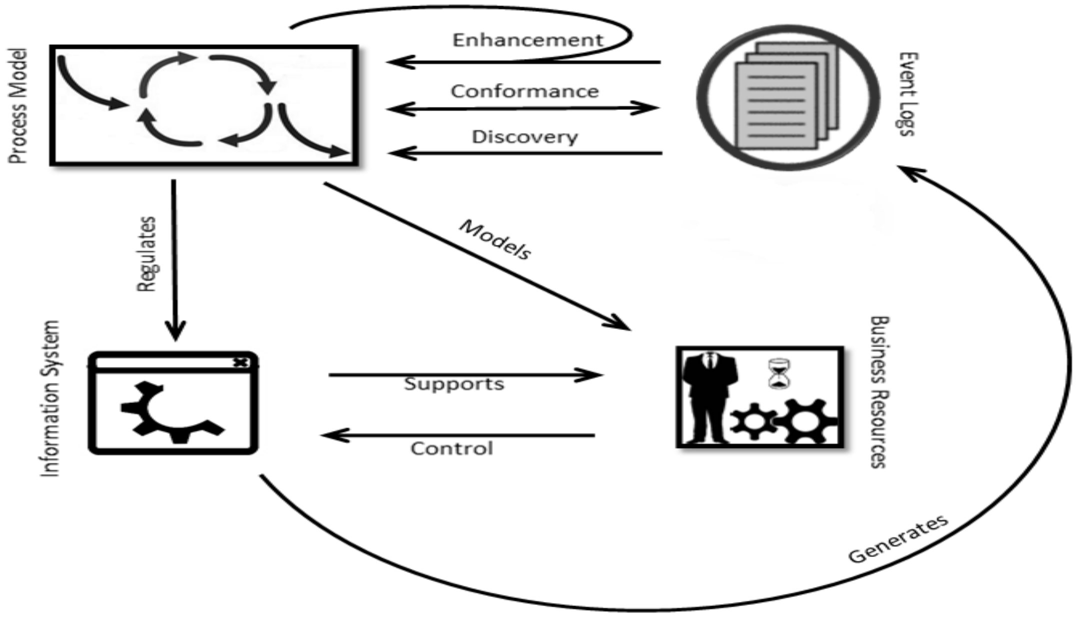


Fig. 1. Process Mining occurrence in organizational process models

In [14] the framework was applied on CoSeLoG project that used municipalities' data in Netherlands. These quality dimensions were studied to show how crucial are they and how important to implement them in [16].The evaluation criteria depends on the four quality dimensions. These quality dimensions are 1: Fitness, 2: Precision, 3: Generalization, and 4: Simplicity. Fitness is represented by this question, is the observed behavior captured by the model? I.e. How many cases can be replayed from the event log over the extracted model? Precision is represented by this question: How precise the process model describes the observed behavior? The less behavior the process model allows that is not observed in the event log, the more precise the process model describes the behavior. Generalization is represented by this question: Does the model allow for more behavior than encountered in reality? Simplicity is represented by this question: How simple, or human-readable, is a process model? [17] A comparison shows the differences between discovery algorithms is shown in TABLE 1 [14]. The tilde (~) sign in the table represents algorithm finish with remaining work. A sound (correct) process models can be produced with ETM*d* algorithm as shown in the last row in TABLE I. . ETM*d* algorithm is a part of ETM (Evolutionary Tree Miner) framework, which is built using an evolutionary algorithm, genetic programming specifically. The best part of evolutionary algorithms is that it finds a solution as long as you have the time to wait for its solution, unless you defined an early exit criterion. In general, Process Mining has a set of characteristics, guiding principles and challenges [18]. ETM framework has covered seven challenges out of eleven challenges mentioned in the manifesto. A full list of challenges that ETM framework has overcome can be read from here [14]. An important challenge is “Improving the Representational Bias Used for Process Discovery”, as mentioned by the process mining manifesto, or “Separation of Visualization and Representational Bias”, as mentioned by [14]. ETM framework uses Process Trees to overcome this challenge. In order to have a deeper sense of this challenge, it should be considered that process models have standard notations like Petri nets, business process management notation (BPMN) [19], and event-driven process chain (EPC) [20]. The discovery algorithms do not mostly use these standards –as mostly each uses its own notation, which makes it hard to business users to read and use such algorithms, which make their own notation. Examples of such algorithms are fuzzy models [13, 12], casual nets [21], and heuristics nets [11, 22]. The crucial point of ETM framework considering the representation challenge is the usage of process tree as it is mentioned before. Process trees notation by itself is not considered a common notation to business people, but process tree notation can be easily converted to a plethora of other common processes modeling notations like Petri nets, EPC, YAWL, casual net, heuristic net, fuzzy model and process algebra. This paper is going to use ETMd as the research's discovery algorithm; in addition, this paper will show the discovered process model using the α-algorithm. This research is going to use software engineering data, based on Scrum process specifically. In fact, up to the moment, software engineering domain is considered a virgin domain for applying Process Mining techniques. The research's results will be visualized using three different forms: process tress,

Petri nets, and BPMN as the OMG standard notation.

The rest of this section discusses the research motivation in subsection C. The process tree notation is showed in subsection D. In addition, the dataset collection and preparation that the research was based on is discussed in section II. Moreover, in section II, this literature has a subsection discussing the Scrum process definition template. Discovery of the both process levels is discussed in section III. The results are compared to the process definition template in section IV. Finally yet importantly, a quick summary is showed in section V.

TABLE I. PROCESS DISCOVERY ALGORITHMS COMPARISON FROM [14]

Algorithm	**Error-Free?**	**Replay Fitness**	**Precision**	**Generalization**	**Simplicity**
α-algorithm	Yes	~	Yes	No	No
Genetic miner	~	~	~	No	No
Heuristics miner	~	No	Yes	No	No
ILP miner	~	Yes	No	Yes	No
Inductive miner	Yes	Yes	No	Yes	Yes
Language-based region theory	Yes	No	No	No	No
Multi-phase miner	~	Yes	No	No	No
State-based region theory	Yes	Yes	No	No	No
ETM*d*	Yes	Yes	Yes	Yes	Yes

B. *Process Tree Notation*

Process trees have six different operators: sequence (→), the reversed sequence (←), exclusive-choice (×), parallelism (∧), non-exclusive choice (∨) and the loop (↻) [11]. Sequence operator (→) forces the mentioned sequence of child nodes from left to right and vice versa considering the reversed sequence operator (←). Exclusive-choice operator (×) of child nodes a, b, and c results in one of the following combinational sequence <a>, <b>, or <c>.

Parallelism operator (∧) of child nodes a, b, and c results in the set of {a, b, c} in all possible orders, i.e. the result can be one of the following combinational sequences <a, b, c>, <a, c, b>, <b, a, c>, <b, c, a>, <c, a, b>, or <c, b, a>. Nonexclusive choice operator (∨) of child nodes a, and b results in one of the following combinational sequences <a>, <b>, <a, b>, or <b, a>. On one hand, any of the previous operators could have any number of nodes starting from two nodes – of course; one child node has no meaning since it should be cloned to its parent. On the other hand the loop operator (↻) has, at least, three child nodes. Loop operator (↻) of –specifically ordered, child nodes a, b, and c results in one the following combinational sequences <a, c>, <a, b, c>, <a, b, a, c>, <a, b, b, b..., c>.

III. RESEARCH MOTIVATION

No previous research effort tried to use the aforementioned Process Mining techniques and algorithms in order to be used in the software engineering domain. According to Forrester's State of Agile 2015 report, 40% of the developed software is doing wrong things due to deviations from applying the lifecycle process [23]. In 59% of cases, the major impediment that prevents a correct Agile process adoption is the lack of skilled people [23]. If there were an intelligent way to discover these deviations in the applied lifecycle process, this would give the software factories the opportunity to bring their development back to the track. Thus, decreasing the possibility of producing the incorrect thing. This research utilizes Process Mining techniques to discover the deviation of the actually implemented software engineering process. Since, 86% of the Agile, teams are using Scrum [23]; this research was applied on a project log file of a team that is supposed to be working using the Microsoft Scrum process template definition.

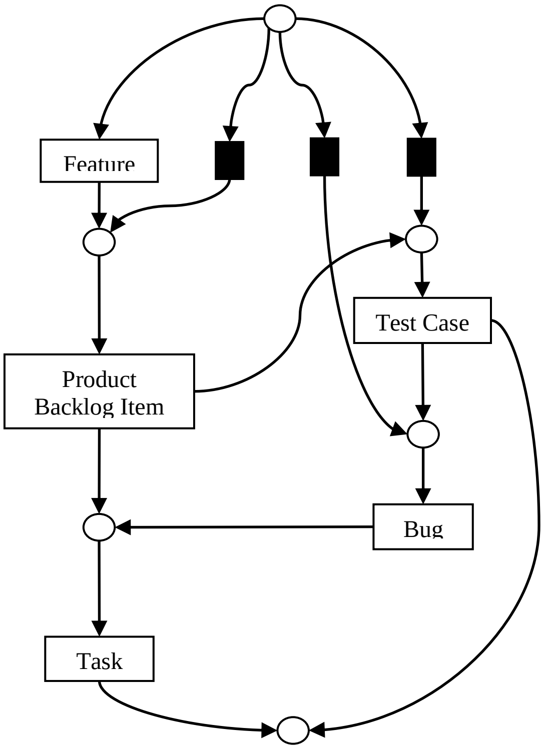


Fig. 2. Scrum Process Work Items Workflow in Petri Net

IV. RESEARCHED DATASET ACQUISITION

Scrum process categorizes all work in the project life cycle into six categories applied by the Microsoft Visual Studio Scrum 3.0. These six categories are feature, product backlog item (PBI), task, test case, bug, and impediment. Each one of them is called work item (WI) and has its own workflow states' model. Altogether, they formulates the software engineering Scrum process. According to the held data, this study is going to concentrate on subset of these WIs, which are feature, product backlog item, bug, and task. The Scrum process as defined by Microsoft is shown in Figure 2. The workflow model of each one of them is depicted in Fig. 3 [24]. The CASE tool used to capture the team's data and apply the desired process is Microsoft Team Foundation Server (TFS) 2013.

TFS data resides in Microsoft SQL Server. In order to discover the generated the event log, this study used ProM tool for generating the event logs, and discovering process off the data [25]. This research used SQL queries to extract the required process attributes and data, and then it used XESame to generate event logs in order to able to discover the process using discovery algorithms implemented in ProM. XESame is an integrated software in ProM responsible for extracting event it uses JDBC. XESame generates the event logs in defined formats with extensions (.XES), and the event log should be composed of events, cases, traces, and attributes.

For example, if you have two product backlog items, so, each one of them is called a case. Each case is composed of a set of chronologically ordered events using timestamp, called a trace. Each PBI has a team member changes its state. Both timestamp and the team member are called attributes. From TABLE II. and TABLE III. you can see how the event logs are composed. For the extraction phase, in order to discover the workflow of each work item this research defined the work item type and work item ID as the case, and this research defined work item state as the event transition life cycle. Data is filtered for the WIs workflow that is a single process too, which composes of 1482 cases, 12 event types, and event's frequency of 8609. In addition, to discover the life cycle of the scrum process, iteration ID is selected as the case, and work item types is selected as the event transition life cycle. Data is filtered for the lifecycle of the scrum process that is a single process, which composes of 500 cases, 5 event types, and event's frequency of 2496. After this step, it is capable of extracting the two XES files to discover each of their process models.

A. Scrum Process WI's Petri Net Discussion

In Figure 2 depicts the work item's (WI) process. The black transitions called siltent transitions. This literature used this transitions to be able to depict PBI WI, Bug WI, and Test Case WI can come as the first transition. In addition, PBI WI can come after Feature WI. Test Case WI can be the child of the PBI WI as well as the Task WI. Task WI's parent can be also a Bug WI. Bug WI's parent can be Test Case WI. Any case happens other than the mentioned cases is a clear devision from the definition template. An example of the deviation is a Task WI is a child of Feature WI.

V. Discovery of the Actually Applied Process Model

The most common and well-known discovery algorithm is α-algorithm [9]. Using 'Mine for a Petri Net using Alpha-algorithm' plug-in from ProM, which implements the α-algorithm and applying it over the extracted XES file that contains the data of different work items workflows, and visualize the discovered in petri net. Fig. 4 depicts the discovered workflow of different work items. Fig. 4 consolidates all work items workflows in single petri net.

Another way to discover this event log is using 'Mine a Pareto Front with ETMd' plug-in from ProM. Pareto Front is multi-objective optimization technique that offers a set of the best-optimized solutions. The trade-off between objectives in this study is the four quality dimensions. The plug-in offers a set of process models of Pareto fitness equals to 0.991326, which means that most of the offered solutions are feasible. Pareto fitness is the average fitness of the whole Pareto Front.

For the research's experiment, it has only two models of fitness equals to one. In this paper, only one model is presented –out of 193 models, of replay-fitness equals to one. This plug-in has the power of visualizing its discovered process model as BPMN as default as depicted in Fig. 5. In addition, the plug-in provides the process tree string, and the string is depicted in Fig. 6.

TABLE II. DETAILED EVENT LOG

WI Type	State	Timestamp	Changed By
PBI (1)	New (N)	Day (1)	Team Member A
PBI (1)	Approved (A)	Day (2)	Team Member A
PBI (2)	New (N)	Day (2)	Team Member A
PBI (2)	Approved (A)	Day (3)	Team Member B
PBI (1)	Committed (C)	Day (3)	Team Member B
PBI (2)	New (C)	Day (4)	Team Member A

TABLE III. TWO-WAYS SUMMARIZED EVENT LOG

Case	Trace of Events
PBI (1)	N, A, C
PBI (2)	N, A, N

Trace	Occurrence Count
N, A, C	1
N, A, N	1

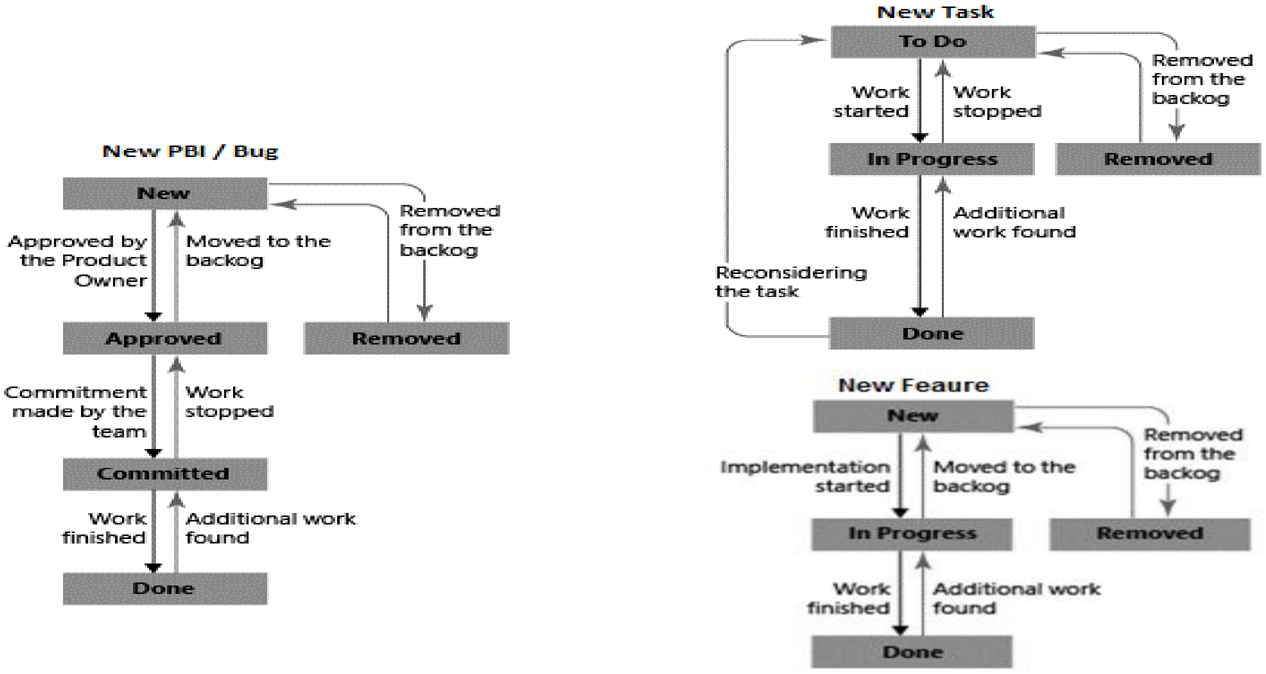


Fig. 3. Scrum Work Items Workflow [24]

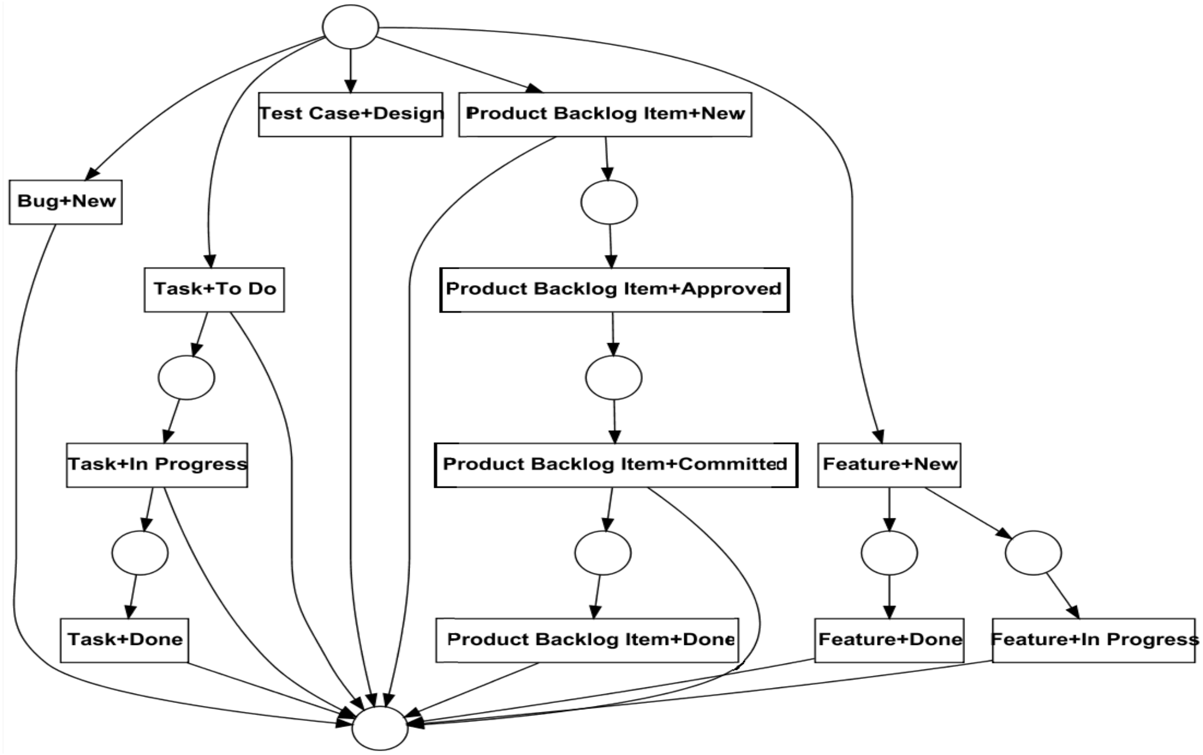


Fig. 4. Discovered Petri Net for Scrum WIs workflows

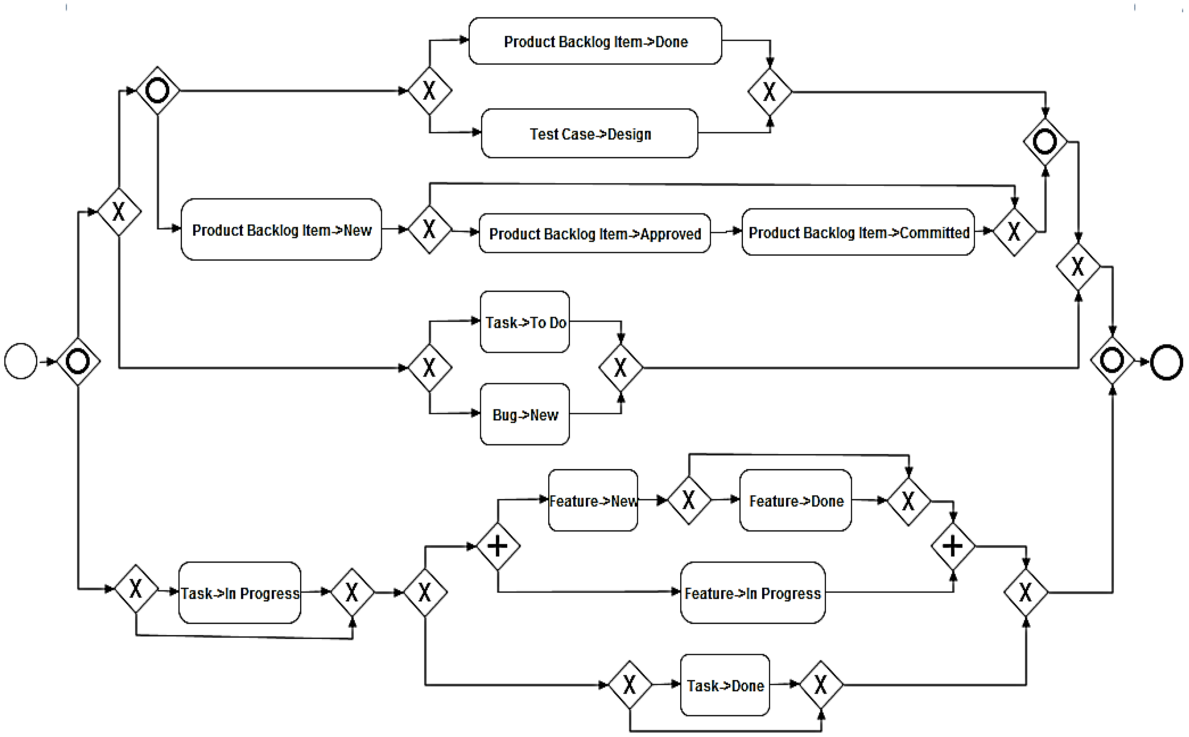

Product Backlog Item->Done
Test Case->Design
Product Backlog Item->New
Product Backlog Item->Approved
Product Backlog Item->Committed
Task->To Do
Bug->New
Feature->New
Feature->Done
Task->In Progress
Feature->In Progress
Task->Done

Fig. 5. Discovered Pareto Front as BPMN for Scrum WIs workflows

VI. Results Discussion

The discovered Pareto Front model's quality dimensions values are 1: Fitness = 1.0, 2: Precision = 0.771, 3: Generalization = 0.579, and 4: Simplicity = 0.906. These quality dimensions' values reflect the logged events are all applicable by the discovered model that is a positive aspect, the discovered process model describes the log well but it can allow much more behavior than recorded that is a down side, and the discovered process model is readable. Comparing the discovered model with Fig. 3 a huge gap away of the standard process model definition is still noticed. This can be deduced, because the scrum team is not adhering to the process definition as it is obviously discovered from the two algorithms. On one hand, the discovered PBI WI, Task WI, and Feature WI states are highly correlate to the process definition in Figure 3. On the other hand, it is found that the discovered Case WI and Bug WI states are not correlating with Figure 3 at all.

After discovering a detailed view of the WIs workflows, it is needed to know if conformance to the Scrum process as whole occurs. This study used α-algorithm [9] by applying 'Mine for a Petri Net using Alpha-algorithm' plug-in from ProM. The output result is visualized in the petri net of Fig. 7. Obviously, the discovered model is away from conformance to the Scrum process model depicted in Fig. 2. This nonconformance may be mainly due to the same reason mentioned for the WIs workflow that is the team overlooks the defined process is still applicable for noncomplying the Scrum process. The overlook here can reflect the team's lack of knowledge by the purpose of work item types and the workflow that organizes them. Deviations can be categorized into two categories. The first is fatal deviation, the second one is some events may still in progress so you can see it in the discovered model as deviated, this can be called slight deviations. Considering the fatal deviations, results showed in the discovered model that the PBI WI's parent is Bug WI as depicted in Figure 6 while in the template definition in Figure 2 PBI's parent is the Feature WI. In addition, Bug WI's child is Task WI according to the template definition not PBI as discovered. Considering the slight deviations, Feature WI has no child PBI WI. Bug WI has no child Task WI. Test Case WI has no PBI WI parent. There is no Bug WI generated from a Test Case WI, this point could not be considered a deviation as bugs could not be generated from the defined test cases but from just exploratory testing [26].

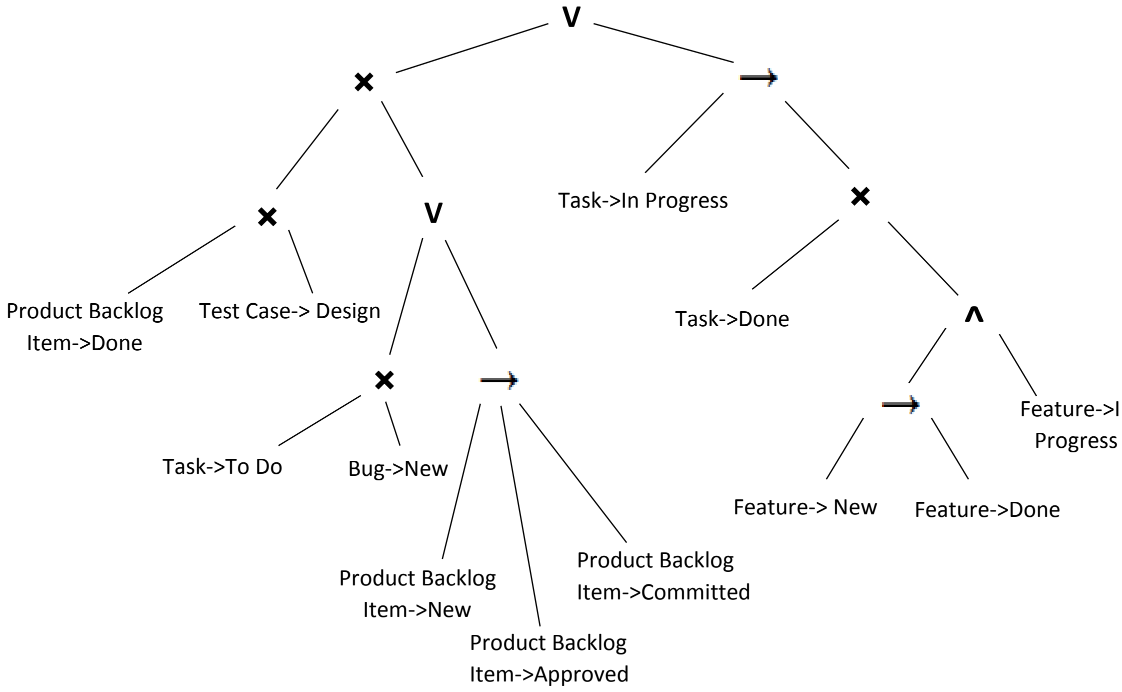


Fig. 6. Discovered Pareto Front as Process Tree for Scrum WIs workflows

VII. Future Work

We target applying conformance checking algorithms to acquire detailed knowledge about the deviations positions concerning software engineering domain. In addition, we are planning to apply enhancements algorithms to provide some recommendation in order to advise the team to conform to the process definition.

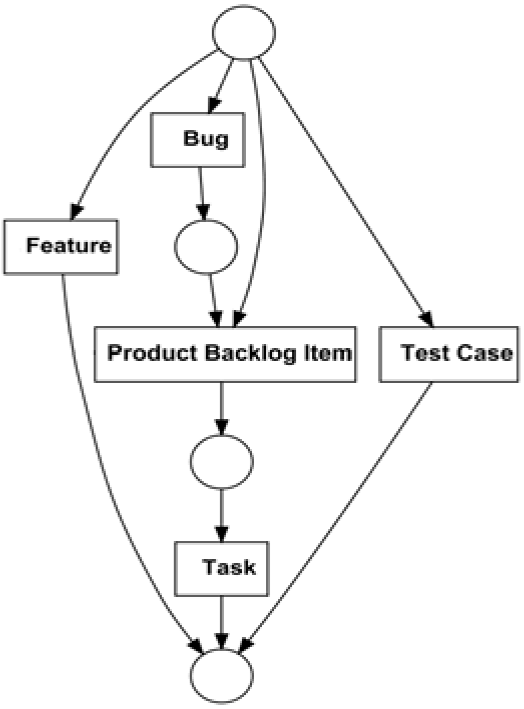


Fig. 7. Discovered Petri-Net for Scrum Process workflow

VIII. Conclusion

Research proved that Process Mining could unveil the deviation from the standard process and define the currently applied of the mostly applied Agile technique that is Scrum [23]. Software Engineering is considered an industry [27, 28, 29]. Investment in this industry is enormous and a process follow-up is needed in order to overcome any costly deviations. The study showed a significant gap between the actually applied process model and the standard process definition, which is obviously helpful. To discover the actual process model the research used ETM*d* Pareto Front algorithm [14] and α-algorithm [9]. In addition, this research used Petri nets, BPMN, and process trees to visualize the discovered models, which proves the overcome of the representational bias by separating representation and visualization.

Acknowledgment

Authors of this paper would like to thank the team members contributed ProM (www.ProcessMining.org) development, and (TU/e) for its support for such great tool. We also would like to thank the contributors who are implementing the extension plug-ins to broaden the utilization of Process Mining in their life.

This work has been done as part of the research and development efforts of CompuPharaohs software engineering team.

References

[1] W. M. P. v. d. Aalst, A. H. M. t. Hofstede and M. Weske, "Business Process Management: A Survey," in International Conference on Business Process Management, Berlin, 2003.

[2] W. v. d. Aalst and K. v. Hee., Workflow Management: Models, Methods, and Systems, MIT press, Cambridge, MA, 2002.

[3] S. Jablonski and C. Bussler, Workflow Management: Modeling Concepts, Architecture, and Implementation, London, UK: International Thomson Computer Press, 1996.

[4] D. Marinescu, Internet-Based Workflow Management: Towards a Semantic, A. Y. Zomaya, Ed., New York: Wiley Interscience, 2002.

[5] W. v. d. Aalst and P. Berens, "Beyond workflow management: product-driven case handling," in International ACM SIGGROUP Conference on Supporting Group Work (GROUP 2001), New York, 2001.

[6] W. v. d. Aalst, Process Mining - Discovery, Conformance and Enhancement of Business Processes, Springer, 2011.

[7] A. Adriansyah, B. v. Dongen and W. v. d. Aalst, "Towards Robust Conformance," in BPM 2010 Workshops, Proceedings of the 6th Workshop on Business Process Intelligence, Berlin, 2011.

[8] W. v. d. Aalst, "Mediating between modeled and observed behavior: The quest for the "right" process: Keynote," in Research Challenges in Information Science (RCIS), Paris, 2013.

[9] W. v. d. Aalst, A. Weijters and L. Maruster, "Workflow Mining: Discovering Process Models from Event Logs.," IEEE Transactions on Knowledge and Data Engineering (TKDE), 2003.

[10] A. d. Medeiros, B. v. Dongen, W. v. d. Aalst and A. Weijters, "Process Mining: Extending the α-algorithm to Mine Short Loops," International Journal Of Engineering And Computer Science (IJESC), 2004.

[11] A. Weijters, W. v. d. Aalst and A. d. Medeiros, Process mining with the heuristics miner-algorithm, Technische Universiteit Eindhoven, 2006.

[12] C. Günther and W. v. d. Aalst, "Fuzzy mining - adaptive process simplification based on multi-perspective metrics," in Business Process Management (BPM), Lecture Notes in Computer Science, Brisbane, Australia, 2007.

[13] C. Günther, Ph.D. thesis: Process Mining in Flexible Environments., Eindhoven University of Technology, 2009.

[14] J. Buijs, Flexible Evolutionary Algorithms for Mining Structured Process Models, Eindhoven: Technische Universiteit Eindhoven, 2014.

[15] A. Rozinat, A. A. d. Medeiros, C. Gunther, A. Weijters and W. v. d. Aalst, "The Need for a Process Mining Evaluation Framework," in Business Process Management Workshops, Berlin, 2008.

[16] J. Buijs, B. v. Dongen and W. v. d. Aalst, "Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity," 2014.

[17] W. M. v. d. Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer, 2011.

[18] Technische Universiteit Eindhoven (TU/e), "Process Mining Manifesto," 2012. [Online]. Available: http://www.win.tue.nl/ieeetfpm/doku.php?id=shared%3Aprocess_mining_manifesto. [Accessed December 2015].

[19] OMG, "Business Process Model and Notation (bpmn) version 2.0," Object Management Group, 3 January 2011. [Online]. Available: http://www.omg.org/spec/BPMN/2.0/PDF. [Accessed December 2015].

[20] A. Scheer, Business Process Engineering, Reference Models for Industrial Enterprises, Berlin: Springer-Verlag Berlin Heidelberg, 1994.

[21] W. v. d. Aalst, A. Adriansyah and B. v. Dongen, "Causal nets: A modeling language tailored towards process discovery," in CONCUR, 2011.

[22] A. Weijters and J. Ribeiro, "Flexible heuristics miner (FHM)," in Computational Intelligence and Data Mining (CIDM), 2011.

[23] D. L. Giudice, "The 2015 State Of Agile Development," Forrester Research, 2015.

[24] Microsoft, "Scrum process work item types and workflow | VS 2013," 2013. [Online]. Available: https://msdn.microsoft.com/library/jj920147%28v=vs.120%29.aspx. [Accessed December 2015].

[25] TU/e, "ProM," Process Mining Group, Math&CS department, Eindhoven University of Technology., 2015. [Online]. Available: http://www.processmining.org/prom/start. [Accessed December 2015].

[26] C. Kaner, A Tutorial in Exploratory Testing, 2008.

[27] C. Baum, The system builders: The story of SDC, System Development Corp, 1981.

[28] M. Campbell-Kelly, From Airline Reservations to Sonic the Hedgehog: A History of the Software Industry, The MIT Press, 2003.

[29] M. A. Cusumano, Microsoft Secrets: How the World's Most Powerful Software Company Creates Technology, Shapes Markets and Manages People, Touchstone, 1998.

A Novel Neural Network Based Method Developed for Digit Recognition Applied to Automatic Speed Sign Recognition

Hanene Rouabeh
Micro Electro Thermal Systems
Research Group
National Engineers School of Sfax
University of Sfax, Tunisia

Chokri Abdelmoula
Micro Electro Thermal Systems
Research Group
National Engineers School of Sfax
University of Sfax, Tunisia

Mohamed Masmoudi
Micro Electro Thermal Systems
Research Group
National Engineers School of Sfax
University of Sfax, Tunisia

***Abstract*—This Paper presents a new hybrid technique for digit recognition applied to the speed limit sign recognition task. The complete recognition system consists in the detection and recognition of the speed signs in RGB images. A pretreatment is applied to extract the pictogram from a detected circular road sign, and then the task discussed in this work is employed to recognize digit candidates. To realize a compromise between performances, reduced execution time and optimized memory resources, the developed method is based on a conjoint use of a Neural Network and a Decision Tree. A simple Network is employed firstly to classify the extracted candidates into three classes and secondly a small Decision Tree is charged to determine the exact information. This combination is used to reduce the size of the Network as well as the memory resources utilization. The evaluation of the technique and the comparison with existent methods show the effectiveness.**

Keywords—Image processing; Road Sign Recognition; Neural Networks; Digit Recognition

I. INTRODUCTION

Nowadays, intelligent driver assistance systems stand as an important component in new vehicle generation. These systems assist the driver in many driving situations and contribute to decreasing the risk of accidents. Artificial perception modules are used to develop various and different driver assistance tasks. Vision-based modules are involved in many types of research in this topic. Numerous assistance applications were exploited such as automatic obstacle avoidance, vehicle detection and recognition, line following, traffic lights detection and many others. This work, deals with the development of an intelligent vision-based system to recognize speed limit signs. In purpose to achieve real-time processing constraints, the complete image processing system is decomposed into sub-processing levels. This paper discusses a new method developed for the identification of the speed limit value. The main goal consists in recognizing the nature of the interior symbol: whether it is a positive speed value candidate and so identifying the exact value or not a speed value candidate. The input data for this task is a small binary image presenting the pictogram extracted from a detected circular road sign with a red border. A detection process precedes the recognition one to detect and extract this type of road signs. The speed limit recognition problem can be addressed as a digit candidate identification which is the subject of this work. The developed approach uses as a first step a Neural Network that classifies the detected candidates into three classes and as a second step a small Decision Tree is used to identify the exact information: a positive digit candidate and so gives the value or a negative candidate. The main contributions proposed in this work consist in reducing the Neural Network architecture and the development of a technique that is well-suited for real-time processing and hardware implementation.

The remainder of this paper is organized as follows: In Section II, a review of previous work is presented. The proposed method is discussed in Section III, whereas the performance evaluation is proposed in Section IV. Concluding points and future work are drawn in the conclusion.

II. RELATED WORK

The majority of existing automatic road sign classification and recognition, as well as candidate identification especially digit candidate use artificial intelligence based algorithms. Many issues have involved Support Vector Machine (SVM) in the recognition stage. Arroyo et al [1] have proposed a method based on SVM to classify signs according to their geometric shapes. Authors in [2] have also used SVM to classify road signs based on shapes. A group of cascade SVM classifiers was used in [3] to classify detected road signs. A new method for pictogram classification was presented in [4] using Decision-tree-based support vector multi-class classifiers. Authors in [5] have used self-organizing maps for the detection and recognition of road signs. In the recognition stage, the nature of information is identified using the distribution of dark pixels in the pictogram of the detected sign. Neural Networks have also been involved in various recognition tasks [6-9]. In [10] a Neural Network was designed to recognize traffic sign patterns. In purpose to minimize the number of input units from 2700 units that presents the number of all pixels in the three channels R,G and B of the detected pictogram to a reduced number, they calculate three normalized averages, 30 inputs from the vertical histogram and 30 inputs from the horizontal histogram. These 63 parameters present then the total input units of their designed Network. Maurice et al [11] have used Convolutional Neural Networks for speed signs recognition. The input layer is composed of 32x32 units presenting the pixels of the extracted pictogram. The recognition of the speed value presents a challenging problem among these researches. There are issues that use the detected sign matrix as input to the recognizer [12]. Others consider the extracted pictogram matrix as input [11]. In [13] the binary matrix of the first extracted digit is used, whereas in [14] authors prefer to recognize each extracted candidate separately. In spite of the many advantages shown by Neural Networks in recognition problems due to their high learning capabilities, the real-time processing and memory resources utilization is related to the complexity and the architecture of the designed Network. One of the key points that well contribute to simplify the architecture is reducing the number of input units. This research interest the conception of a simple and performing technique for digit candidate recognition applied to speed sign recognition. The proposed method is presented in the next section.

III. Proposed Method

This section describes the developed digit recognition technique. This technique is part of a complete recognition system which is developed for rounded speed limit signs with red border detected on an RGB image. The detection stage allows detecting this type of signs and extracting the first character of the interior information which can be the first digit of a speed value or another type of information. The extraction step is presented in Fig. 1. Where (a) is the detected circular road signs in the RGB image. The size of this extracted pattern is not fix for different images, for this example the size is 33x33. A resizing is applied then to all detected signs to normalize the size to 32x32. (b) is the Red component of the resized sign. (c) is the result of converting (b) to binary image using a simple thresholding method. (d) is the first character extracted from (c). The size of (d) is 32xM where M is a scalar less than or equal to 32.

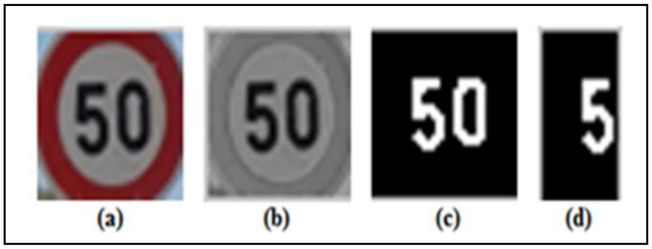


Fig. 1. Digit candidate extraction from a detected sign

The use of Neural Networks in the recognition stage for this application was investigated in many researches. Authors in [15-16] have developed Networks were the input layer consists in the extracted digit pattern. The total number of input neurons is equal to the total number of pixels. Respectively 400 pixels (20x20) for [15] and 72 pixels (6x12) for [16]. The use of a large number of input units requires more computation time and more memory resources. A simple method that reduces the number of inputs and uses a simple Neural Network architecture was tried and developed. The technique is divided into two parts: In the first one a Neural Network is designed to determine the class of the digit candidate from three probably classes. Then in the second part a small Decision Tree is charged to recognize the exact signification (a number from 0 to 9 or not a number). The number of units in the input layer is equal to 11 units. These input values are extracted from the digit candidate binary image. The realization of the recognition task is described in the following steps.

A. Neural Network creation

This section describes the different steps achieved in the design of the Network architecture. The used Net is a feed-forward multi-layer Network.

A pretreatment is carried on the binary image of Fig. 1 (d) to extract parameters to be used as the Network input.

As shown in Fig. 2 a left projection of the candidate image is done to calculate at each row the distance to the first foreground pixel.

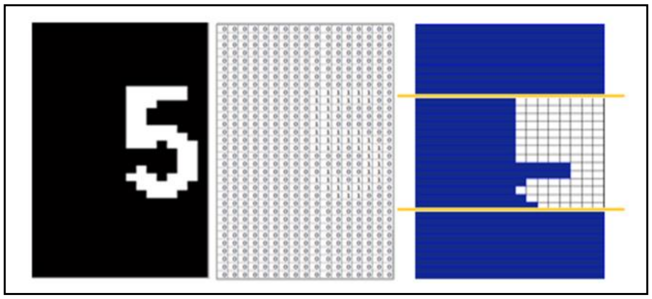


Fig. 2. Left projection and distance calculation

An array of 32 distances ‘**dis**’ is obtained as follows:

From row 1 to 10

17	17	17	17	17	17	17	17	17	9

From row 11 to 20

9	9	9	9	9	9	9	14	14	10

From row 21 to 32

9	10	11	17	17	17	17	17	17	17	17	17

This array is used to fill the ‘**dr**’ array consisting in the 11 Neural Network inputs as described in TABLE I, where D_{-1}, D_0 and D_1 present the difference between two consecutive distances. If these distances are equal then the ‘**dr**’ case corresponding to D_0 is incremented. D_{-1} is incremented if the left distance is greater than the right one and D_1 is incremented in the opposite situation. Only the part containing the digit is considered for which the corresponding distances are marked in orange color. Fig. 3 explains the idea. The following array describes the obtained input array of the preceding example.

D_{-1}	D_0	D_1	D_{-1}	D_0	D_1	D_{-1}	D_0	D_1	D_{-1}	D_0
0	7	1	2	1	2	0	0	0	0	0

The Pseudo-code of the process is exhibited as follows:

```
k= 1
for i in 1 to 31
    if dis(i) < number of columns
        if dr(k+2)>0 & dis(i+1)<=dis(i)
            k=k+3
        end
        if dis(i+1) < dis(i) then
            %%D-1: incremented %%
            dr(k) =dr(k)+1
        else if dis(i+1) = dis(i)
            %%D0: is incremented%%
            dr(k+1) =dr(k+1)+1
        else
            %%D1: is incremented%%
            dr(k+2) =dr(k+2)+1
        end
    end
end
```

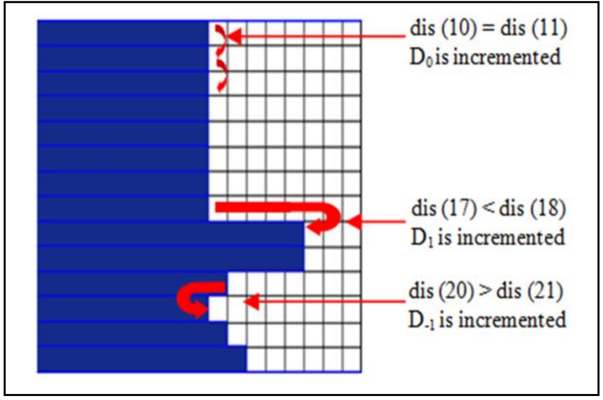


Fig. 3. Process to fill the input array

Fig. 4 shows extracted digit candidates for numbers varing from 0 to 9. These candidates are extracted from detected road signs in road images. TABLE I exhibits the corresponding Neural Network input arrays determined as explained previously for these candidates. The analysis of a large number of extracted digits have shown that using this input vector allows to classify them into three classes. The first one is including 1,2,4 and 7 , the second one is including 0 and 6 and the third class includes 3,5,8 and 9. The designed Neural Network is charged to determine the class of the digit candidate. But given that the detection module can detect not only speed signs but some other signs can be detected due to their rounded geometry and red border like in Fig. 5, the separated characters of the detected pictogram are extracted and processed seperately to guarantee the accuracy of the method.

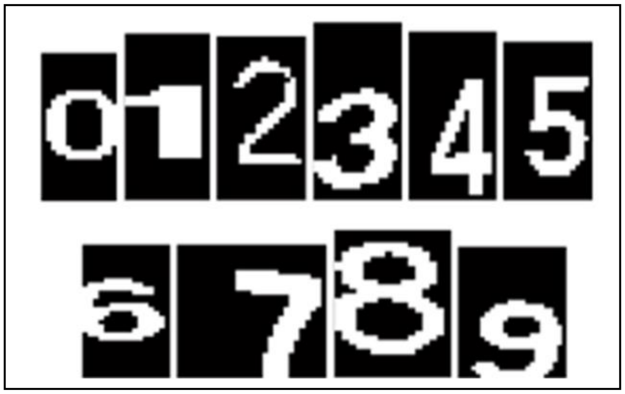


Fig. 4. Examples of extracted digit candidates

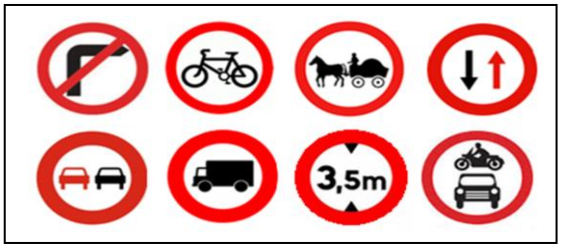


Fig. 5. Other circular road signs with red border

TABLE I. CORRESPONDING NN INPUT VECTORS RELATED TO SAMPLES OF FIG. 4

Value	NN input array 'dr'
0	3 7 1 0 1 2 0 0 0 0 0
1	2 1 1 0 9 0 0 0 0 0 0
2	4 1 1 7 7 0 0 0 0 0 0
3	4 1 1 2 1 2 2 1 3 0 0
4	7 1 0 1 0 3 0 0 0 0 0
5	1 9 1 1 3 1 0 1 2 0 0
6	2 1 0 2 0 0 0 0 0 0 0
7	0 4 2 5 1 4 0 0 0 0 0
8	3 2 1 0 2 2 2 6 1 0 1
9	4 4 4 2 2 2 0 0 0 0 0

The final Neural Network (NN) architecture is shown in Fig. 7. The input layer is composed of 11 neurons. The output layer consists of two neurons presenting the candidate class as follows:

class 1: 0 0 (1,2,4,7)
class 2: 0 1 (0, 6)
class 3: 1 1 (3,5,8,9)

The number of hidden layers and hidden units was fixed after empirical training of the Network changing the training and transfer functions and other parameters. The final architecture was fixed when desired mean-squared error was reached taking into account reducing the network complexity. In purpose to simplify the hardware implementation of the designed system the elliot symmetric transfer function [17] was used in this work. This function is defined by equation 1. Its main advantage that is suitable for hardware implementation. Contrary to other sigmoid function, it does not require the implementation of any exponential function which requires high memory resources in hardware implementation.

$$\text{Elliotsig}(x) = \frac{x}{1+\text{abs}(x)} \qquad (1)$$

Fig. 6 shows the comparison between the hyperbolic tangent sigmoid transfer function and elliot symmetric one.

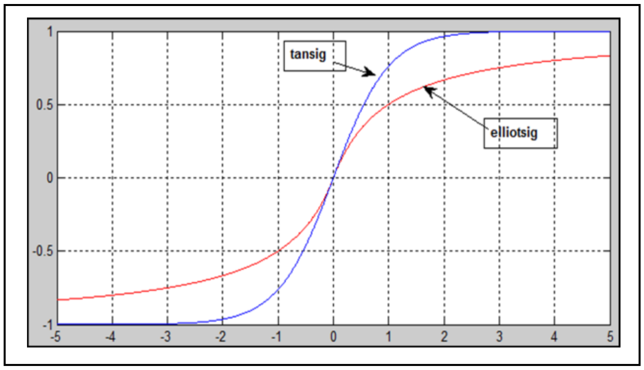


Fig. 6. Hyperbolic tangent sigmoid transfer function Vs Elliot symmetric one

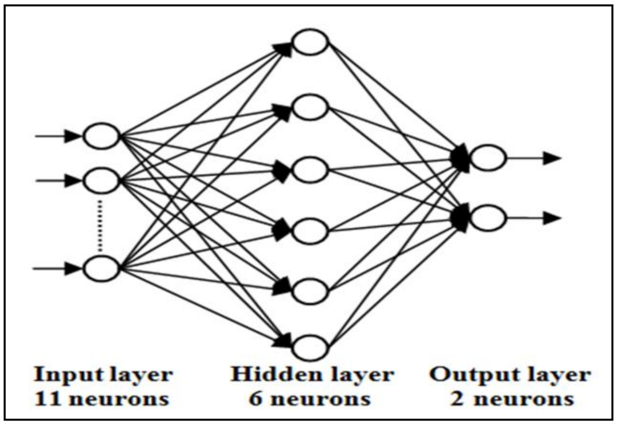


Fig. 7. Final architecture of the Neural Network Classifier

TABLE II. NEURAL NETWORK EVALUATION

Class	Total samples	Correct classification	Classification %
1	69	67	97.10 %
2	34	31	91.18 %
3	65	61	93.85 %

B. Decision Tree realization

The second part of the recognition system is accomplished with a small Decision Tree (DT) shown in Fig. 8. This DT is designed to determine the nature of the extracted candidate firstly classified with the NN. Some parameters are used in the DT which are shown in Fig. 9 and are respectively:

Nbi: Presents the number of intersection of the ith column with the character, where the ith column contains the center of geometry.

Dr(1) is the value of the first value of the "**dr**" array.

Nbt : Presents the maximum difference between two consecutive distances where the right one is greater than the left one.

b: Presents the number of consecutive zeros in the central column below the central pixel.

h: Presents the number of consecutive zeros in the central column above the central pixel.

ND: Means Not a digit.

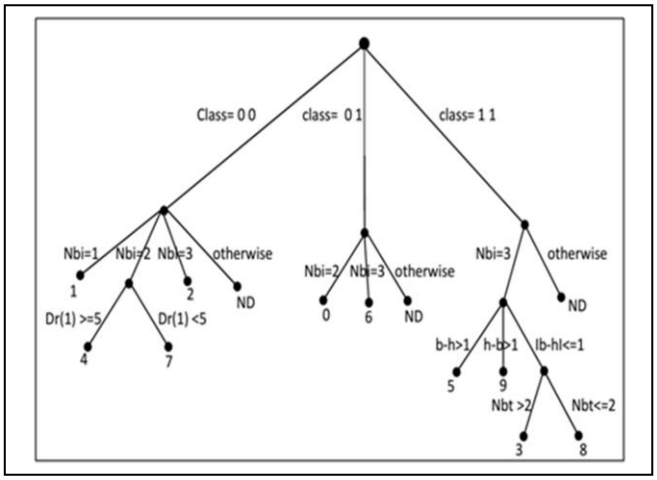


Fig. 8. Architecture of the Decision Tree

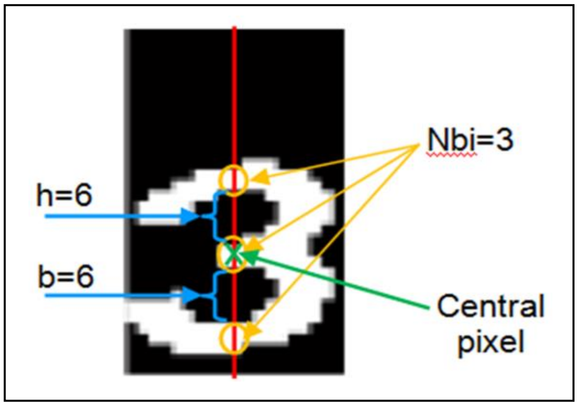


Fig. 9. Extraction of the different parameters

IV. EVALUATION

Fig. 10 shows the complete recognition process where Nc is the number of separated characters in the extracted pictogram. Given the binary matrix of the extracted first digit candidate, NN inputs are calculated and passed throw the network for the first classification. Then the DT is charged to identify whether it is a positive or a negative digit candidate.

The second character is then processed similarly if the first candidate is identified as a positive digit candidate to obtain the final decision.

If the information contains only one symbol it will be rejected and decided as no speed sign given that a speed value should be composed of 2 or 3 separate characters.

TABLE III presents different Network architectures and their relative mean-squared error. TABLE IV presents the classification results carried on a number of testing images for the best Network architecture, the one with one hidden layer and 6 hidden units. This table exhibits the classification percentage for the NN and the DT separately and then for the whole system.

TABLE V shows the performance comparison of the proposed method with the method discussed in [16].

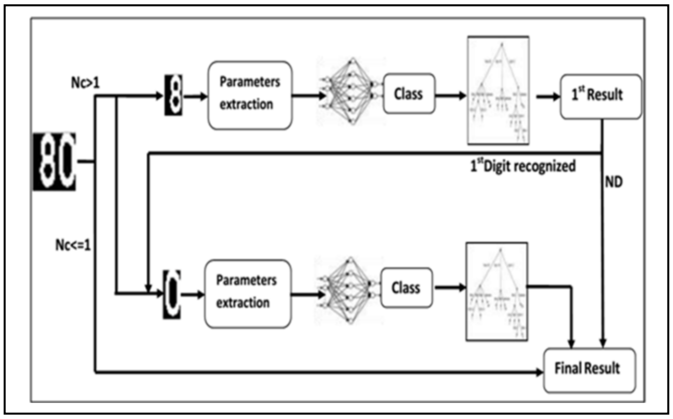


Fig. 10. Speed value recognition process

TABLE III. NETWORK ARCHITECTURES

NN Architecture	Mean Squared error
11- 5 - 2	0.0980
11- 6 - 2	0.0298
11- 7 - 2	0.0777
11- 8 - 2	0.1115
11-10 - 2	0.1081
11-12 - 2	0.0811
11-10 - 7 -2	0.0473
11-13 - 5 -2	0.0541

TABLE IV. CLASSIFICATION RESULTS

Proposed Method		
NN	DT	NN+DT
94.64%	98.55%	93.45%

TABLE V. PERFORMANCE COMPARISON

	Architecture	Total signs number	Recognition %
This work	NN+DT 11-6-2	168	93.45%
Method in [16]	NN 72 - 10 -10	128	82.8%

TABLE IV shows the classification rate of the proposed method carried on a 168 testing samples. The recognition percentage of the DT is the percentage of the correct recognized samples among the 159 correct classified samples by the NN. These results show the robustness and accuracy of the developed method. TABLE V exhibits the advantages of the method compared to [16]. The developed technique gives a high recognition rate with simple and soft architecture.

V. CONCLUSION

This paper has presented a new, robust and efficient approach to identify positive digit candidates extracted from detected road signs. The main advantages of this work consist in the simplicity of the Network architecture and the short computation time. Compared to approaches [12, 13, 15, and 16] in which an important number of input units was used that complicate calculations and require a high number of hidden neurons. The use of the DT as a complementary task didn't complicate the system. Indeed it is composed of a minimum number of branches and leaves that require a very short computation time. In order to minimize the hardware implementation of the designed system the Elliot symmetric transfer function is used in the NN. This function requires simple hardware calculations contrary to other sigmoid functions that need high memory resources to implement exponential and other geometric functions. The recognizer discussed in this paper presents a part of a complete intelligent speed sign recognition system that is designed to be implanted on an FPGA hardware architecture in future works.

REFERENCES

[1] S. Lafuente-Arroyo, P. Gil-Jimenez, R. Maldonado-Bascon, F. Lopez-Ferreras, and S. Maldonado-Bascon "Traffic sign shape classification evaluation I: SVM using distance to borders", Intelligent Vehicles Symposium, 2005, Proceedings. IEEE, Las Vegas, pp. 557-562, June 2005.

[2] P. Gil-Jimenez, H. Gomez-Moreno, P. Siegmann, S. Lafuente-Arroyo, and S. Maldonado-Bascon, "Traffic sign shape classification based on Support Vector Machines and the FFT of the signature of blobs", Intelligent Vehicles Symposium, 2007 IEEE, Istanbul, pp. 375-380, June 2007.

[3] Jack Greenhalgh and Majid Mirmehdi "Real-Time Detection and Recognition of Road Traffic Signs" IEEE Transactions on Intelligent Transportation Systems, VOL. 13, NO. 4, December 2012

[4] Hossein Pazhoumand-Dar and Mehdi Yaghobi, "DTBSVMs: a New Approach for Road Sign Recognition," 2010 Second International Conference on Computational Intelligence, Communication Systems and Networks.

[5] Miguel S. Prietoa and Alastair R. Allen, Using self-organising maps in the detection and recognition of road signs , Image and Vision Computing Volume 27, Issue 6, 4 May (2009), pp. 673-683 .

[6] M. Rahman, F. Mousumi, E. Scavino, A. Hussain, and H. Basri, "Realtime road sign recognition system using artificial neural networks for Bengali textual information box," in Proc. ITSim, 2008, vol. 2, pp. 1–8.

[7] E. Cardarelli, P. Medici, P. P. Porta, and G. Ghisio, "Road sign shapes detection based on Sobel phase analysis," in Proc. IEEE IVS, 2009, pp. 376–381.

[8] M. Hossain, M. Hasan, M. Ali, M. Kabir, and A. Ali, "Automatic detection and recognition of traffic signs," in Proc. RAM, 2010, pp. 286–291.

[9] R. Vicen-Bueno, R. Gil-Pita, M. Rosa-Zurera, M. Utrilla-Manso, and F. Lopez-Ferreras, "Multilayer Perceptrons Applied to Traffic Sign Recognition Tasks", LNCS 3512, IWANN 2005, J. Cabestany, A. Prieto, and D.F. Sandoval (Eds.), Springer-Verlag, Berlin, Heidelberg, 2005, pp. 865-872.

[10] Auranuch Lorsakul and Jackrit Suthakorn "Traffic sign recognition using neural network on open CV: toward intelligent vehicle/driver assistance system" 4th International Conference on Ubiquitous Robots and Ambient Intelligence ,2007 http://dspace.li.mahidol.ac.th/handle/123456789/2714

[11] Peemen, Maurice, Mesman, Bart, Corporaal, Henk "Speed Sign Detection and Recognition by Convolutional Neural Networks"http://parse.ele.tue.nl/system/attachments/11/original/paperspeedsigncnn.pdf

[12] Damavandi, Y.B., Mohammadi, K., (2004), "Speed limit traffic sign detection and recognition", IEEE Conference on Cybernetics and Intelligent Systems, pp. 797 – 802

[13] Torresen J.. Bakke J.W., Sekanina L. (2004) **"**Efficient recognition of speed limit signs", IEEE Conference on Intelligent Transportation Systems, pp. 652 – 656

[14] Moutarde F., Bargeton A., Herbin A, Chanussot L., (2007), "Robust on-vehicle real-time visual detection of American and European speed limit signs, with a modular Traffic Signs Recognition system" Proceedings of IEEE Intelligent Vehicles Symposium, pp. 1122-1126

[15] Marcin L. Eichner, Toby P. Breckon "Integrated Speed Limit Detection and Recognition from Real-Time Video" 2008 IEEE Intelligent Vehicles Symposium.

[16] Martinović, A., Glavaš, G. , Juribašić, M. , Sutić, D. and Kalafatić, Z."Real-time detection and recognition of traffic signs" MIPRO, 2010 Proceedings of the 33rd International Convention, Opatija, Croatia

[17] http://www.mathworks.com/help/nnet/ref/elliotsig.html?requestedDomain=www.mathworks.com

A Secure Cloud Computing Architecture Using Homomorphic Encryption

Kamal Benzekki
Laboratory of Computer Networks and Systems
Department of Mathematics and Computer Science
Moulay Ismail University, Faculty of Sciences, Meknes, Morocco

Abdeslam El Fergougui
Laboratory of Computer Networks and Systems
Department of Mathematics and Computer Science
Moulay Ismail University, Faculty of Sciences, Meknes, Morocco

Abdelbaki El Belrhiti El Alaoui
Laboratory of Computer Networks and Systems
Department of Mathematics and Computer Science
Moulay Ismail University, Faculty of Sciences, Meknes, Morocco

***Abstract*—The Purpose of homomorphic encryption is to ensure privacy of data in communication, storage or in use by processes with mechanisms similar to conventional cryptography, but with added capabilities of computing over encrypted data, searching an encrypted data, etc. Homomorphism is a property by which a problem in one algebraic system can be converted to a problem in another algebraic system, be solved and the solution later can also be translated back effectively. Thus, homomorphism makes secure delegation of computation to a third party possible. Many conventional encryption schemes possess either multiplicative or additive homomorphic property and are currently in use for respective applications. Yet, a Fully Homomorphic Encryption (FHE) scheme which could perform any arbitrary computation over encrypted data appeared in 2009 as Gentry's work. In this paper, we propose a multi-cloud architecture of N distributed servers to repartition the data and to nearly allow achieving an FHE.**

Keywords—multi-cloud; privacy; fully homomorphic encryption; distributed System; confidentiality

I. Introduction

Cryptosystems supply mechanisms to ensure data confidentiality and integrity. If the data is always encrypted in the cloud, then control is not lost, and the concerns are removed. When an encryption algorithm does not allow arbitrary computation over encrypted data, the encrypted data must be decrypted before the computation, and the decrypted data is no longer under control.

Cloud computing is infeasible for many business organizations if they need to download sensitive data from the cloud to a trusted computer in order to perform operations, and then send the encrypted results backed to the cloud. Encrypted data has historically been impossible to operate on without first decrypting them. There are some encryption algorithms that allow arbitrary computation on encrypted data. For instance, RSA is a multiplicatively homomorphic encryption algorithm where the decryption of the product of two encrypted data will be the product of the two plain data. However, RSA doesn't allow addition operation nor the combination of multiplications and additions. Later, FHE has appeared [1] to perform unlimited chaining of algebraic operations in the cipherspace, which means that an arbitrary number of additions and multiplications can be applied to encrypted operands. Unfortunately, all implementations of FHE schemes showed that this technique is still much too slow for practical applications.

In this work, we will be focusing on the application of N distributed servers and N cloud systems to homomorphic encryption in order to perform a nearly FHE scheme for the security of data and applications, particularly the possibility to execute the calculations of encrypted confidential data without decrypting them.

The remainder of the paper is organized as follows. Homomorphic encryption and related definitions are introduced in section II. In section III, we discuss the Somewhat Homomorphic Scheme. In section IV, we present some examples of partially homomorphic cryptosystems. Finally in the section V, we propose a secure multi-cloud architecture for processing encrypted data. The perspectives are mentioned in section VI with the conclusion.

II. Toward Homomorphic Encryption

The security requirements for data and algorithms have become very strong in the last few years. Due to the vast growth of technology, a great variety of attacks on digital goods and technical devices are enabled. For storing and reading data securely, there exist several possibilities like secure data encryption. The problem becomes more complex when asking for the possibility to compute (publicly) with encrypted data or to modify functions in such a way that they are still executable while the privacy is ensured. That is where homomorphic cryptosystems can be used.

The notion and idea of fully homomorphic schemes was introduced by Rivest, Adle-man and Dertouzos in [2] shortly after the invention of RSA [3]. They asked for an encryption function that permits encrypted data to be operated on without preliminary decryption of the operands, and they called those schemes privacy homomorphisms. Even in 1978 this was a highly important matter, it is even more important nowadays. While the partially homomorphic properties of schemes like RSA, Paillier, ElGamal, etc. have been acknowledged ever since, it was not before 2009 when a young IBM researcher published the first working fully homomorphic cryptosystem based on lattices.

Among the homomorphic encryption we distinguished according to their operation to assess on raw data. The additive homomorphic encryption (addition of the raw data) is the Pailler [4] and Goldwasser-Micalli[5] cryptosystems and the multiplicative homomorphic encryption (only products on raw data) is the RSA [6] and El Gamal [7] cryptosystems.

A. Definition of a Homomorphic Encryption Scheme

A public-key encryption scheme E=(KeyGen, Enc, Dec) is homomorphic if for all k and all (pk,sk) output from KeyGen(k), it is possible to define groups M, C so that:

- The plaintext space M, and all ciphertexts output by Enc_{pk} are elements of C.
- For any m_1 , $m_2 \in M$ and c_1 , $c_2 \in C$ with $m_1 = Dec_{sk}(c_1)$ and $m_2 = Dec_{sk}(c_2)$ it holds that:

$Dec_{sk}(c_1 * c_2) = m_1 * m_2$

Where the group operations ∗ are carried out in C and M, respectively.

In other words, a homomorphic cryptosystem is a PKS with the additional property that there exists an efficient algorithm (Eval) to compute an encryption of the sum or/and the product of two messages given the public key and the encryptions of the messages, but not the messages themselves.

Moreover, a fully homomorphic scheme is able to output a ciphertext that encrypts f $(m_1,...,m_t)$, where f is any desired function, which of course must be efficiently computable. No information about $m_1,..., m_t$ or f $(m_1,...,m_t)$, or any intermediate plaintext values should leak. The inputs, outputs and intermediate values are always encrypted, and therefore useless for an adversary. Before we take a closer look on fully homomorphic encryption schemes, we will need another important notion from information theory.

B. Circuits

Informally speaking, circuits are directed, acyclic graphs where nodes are called gates and edges are called wires. Depending on the nature of the circuit the input values are integers, boolean values, etc. and the corresponding gates are set operations and arithmetic operations or logic gates (AND, OR, NOR, NAND, ...). In order to evaluate a function f, we express f as a circuit and topologically arrange its gates into levels which will be executed sequentially.

Example. Assume the function f outputs the expression: A·B+B·C·(B+C) on input (A,B,C). Then the following circuit represents the function f, with the logic gates AND and OR.

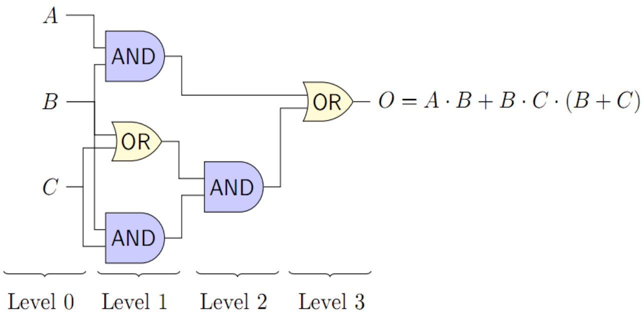


Fig. 1. Example for circuit representation

Two important complexity measures for circuits are size and depth.

The size of a circuit C is the number of its non-input gates. The depth of a circuit C is the length of its longest path, from an input gate to the output gate, of its underlying directed graph.

This yields to another definition of fully homomorphic encryption [8]:

ciphertexts $\Psi = \{c_1 , ..., c_t \}$ where $c_i \leftarrow Enc_{pk}(m_i)$, outputs

$$c \leftarrow Eval_{pk}(C, \Psi)$$

under pk.

There is another way to construct fully homomorphic encryption schemes. To understand how this transformation works, we need the following definitions and corollaries.

Definition : A homomorphic encryption scheme E is said to be correct for a family C_E of circuits if for any pair (sk, pk) output by $KeyGen_E(\lambda)$ any circuit $C \in C_E$, any plaintext $m_1,...,m_t$, and any ciphertexts $\Psi = c_1, ...,c_t$

with $c_i \leftarrow Enc_{pk}(m_i)$, it is the case that:

If $c \leftarrow Eval_E(pk, C, \Psi)$, then $Dec_E(sk, c) \rightarrow C(m_1, ...,m_t)$

Except with negligible probability over the random coins in $Eval_E$.

Definition: A homomorphic encryption scheme E is compact, if there is a polynomial f so that, for every value of the security parameter λ, E's decryption algorithm can be expressed as a circuit D_E of size at most f (λ).

A homomorphic encryption scheme E compactly evaluates circuits in C_E if E is compact and also correct for circuits in C_E.

Corollary: A homomorphic encryption scheme E is fully homomorphic if it compactly evaluates all circuits.

This demand is considered to be almost too strong for practical purposes, hence it uses a certain relaxation to include leveled schemes, which only evaluate circuits of depth up to some d, and whose public key length may be poly(d).

Definition: (leveled fully homomorphic). A family of homomorphic encryption schemes $\{E^{(d)} : d \in Z_+ \}$ is said leveled fully homomorphic if, for all $d \in Z_+$, it all uses the same decryption circuit, $E^{(d)}$ compactly evaluates all circuits of depth at most d (that use some specified set of gates), and the computational complexity of $E^{(d)}$'s algorithms is polynomial in λ, d, and (in the case of $Eval_E$) the size of the circuit C.

An encryption scheme which supports both addition and multiplication (a fully homomorphic scheme) thereby preserves the ring structure of the plaintext space and is therefore far more powerful. Using such a scheme makes it possible to let an untrusted party do the computations without ever decrypting the data, and therefore preserving their privacy.

A widely esteemed application of homomorphic encryption schemes is cloud computing. Presently, the need for cloud computing is increasing fast, as the data we are processing and computing on is getting bigger and bigger every day, with the effect that a single person's computation power does not suffice anymore. Hence, it is favorable to use someone else's power without losing the privacy we seek.

Say, Alice wants to store a sensitive file $m \in \{0, 1\}_n$ on Bob's server. So she sends Bob Enc(m_1), ..., Enc(m_n). Assume that the file is a database (a list of people with specific information about them) and Alice wants to find out how many of them are 25 years old. Instead of retrieving the data from Bob, decrypting it and searching for the wanted information, she will ask Bob to do the computations, without him knowing what or who he is computing on.

The answer from Bob comes in form of a ciphertext which only she can decrypt with her secret key.

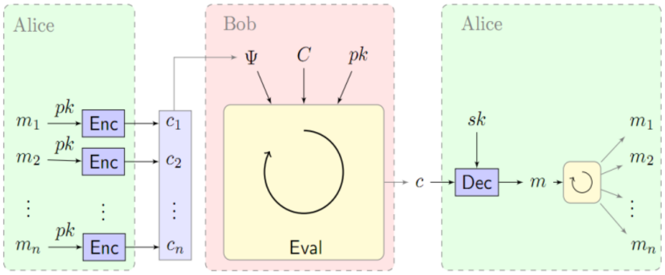


Fig. 2. Diagram of a homomorphic encryption scheme

The benefit of fully homomorphic encryption has long been recognized. The question for constructing such a scheme arose within a year of the development of RSA [2].

For more than 30 years, it was unclear whether fully homomorphic encryption was even achievable. During this period, the best encryption system was the Boneh-Goh-Nissim cryptosystem [9] which supports evaluation of an unlimited number of addition operations but one multiplication at the most.

A common reason why a scheme cannot compute circuits of a certain depth is that after a certain amount of computations too much error accumulates, which causes the decryption to obtain a wrong value. The decryption usually is able to handle small amounts of error within a certain range and bootstrappable encryption enables "refreshing" after some time. The basic idea of "refreshing" is to encrypt under a first key. Compute until right before the error grows too large. Encrypt under a second key. Compute the decryption circuit, which since it stopped before the error grew too large, gives the correct value encrypted under the second key. The first key is no longer required. Continue computation under the second key, and repeat with a new key as often as needed. When the computation has finished, decrypting with the last used key gives the original plaintext.

Gentry's method can be broken down into three major steps:

Step 1: Constructing an encryption scheme using ideal lattices that is somewhat homomorphic, which means it is limited to evaluating low-degree polynomials over encrypted data. This scheme is very similar to the Goldreich-Goldwasser-Halevi scheme published in 1997 [10] which is based on lattice problems as well.

Step 2: "Squashing" the decryption circuit of the original somewhat homomorphic scheme to make it bootstrappable.

Step 3: Bootstrapping the slightly augmented original scheme of step 2 to yield the fully homomorphic encryption scheme. This will be done with a "refreshing" procedure.

The innovative idea of Gentry's method of creating a fully homomorphic scheme out of a somewhat homomorphic scheme is the method of squashing and boot-strapping. Mathematically the most appealing step is the first step.

III. The Somewhat Homomorphic Scheme

The aim of this somewhat homomorphic scheme (SHS) is to construct an encryption scheme that is "almost" bootstrappable with respect to a universal set of gates. The first step is to design a SHE scheme which is a scheme that supports *some* computations over encrypted data. Gentry then showed that if you can manage to design a SHE scheme that supports the evaluation of its own decryption algorithm (and a little more), then there is a general technique to transform the SHE scheme into a FHE scheme. A SHE that can evaluate its own decryption algorithm homomorphically is called *bootstrappable* and the technique that transforms a bootstrappable SHE scheme into a FHE scheme is called *bootstrapping*.

Bootstrapping. So how does bootstrapping work and why is bootstrappability such a useful property? To understand this, you first have to know how the currently-known SHE schemes work. Roughly speaking, the ciphertexts of all these schemes have noise inside of them and unfortunately this noise gets larger as more and more homomorphic operations are performed. At some point, there is so much noise that the encryptions become useless (i.e., they do not decrypt correctly). This is the main limitation of SHE schemes and this is the reason that they can only perform a restricted set of computations. Bootstrapping allows us to control this noise.

The idea is to take a ciphertext with a lot of noise in it and an encryption of the secret key and to homomorphically decrypt the ciphertext. Note that this can only work if the SHE scheme has enough homomorphic capacity to evaluate its own decryption algorithm which is why we need the SHE scheme to be bootstrappable. This homomorphically computed decryption will result in a new encryption of the message but without the noise (or at least with less noise than before). More concretely, say we have two ciphertexts:

$c_1 = E_{pk}(m_1)$ *and* $c_2 = E_{pk}(m_2)$

with noise n_1 and n_2, respectively. We can multiply these encryptions using the homomorphic property of the SHE scheme to get an encryption*:*

$C_3 = E_{pk}(m_{1} \times m_{2})$

of $m_1 \times m_2$ under key *pk* ,but C_3 will now have noise $n_1 \times n_2$. The idea behind bootstrapping is to get rid of this noise as follows. First, we encrypt C_3 and *sk* under *pk* .This results in two new ciphertexts

$C_4 = E_{pk}(C_3) = (E_{pk}(m_1 \times m_2))$ *and* $C_5 = E_{pk(sk)}$

Given C_4 and C_5, we now *homomorphically* decrypt C_4 using C_5. In other words, we compute the following operation over C_4 and C_5: “decrypt $c_3 = E_{pk}(m_1 \times m_2)$ using *sk*“. This is allowed since the scheme has enough homomorphic capacity to evaluate its own decryption algorithm.

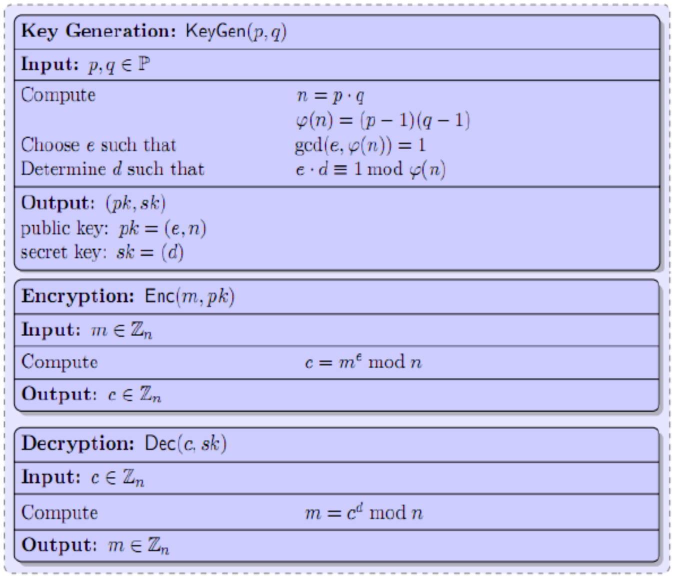
By using this technique throughout a computation whenever the ciphertexts get too noisy, we can remove the main limitation of the SHE scheme and turn it into a FHE scheme.

It turns out that constructing a bootstrappable SHE scheme is difficult. To do this, Gentry had to build his scheme using sophisticated techniques [1] so a lot of the recent work in FHE has tried to figure out how to design simpler bootstrappable SHE schemes.

IV. Partially Homomorphic Cryptosystems

A. RSA-A Multiplicatively Homomorphic Scheme

In 1978, Rivest, Shamir, and Adleman published their public-key cryptosystem which only uses elementary ideas from number theory, in their paper "A Method for Obtaining Digital signatures and Public-Key Cryptosystems" [3]. It was one of the first homomorphic cryptosystem. The RSA cryptosystem is the most widely used public-key cryptosystem. It may be used to provide both secrecy and digital signatures and its security is based on the intractability of the integer factorization problem.



Key Generation: KeyGen(p, q)

Input: $p, q \in \mathbb{P}$

Compute	$n = p \cdot q$ $\varphi(n) = (p-1)(q-1)$
Choose e such that	$\gcd(e, \varphi(n)) = 1$
Determine d such that	$e \cdot d \equiv 1 \bmod \varphi(n)$

Output: (pk, sk)
public key: $pk = (e, n)$
secret key: $sk = (d)$

Encryption: Enc(m, pk)

Input: $m \in \mathbb{Z}_n$

Compute $c = m^e \bmod n$

Output: $c \in \mathbb{Z}_n$

Decryption: Dec(c, sk)

Input: $c \in \mathbb{Z}_n$

Compute $m = c^d \bmod n$

Output: $m \in \mathbb{Z}_n$

Fig. 3. RSA Algorithm

The **encryption** algorithm takes as input a message m from the plaintext space Z_n and computes the according ciphertext $c = m^e \bmod n$. This integer $c \in Z_n$ cannot be traced back to the original message without the knowledge of p and q, which will be proved later in this section.

Decryption takes as input the ciphertext c and the secret key (d, n) and computes $m = c^d \bmod n$. Since d is the inverse of e in Z_n this is indeed the original message.

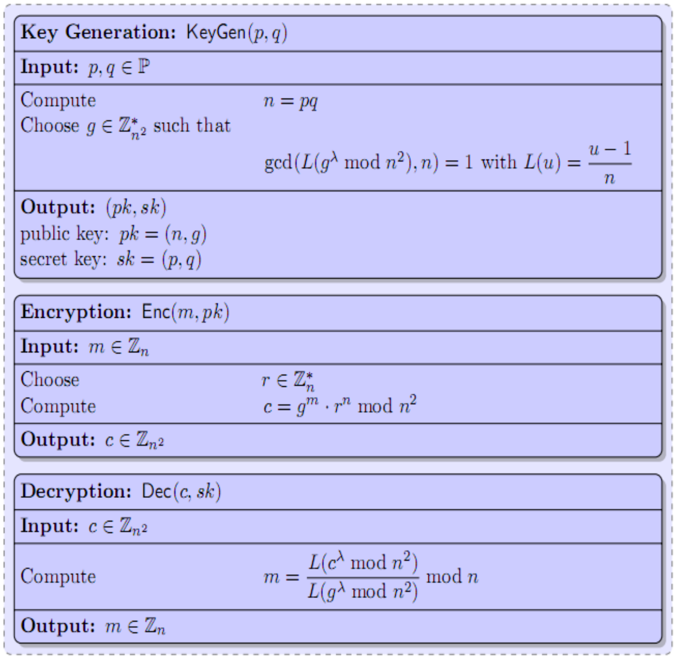
The three steps (key generation, encryption and decryption) can be found in the following table.

B. Paillier - An Additively Homomorphic Scheme

Pascal Paillier introduced his cryptosystem in 1999 published paper "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes" [11]. The proposed technique is based on composite residuosity classes, whose computation is believed to be computationally difficult. It is a probabilistic asymmetric algorithm for public key cryptography and inherits additive homomorphic properties.

The encryption procedure takes as input a message $m \in Z_n$ and randomly chooses an integer r in $Z*$, this random number is used to fulfill the probabilistic algorithm’s n property, that one plaintext can have many ciphertexts. It is later shown that this random variable does not impede the correct decryption, but has the effect of changing the corresponding ciphertext.

The three steps (key generation, encryption and decryption) can be found in the following table:



Key Generation: KeyGen(p, q)

Input: $p, q \in \mathbb{P}$

Compute	$n = pq$
Choose $g \in \mathbb{Z}^*_{n^2}$ such that	$\gcd(L(g^\lambda \bmod n^2), n) = 1$ with $L(u) = \frac{u-1}{n}$

Output: (pk, sk)
public key: $pk = (n, g)$
secret key: $sk = (p, q)$

Encryption: Enc(m, pk)

Input: $m \in \mathbb{Z}_n$

Choose	$r \in \mathbb{Z}^*_n$
Compute	$c = g^m \cdot r^n \bmod n^2$

Output: $c \in \mathbb{Z}_{n^2}$

Decryption: Dec(c, sk)

Input: $c \in \mathbb{Z}_{n^2}$

Compute $m = \frac{L(c^\lambda \bmod n^2)}{L(g^\lambda \bmod n^2)} \bmod n$

Output: $m \in \mathbb{Z}_n$

Fig. 4. Paillier Algorithm

V. Our Architecture

The fully homomorphic encryption schemes [1] are very time consuming. Considering the evaluation of one gate demanding a refresh, the run-time will be significant as well as the processing of security parameters.

A suggestion of a nearly FHE scheme based architecture for enabling the evaluation of any function and processing encrypted data is illustrated in Figure 6. In our proposed architecture, the service provider repartitions the processing among the servers to fasten the evaluation process of any function.

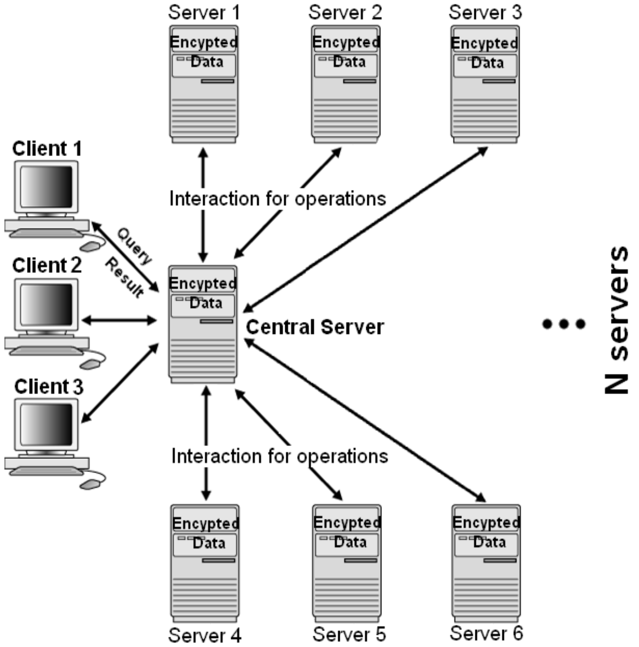


Fig. 5. An architecture of distributed servers for processing encrypted data

In this system, we provide a high leveled architectural scheme through the usage of multiple servers in the computation. This computational system will nearly allow achieving a FHE, and thus a large number of operations including multiplications and additions can be performed. For instance, in Fig. 5, Client 1 requests the results of a given function, let’s say f(x)=ax²+bx+c. In this case the function elements are encrypted and divided into several chunks depending on the number of operations (Multiplication and addition), and will be processed separately on N different servers, equivalent to the number of addition operations Finally the result is sent back to a Central Server in order to be forwarded to Client 1 and then decrypted.

The benefit is that no longer chipertext after encryption unlike the classical method. The keys are easily handled and more security is maintained since is it impossible to read relevant information in distributed systems. In the cloud the N servers consists of hypervisors hosting multiple virtual machines which help improving the response time and augment the number of the involved computational entities in the distributed system.

In this suggestion, we analyze the added value of the distributed systems in processing operations requested by clients. The scheme of homomorphic encryption is dispatched within the servers and this can be practical and help improving the security of the cloud in terms of confidentiality of data and performance.

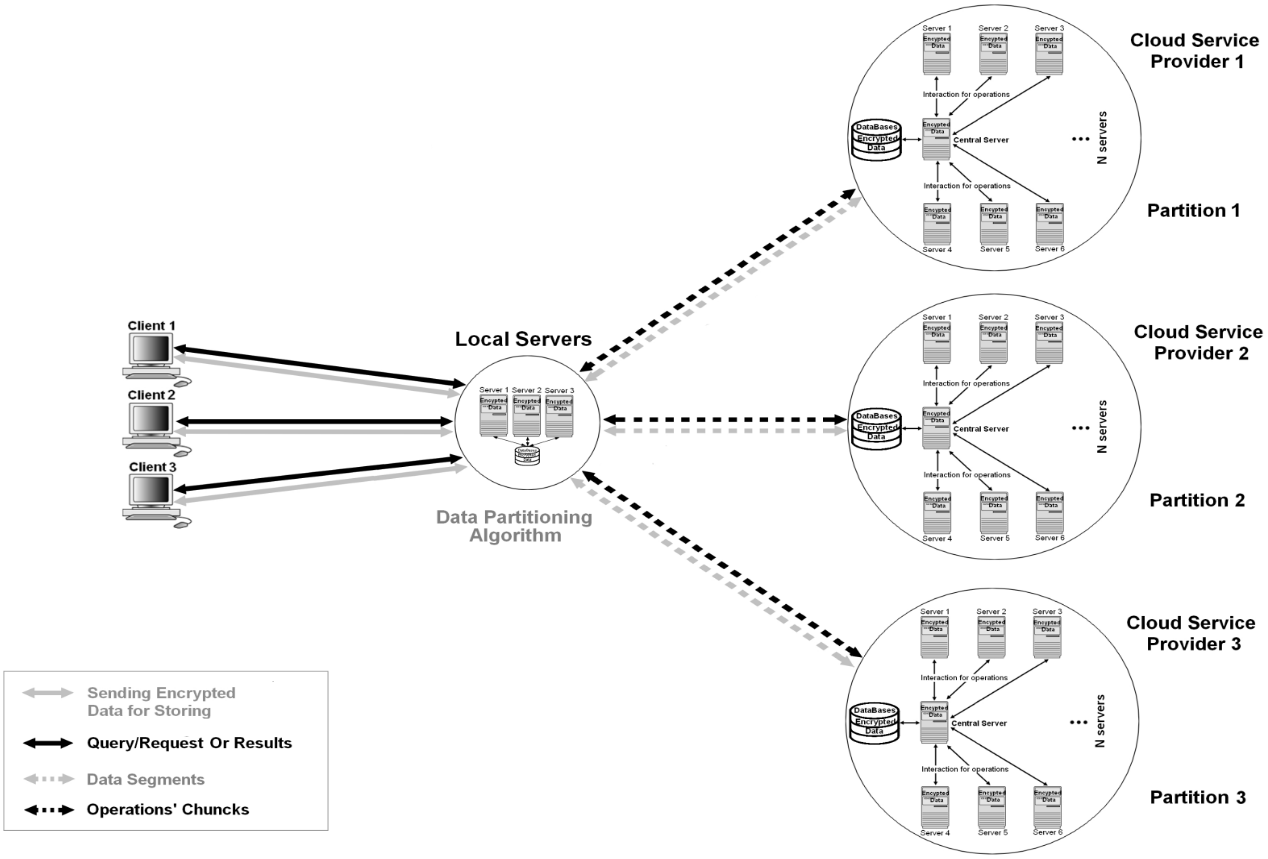


Fig. 6. The proposed architecture to secure data using homomorphic encryption

Another concern that should be considered in our architecture is the confidentiality of the processed data over the distributed systems, which is the main concern of most organization when using third-party hosting. Our approach regarding this matter is the split the stored data among multiple Cloud service providers to decrease the risk of data breaches and increase the parallel processing as well as the number of the servers involved in performing homomorphic encryption. Partitioning and outsourcing the data, applications onto different cloud infrastructures has the advantage of making them ambiguous for third-parties and adversaries, and thus this helps enhancing the confidentiality as well as the privacy.

As the stored encrypted data is repartitioned among a Multi-Cloud Architecture belonging to different Cloud Service Providers (See Fig. 6), Client 1 can perform operations on them and transparently retrieve the intended results. The data is segmented through a Data Partitioning Algorithm (DPA) which allows partitioning, collecting and reconstructing the data. The main operations are chunked into subsets to be handled by the N Clouds/N Servers. The combination of N Clouds and homomorphic encryption using N servers provides an enhanced security strategy which is a safe approach to prevent any potential data breaches even if the data have been already encrypted.

Choosing a trusted CSP requires a Service Level Agreement (SLA), contract negotiation and risk assessments. In most cases it may be logical to believe that a CSP to be trustworthy and handling the clients' sensitive data and applications in a responsible manner.

VI. Conlusion

As Gentry proposed his construction and blueprint in 2009, there has been a huge effort to make FHE more practical. While a lot of progress has been made, unfortunately, we are still some way from truly practical FHE.

Most FHE schemes are based on Gentry's blueprint which consists of first constructing a SHE and then using Gentry's bootstrapping technique to turn it into a FHE scheme. It turns out that bootstrapping is a major bottleneck and that SHE is actually reasonably efficient. So, if we care about practical applications, then it may be worthwhile to explore what exactly we can do with SHE instead.

Distributed systems and multi-could architectures could bring lots of benefits to the application of homomorphic encryption and making it more practical in the case of the security of data and applications.

In a future work, we will focus on the implementation of our proposal and conduct security and performance tests in order to show its practicability.

References

[1] C. Gentry, "A fully homomorphic encryption scheme," Doctoral dissertation, Stanford University, 2009.

[2] R. Rivest, L. Adleman, and M. Dertouzos, "On data banks and privacy homomorphisms," In Foundations of Secure Computation, pages 169-180, 1978.

[3] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," Communications of the ACM, 21(2):120-126, 1978.

[4] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," In 18th Annual Eurocrypt Conference (EUROCRYPT'99) Prague, Czech Republic , volume 1592, 1999.

[5] J. Bringe and al., "An Application of the Goldwasser-Micali Cryptosystem to Biometric Authentication", Springer-Verlag, 2007.

[6] R. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public key cryptosystems," Communications of the ACM, 21(2):120-126, 1978. Computer Science, pages 223-238. Springer, 1999.

[7] T. ElGamal, "A public key cryptosystem and a signature sche based on discrete logarithms," IEEE Transactions on Information Theory, 469-472, 1985.

[8] C. Gentry, "Fully homomorphic encryption using ideal lattices," InSTOC, Vol. 9, pp. 169-178, 2009.

[9] D. Boneh, E. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," In Proceedings of Theory of Cryptography (TCC) '05, LNCS 3378, pages 325-341, 2005.

[10] O. Goldreich, S. Goldwasser, and S. Halevi, "Public-key cryptosystems from lattice reduction problems," In Proceedings of the 17th Annual International Cryptology Conference on Advances in Cryptology, pages 112-131. Springer-Verlag, 1997.

[11] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," Advances in Cryptology Eurocrypt, 1592:223-238, 1999.

[12] S. Goluch, "The development of homomorphic cryptography: From RSA to Gentry's privacy homomorphism" Doctoral dissertation, Vienna university of Technology, 2010.

Nonquadratic Lyapunov Functions for Nonlinear Takagi-Sugeno Discrete Time Uncertain Systems Analysis and Control

Ali Bouyahya
University of Tuins El Manar,
National Engineering School of
Tunis, Laboratory of research in
Automatic Control, BP 37,
Belvédère, 1002
Tunis, Tunisia

Yassine Manai
University of Tuins El Manar,
National Engineering School of
Tunis, Laboratory of research in
Automatic Control, BP 37,
Belvédère, 1002
Tunis Tunisia

Joseph Haggège
University of Tuins El Manar,
Departement of Electrical
Engineering, National Engineering
School of Tunis, Laboratory of
research in Automatic Control, BP
37, Belvédère, 1002
Tunis, Tunisia

***Abstract*—This paper deals with the analysis and design of the state feedback fuzzy controller for a class of discrete time Takagi -Sugeno (T-S) fuzzy uncertain systems. The adopted framework is based on the Lyapunov theory and uses the linear matrix inequality (LMI) formalism. The main goal is to reduce the conservatism of the stabilization conditions using some particular Lyapunov functions. Four nonquadratic Lyapunov Functions are used in this paper. These Lyapunov functions represent an extention from two Lyapunov functions existing in the literature. Their influence in the stabilization region (feasible area of stabilization) is shown through examples, the stabilization conditions of controller for discrete time T-S parametric uncertain systems is demonstrated with the variation of the lyapunov functions between (k, k+1) and (k, k+t) sample times. The controller gain can be obtained via solving several linear matrix inequalities (LMIs). Through the examples and simulations, we demonstrate their uses and their robustness. Comparative study verifies the effectiveness of the proposed methods.**

***Keywords*—Nonquadratic Lyapunov functions; Non-PDC; Linear Matrix Inequality; Parametric Uncertain Systems; Takagi-Sugeno**

I. Introduction

Fuzzy control systems have experienced a big growth of industrial applications in the recent decades, because of their reliability and effectiveness.

In recent years, there has been growing interest in the study of stability and stabilization of Takagi–Sugeno (T–S) fuzzy system[1, 2, 3,4, 5] due to the fact that it provides a general framework to represent a nonlinear plant by using a set of local linear models which are smoothly connected through nonlinear fuzzy membership functions.

Nonlinear systems are difficult to describe. Takagi-Sugeno fuzzy model is a multimodel approach much used to modelise non linear sytems by construction with identification of input-output data [6,7]. The merit of such fuzzy model-based control methodology is that it offers an effective and exact representation of complex nonlinear systems in a compact set of state variables.With the powerful T–S fuzzy model, a natural, simple, and systematic design control approach can be provided to complement other nonlinear control techniques that require special and rather involved knowledge. Nowadays, T–S fuzzy model-based control approaches have been applied successfully in a wide range of applications.

One of the most important issues in the study of T–S fuzzy systems is the stability and stabilization analysis problems [8]. Via various approachs, a great number of stability and stabilization results for T–S fuzzy systems in both the continuous and discrete time have been reported in the literature [9,10].

Two classes of Lyapunov functions are used to analyze these systems: quadratic Lyapunov and nonquadratic Lyapunov functions. The second class of function is less conservative than the first. Many researches have investigated nonquadratic Lyapunov functions [11, 12, 13, 14, 15, 16].

Many works try to reduce the conservatism of quadratic form. Several approaches have been developed to overcome the above mentioned limitations. Piecewise quadratic Lyapunov functions were employed to enrich the set of possible Lyapunov functions used to prove stability [11]. Multiple Lyapunov functions have been paid a lot of attention due to avoiding conservatism of stability and stabilization. Some works try to enrich some properties of the membership functions [17, 18], others introduce decisions variables (slack variables) in order to provide additional degrees of freedom to the LMI problem [19, 20].

For every case, The Lyapunov function used to prove the stability has the most important effect to the results. To leave the quadratic framework, some works have dealt with nonquadratic Lyapunov functions. In this case, some results are available in the continuous and the discrete cases [21],[22],[23]. In the discrete case, new improvements has been developed in [24], by replacing the classical one sample time variation of the Lyapunov function by its variation over several samples (k samples times variations). This condition reduces the conservatism of quadratic form and give a large sets of solutions in terms of linear matrix inequality LMI. The relaxed conditions admitted more freedom in guaranteeing the stability and stabilization of the fuzzy control systems and were found to be very valuable in designing the fuzzy controller, especially when the design problem involves not only stability, but also the other performance requirements such as the speed of response, constraints on control input and output .

In this paper, a new stabilization conditions for discrete time Takagi Sugeno parametric uncertain fuzzy systems with the use of [25,26, 27] new nonquadratic Lyapunov functions are discussed. This condition was reformulated into LMI. [28,29,30,31,32,33,34,35], which can be efficiently solved by using various convex optimization algorithms.

The organization of the paper is as follows. First, T-S fuzzy modeling is discussed. Second, we discuss the proposed approachs to stabilize a T-S fuzzy system in closed loop with the new lyapunov functions. Third, simulation results show the robustness of this approachs and their influence in the stabilization region (feasible area of stabilization). We finish by a conclusion.

II. System Description and Preliminairies

In this section, we describe the concept of the Takagi-Sugeno parametric uncertain system. It's based on the state space representation.

Consider the discrete time fuzzy model T-S parametric uncertain systems for nonlinear systems given as follows.

If $z_1(t)$ is M_{i1} and$\ldots$ *and* $z_p(t)$ *is* M_{ip} then

$$\begin{cases} x(k+1) = (A_i + \Delta A_i)x(k) + (B_i + \Delta B_i)u(k) \\ y(k+1) = (C_i + \Delta C_i)x(k) \end{cases} \quad (1)$$

$$i = 1......r$$

Where $M_{ij}(i = 1,2....r,\ j = 1,2.....p)$ is the fuzzy set and r is the number of model rules, $x(k) \in \Re^n$ is the states vector ; $u(k) \in \Re^m$ is the input vector; $A_i \in \Re^{n\times n}$,the states matrix, $B_i \in \Re^{n\times m}$ the control matrix and $z_1(k),......, z_p(k)$ are known premise variables.

The T-S fuzzy model is written under the following form:

$$\begin{cases} x(k+1) = \dfrac{\sum_{i=1}^{r} w_i(z(k))\left((A_i + \Delta A_i)x(k) + (B_i + \Delta B_i)u(k)\right)}{\sum_{i=1}^{r} w_i(z(k))} \\ y(k) = \dfrac{\sum_{i=1}^{r} w_i(z(k))(C_i + \Delta C_i)\,x(k)}{\sum_{i=1}^{r} w_i(z(k))} \end{cases} \quad (2)$$

r : is the number of model rules.

With

$$\begin{cases} w_i(z(k)) = \prod_{j=1}^{r} M_{ij}(z_j(k)) \\ h_i(z(k)) = \dfrac{w_i(z(k))}{\sum_{i=1}^{r} w_i(z(k))} \qquad i = 1, 2, \ldots, r \end{cases} \quad (3)$$

The term $M_{ij}(z_j(k))$ is the membership degree of $z_j(k)$ in M_{ij} .

Since

$$\begin{cases} \sum_{i=1}^{r} w_i(z(k)) > 0 \\ w_i(z(k)) \geq 0 \qquad i = 1......r \end{cases} \quad (4)$$

we have

$$\begin{cases} 0 < h_i(z(k)) < 1 \\ \sum_{i=1}^{r} h_i(z(k)) = 1 \end{cases} \quad (5)$$

The final output can be written under the following form

$$\begin{cases} x(k+1) = \sum_{i=1}^{r} h_i(z(k))(A_i + \Delta A_i)x(k) + (B_i + \Delta B_i)u(k) \\ \qquad = (A_{z(k)} + \Delta A_{z(k)})x(k) + (B_{z(k)} + \Delta B_{z(k)})u(k) \\ y(k) = \sum_{i=1}^{r} h_i(z(k))(C_i + \Delta C_i)x(k) = (C_{z(k)} + \Delta C_{z(k)})x(k) \end{cases} \quad (6)$$

$\Delta A_i, \Delta B_i$ represents parametric uncertainties matrices in the state space representation. These uncertainties matrices are written under the following form.

$$\begin{cases} \Delta A = H_a F E_a \\ FF^T \leq 1 \end{cases}, \begin{cases} \Delta B = H_b F E_b \\ FF^T \leq 1 \end{cases}, \begin{cases} \Delta C = H_c F E_c \\ FF^T \leq 1 \end{cases} \quad (7)$$

With $H_a, H_b, H_c, E_a, E_b, E_c$ are constants matrices.

Lemma 1, 2 and 3 present the techniques and powerful tools used through the development of the next theorems.

Lemma 1 (Schur Complement) [36,37]

Consider A,G,L,P and Q matrices with appropriates dimensions. The next properties are equivalent:

$$1.\ A^T P A - Q < 0\ ,\ P > 0 \quad (8)$$

$$2. \begin{bmatrix} -Q & A^T P \\ PA & -P \end{bmatrix} < 0 \quad (9)$$

$$3. \exists\, G \begin{bmatrix} -Q & A^T G \\ G^T A & -G - G^T + P \end{bmatrix} < 0,\ P > 0 \quad (10)$$

$$4.\exists\ G, L \begin{bmatrix} -Q + A^T L^T + LA & -L + A^T G \\ -L^T + G^T A & -G - G^T + P \end{bmatrix} < 0,\ P > 0 \tag{11}$$

Lemma 2 [38]

Relaxaion : Whatever the choise of the Lyapunov function, the analysis of the stabilization leads us to the inequality (12) with multiple sum

$$\begin{aligned} &\sum_{i_0=1}^{r}\sum_{i_{k-1}=1}^{r}\sum_{i_0=1}^{r}\sum_{i_{k-1}=1}^{r} h_{i_0}\left(z(k)\right)....h_{i_{k-1}}\left(z(2k-1)\right)h_{j_0}\left(z(k)\right)\times ...... \\ &....\times h_{j_{k-1}}\left(z(2k-1)\right)\Upsilon_{i_0,...i_{k-1},j_0,...j_{k-1}} < 0 \end{aligned} \tag{12}$$

Consider $\Upsilon_{i_0,...i_{k-1},j_0,...j_{k-1}} \Box\ \tilde{\Upsilon} + \tilde{\Upsilon}^{(0)}_{i_0,j_0} + ... + \tilde{\Upsilon}^{(k-1)}_{i_{k-1},j_{k-1}}$ matrices and h_i functions having the convex sum properties.

The inequality (12) is verified if the next $\left(0.5r\left(r+1\right)\right)^k$ conditions are verified

$$\begin{aligned} &\forall\left(i_0, j_0\right),.....,\left(i_{k-1}, j_{k-1}\right) \in \{1,2,...,r\} \\ &\Upsilon_{i_0 i_1,...i_{k-1},j_0 j_1,...j_{k-1}} + \Upsilon_{j_0 j_1,...j_{k-1},i_0 i_1,...i_{k-1}} < 0 \end{aligned} \tag{13}$$

where

$i_0 \le j_0,.....,i_{k-1} \le j_{k-1}$

Lemma 3 [39]

Consider X and $Y, Q = Q^T > 0$ matrices of appropriate dimensions, the following inequality is verified

$$XY^T + YX^T \le XQX^T + YQ^{-1}Y^T \tag{14}$$

The use of these lemmas will be shown in the next section.

III. Stabilization Analysis

This section recalls the technique of the stabilization analysis of discrete T-S model based on a nonquadratic Lyapunov function. In the discrete case, we consider the variation of the Lyapunov function between two sample time. If the final equation of this variation is negative, we obtain a sufficient condition of the T-S stabilization with the state feedback controller.

Consider the discrete time fuzzy Takagi-Sugeno system under the following form.

$$\begin{cases} x(k+1) = \sum_{i=1}^{r} h_i\left(z(k)\right)\left(A_i x(k) + B_i u(k)\right) \\ \qquad\quad = \left(A_{z(k)}x(k) + B_{z(k)}u(k)\right) \\ y(k) = \sum_{i=1}^{r} h_i\left(z(k)\right)(C_i)x(k) = C_{z(k)}x(k) \end{cases} \tag{15}$$

The non-PDC control law is described by the following equation:

$$u(k) = -\sum_{i=1}^{r} F_i G^{-1} x\left(k\right) \tag{16}$$

The Lyapunov function used in [40] expressed in equation (17)

$$\begin{aligned} V(x(k)) &= x^T(k)\left(\sum_{i=1}^{r} h_i\left(z\left(k\right)\right)G_i\right)^{-T} \\ &\times\left(\sum_{i=1}^{r} h_i\left(z\left(k\right)\right)\left(P_i + \mu R\right)\right)\times\left(\sum_{i=1}^{r} h_i\left(z\left(k\right)\right)G_i\right)^{-1} x(k) \\ &= x^T(k)G_{z(k)}^{-T}(P_{z(k)})G_{z(k)}{}^{-1}x(k) \end{aligned} \tag{17}$$

The final equation of the Lyapunov function variation obtained by [40] which represent the stabilization condition of discrete time T-S systems is written under the following form.

$$\Upsilon_{i_0 i_1,...i_{k-1},j_0 j_1,...j_{k-1}} =$$

$$\begin{bmatrix} -P & (*) & \ddots & 0 \\ A_{i_0}G - B_{i_0}F_{j_0} & -G - G^T & \ddots & \ddots \\ \ddots & \ddots & \ddots & (*) \\ 0 & \ddots & A_{i_{k-1}}G - B_{i_{j-1}}F_{j_{k-1}} & -G - G^T + P \end{bmatrix} < 0 \tag{18}$$

So [40] propose the following theorem:

Theorem [40]

Consider the discrete Takagi-Sugeno (15), the control law (16) and the $\Upsilon_{i_0 i_1,...i_{k-1},j_0 j_1,...j_{k-1}}$ defined in (18). If it exist a definite positive matrix P and matrices G, F_i , $i = \{1.....r\}$ such that the conditions (12) and (13) of lemma 2 are verified the system is globally asymptotic stable in closed loop.

We propose a new Lyapunov function based on the Lyapunov function in equation (17), by multiplying the Lyapunov matrices $P_{z(k)}$ by a scalar $\alpha > 0$. So the new form of the Lyapunov function is written under the following form in equation (19).

$$V(x(k)) = x^T(k)G_{z(k)}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1}x(k) \tag{19}$$

and the non-PDC control law is written under the following form in equation (20).

$$u(k) = -\sum_{i=1}^{r} F_i G_i^{-1} x\left(k\right) \tag{20}$$

The variation of the Lyapunov function between *k* and *k+t* sample times is given by the next equation (21)

$$\begin{aligned} \Delta_k V\left(x\left(k\right)\right) &= x\left(k+t\right)^T G_{z(k)}{}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1}x\left(k+t\right) \\ &\quad - x\left(k\right)^T G_{z(k)}{}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1}x\left(k\right) \end{aligned} \tag{21}$$

The final output $x\left(k+1\right)$ is written between k and (k+t) samples under the next form.

$$x\left(k+1\right) = (A_{z(k)} - B_{z(k)}F_{z(k)}G_{z(k)}{}^{-1})x\left(k\right) \quad with \quad A_{z(k)} = \sum_{i=1}^{r} h_i\left(z(k)\right)A_i$$

$$x\left(k+2\right) = (A_{z(k+1)} - B_{z(k+1)}F_{z(k+1)}G_{z(k)}{}^{-1})(A_{z(k)} - B_{z(k)}F_{z(k)}G_{z(k)}{}^{-1})x\left(k\right)$$

.

.

.

$$x\left(k+t\right) = (A_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)}G_{z(k)}{}^{-1})\times.......\times(A_{z(k)} - B_{z(k)}F_{z(k)}G_{z(k)}{}^{-1})x\left(k\right)$$

The variation of the Lyapunov function for discrete system should be negative $\Delta_k V\left(x\left(k\right)\right) < 0$. This equation is equivalent to

$$x\left(k\right)^T \begin{pmatrix} (*)G_{z(k)}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1}(A_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)}G_{z(k)}{}^{-1})\times.... \\ ...\times(A_{z(k)} - B_{z(k)}F_{z(k)}G_{z(k)}{}^{-1}) - G_{z(k)}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1} \end{pmatrix} x\left(k\right) < 0 \quad (22)$$

The equation (22) is equivalent to equation (23)

$$\begin{pmatrix} (*)G_{z(k)}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1}(A_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)}G_{z(k+1)}{}^{-1})\times.... \\ ...\times(A_{z(k)} - B_{z(k)}F_{z(k)}G_{z(k)}{}^{-1}) - G_{z(k)}^{-T}(\alpha P_{z(k)})G_{z(k)}{}^{-1} \end{pmatrix} < 0 \quad (23)$$

Consider the next modification

$$A_{z(k+t)} - B_{z(k+t)}F_{z(k+t)}G_{z(k+t)}{}^{-1} = \left(A_{z(k+t)}G_{z(k+t)} - B_{z(k+t)}F_{z(k+t)}\right)G_{z(k+t)}^{-1} \quad (24)$$

Using the congruence with the full rank matrix G, we obtain

$$(*)G_z^{-T}(\alpha P_{z(k)})\left[G_z^{-1}(A_{z(k+t-1)}G_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)})\right]\times.....$$
$$..\times\left[G_{z(k)}^{-1}(A_{z(k)}G_{z(k)} - B_{z(k)}F_{z(k)})\right] - (\alpha P_{z(k)}) < 0$$
(25)

So the variation of the lyapunov function $\Delta_k V\left(x\left(k\right)\right) < 0$ holds if the equation (25) is negative. The use of the Schur Complement (Lemma 1) with the equation (25) give the next equation

$$\begin{bmatrix} -\alpha P_{z(k)} & * \\ \Phi\left[G_{z(k)}^{-T}\left(A_{z(k)}G_{z(k)} - B_{z(k)}F_{z(k)}\right)\right] & \begin{pmatrix} -\Phi - \Phi^T + (*)G_{z(k)}^{-T}\alpha P_{z(k)}G_{z(k)}\times... \\ ..\times\left(A_{z(k+1)}G_{z(k+1)} - B_{z(k+1)}F_{z(k+1)}\right) \end{pmatrix} \end{bmatrix} < 0 \quad (26)$$

For each iteration i={1….r}

Let’s consider the following inequality:

$$\begin{bmatrix} -\alpha P_{z(k)} & * \\ \Phi\left[G_{z(k)}^{-T}\left(A_{z(k)}G_{z(k)} - B_{z(k)}F_{z(k)}\right)\right] & -\Phi - \Phi^T + \Gamma_i \end{bmatrix} < 0 \quad (27)$$

with

$$\Gamma_i = (*)G_{z(k)}^{-T}\alpha P_{z(k)}G_{z(k)}\left(A_{z(k+t-1)}G_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)}\right)$$
$$\times.....\times\left(A_{z(k+i)}G - B_{z(k+1)}F_{z(k+i)}\right)$$

The application of the lemma 1 with equation (27) with $\Phi = G$ give the next inequality

$$\begin{bmatrix} -\alpha P_{i_0} & (*) & 0 \\ A_{i_0}G_{j_0} - B_{i_0}F_{j_0} & -G_{i_0} - G_{i_0}^{T} & (*) \\ 0 & A_{i_{k-1}}G - B_{i_{k-1}}F_{j_{k-1}} & -G - G^T + \Gamma_2 \end{bmatrix} < 0 \quad (28)$$

Recursively by the use of Schur Complement we obtain the inequality (29).

$$\begin{bmatrix} -\alpha P_{z(k)} & (*) & \ddots & 0 \\ A_{z(k)}G - B_{z(k)}F_{z(k)} & -G_{z(k)} - G_{z(k)}^T & \ddots & \ddots \\ \ddots & \ddots & \ddots & (*) \\ 0 & \ddots & A_{z(k+t-1)}G_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)} & -G_{z(k+t-1)} - G_{z(k+t-1)}^T + \alpha P_{z(k+t-1)} \end{bmatrix} < 0$$

(29)

The use of the lemma 2 with the equation (29), give the final condition of discrete time T-S systems stabilization. This condition should be negative.

$$\Upsilon_{i_0 i_1,...i_{k-1}, j_0 j_1,...j_{k-1}} =$$
$$\begin{bmatrix} -\alpha P_{i_0} & (*) & \ddots & 0 \\ A_{i_0}G_{j_0} - B_{i_0}F_{j_0} & -G_{i_0} - G_{i_0}^T & \ddots & \ddots \\ \ddots & \ddots & \ddots & (*) \\ 0 & \ddots & A_{i_{k-1}}G_{j_{k-1}} - B_{i_{k-1}}F_{j_{k-1}} & -G_{i_{k-1}} - G_{i_{k-1}}^T + \alpha P_{i_{k-1}} \end{bmatrix} < 0 \quad (30)$$

Therefore we state the following theorem for the discrete time Takagi-Sugeno fuzzy systems.

Theorem 1

Consider the discrete time Takagi-Sugeno (15), the control law (20) and the $\Upsilon_{i_0 i_1,...i_{k-1}, j_0 j_1,...j_{k-1}}$ defined in (30). If exist a definite positive matrices P_i and matrices G_i, F_i , $i = \{1.....r\}$ and $\alpha > 0$ such that the conditions (12) and (13) of lemma 2 are verified the discrete time T-S system is globally asymptotic stable in closed loop.

These two theorems represent sufficient conditions of the discrete time T-S stabilization with state feedback with k sample times variation of the Lyapunov function. In the next section, we present the analysis of the stabilization of the discrete time T-S parametric uncertain systems.

IV. Parametric Uncertain Systems Stabilization Analysis

A. New Lyapunov Function: First approach

In the next, we treat the case of the discrete time Takagi-Sugeno parametric uncertain systems.

Consider the uncertain system described in equation (6)

In that case, the equation (30) becomes.

$$\begin{bmatrix} -\alpha P_{z(k)} & (*) & \ddots & 0 \\ \bar{A}_{z(k)}G_{z(k)} - \bar{B}_{z(k)}F_{z(k)} & -G_{z(k)} - G_{z(k)}{}^{T} & \ddots & \ddots \\ \ddots & \ddots & \ddots & (*) \\ 0 & \ddots & \bar{A}_{z(k+t-1)}G_{z(k)} - \bar{B}_{z(k+t-1)}F_{z(k+t-1)} & -G_{z(k)} - G_{z(k)}{}^{T} + \alpha P_{z(k)} \end{bmatrix} < 0$$

$$\bar{A}_{z(k)} = A_{z(k)} + \Delta A_{z(k)} \ , \ \bar{B}_{z(k)} = B_{z(k)} + \Delta B_{z(k)}$$

(31)

For the uncertainties $\Delta A_{z(k)}, \Delta B_{z(k)}$, the term $\bar{A}_{z(k)}G - \bar{B}_{z(k)}F_{z(k)}$ is transformed in the following form by introducing two scalars $\tau_0 > 0, \ \mu_0 > 0$. The use of lemma 3 on uncertainties $\Delta A_{z(k)}, \Delta B_{z(k)}$ gives the next two inequalities:

$$\begin{bmatrix} 0 \\ H_a \Delta A_z \end{bmatrix} \begin{bmatrix} E_{az}G_z & 0 \end{bmatrix} + (*) \le \begin{bmatrix} \tau_0{}^{-1} G_z^T E_{az}^T E_{az} G_z & 0 \\ 0 & \tau_0 H_a \Delta A_z \left(\Delta A_z\right)^T H_a^T \end{bmatrix}$$

(32)

$$\begin{bmatrix} 0 \\ H_b \Delta B_z \end{bmatrix} \begin{bmatrix} E_{bz}F_z & 0 \end{bmatrix} + (*) \le \begin{bmatrix} \mu_0{}^{-1} F_z^T E_{bz}^T E_{bz} F_z & 0 \\ 0 & \mu_0 H_b \Delta B_z \left(\Delta B_z\right)^T H_b^T \end{bmatrix}$$

(33)

Taking in consideration the two inequality (32) and (33), the equation (31) become equation (34)

$$\begin{bmatrix} -P_{z(k)} + \Omega_0^1 & (*) & \ddots & 0 \\ A_{z(k)}G_{z(k)} - A_{z(k)}F_{z(k)} & -G_{z(k)} - G_{z(k)}^T + \Omega_0^2 & \ddots & \ddots \\ \ddots & \ddots & \ddots & (*) \\ 0 & \ddots & \Xi & \Theta \end{bmatrix} < 0$$

(34)

With

$$\begin{cases} \Theta = -G_{k-1} - G_{k-1}^T + P_{k-1} + \Omega_{k-1}^2 \\ \Xi = A_{z(k+t-1)}G_{z(k+t-1)} - B_{z(k+t-1)}F_{z(k+t-1)} \\ \Omega_0^i = \tau_i{}^{-1} G_z^T E_{az(k+i)}^T E_{az(k+i)} G_z + \mu_i{}^{-1} F_z^T E_{bz(k+i)}^T E_{bz(k+i)} F_z \end{cases}$$

The use of Schur Complement (lemma1) give the next equation (35) which represents the final condition of stabilization of T-S parametric uncertain systems with the use of the Lyapunov function (19) and the control law (20).

$$\left[\begin{matrix} -\alpha P_{z(k)} & (*) & (*) & (*) & 0 \\ E_{bz(k)}F_{z(k)} & -\mu_0 I & 0 & 0 & \ddots \\ E_{az(k)}G_{z(k)} & 0 & -\tau_0 I & 0 & \ddots \\ A_{z(k)}G_{z(k)} - B_{z(k)}F_{z(k)} & 0 & 0 & -G_{z(k)}^T - G_{z(k)} + \Omega_0^2 & \ddots \\ 0 & & & \ddots & \ddots \\ \vdots & & & & \ddots \\ \vdots & & & 0 & \ddots \\ \vdots & & & 0 & \ddots \\ 0 & \cdots & \cdots & 0 & \ddots \end{matrix}\right.$$

$$\left.\begin{matrix} \cdots & \cdots & \cdots & 0 \\ \ddots & \ddots & \ddots & \vdots \\ & & & \\ \ddots & & & \vdots \\ & 0 & 0 & 0 \\ -G_{z(k)}^T - G_{z(k)} + \Omega_{k-2}^2 & (*) & (*) & (*) \\ E_{bz(k)}F_{z(k)} & -\mu_{k-1} I & 0 & 0 \\ E_{az(k)}G_{z(k)} & 0 & -\tau_{k-1} I & 0 \\ A_{z(k)}G_{z(k)} - B_{z(k)}F_{z(k)} & 0 & 0 & -G_{z(k)}^T - G_{z(k)} + \Omega_{k-1}^2 + \alpha P_{z(k)} \end{matrix}\right]$$

(35)

with

$$\Omega_i^2 = \tau_i H_a H_a^T + \mu_i H_b H_b^T \qquad (36)$$

After using lemma 2 the equation (35) become (37)

$$\Upsilon_{i_0 i_1, \ldots i_{k-1}, j_0 j_1, \ldots j_{k-1}} =$$

$$\left[\begin{matrix} -\alpha P_{i_0} & (*) & (*) & (*) & 0 \\ E_{bi_0} F_{j_0} & -\mu_0 I & 0 & 0 & \ddots \\ E_{ai_0} G_{i_0} & 0 & -\tau_0 I & 0 & \ddots \\ A_{i_0} G_{j_0} - B_{i_0} F_{j_0} & 0 & 0 & -G_{i_0}^T - G_{i_0} + \Omega_0^2 & \ddots \\ 0 & & & \ddots & \ddots \\ \vdots & & & & \ddots \\ \vdots & & & 0 & \ddots \\ \vdots & & & 0 & \ddots \\ 0 & \cdots & \cdots & 0 & \ddots \end{matrix}\right.$$

$$\left.\begin{matrix} \cdots & \cdots & \cdots & 0 \\ \ddots & \ddots & \ddots & \vdots \\ & & & \\ \ddots & & & \vdots \\ & 0 & 0 & 0 \\ -G_{i_{k-1}}^T - G_{i_{k-1}} + \Omega_{k-2}^2 & (*) & (*) & (*) \\ E_{bi_{k-1}} F_{j_{k-1}} & -\mu_{k-1} I & 0 & 0 \\ E_{ai_{k-1}} G_{j_{k-1}} & 0 & -\tau_{k-1} I & 0 \\ A_{i_{k-1}} G_{j_{k-1}} - B_{i_{k-1}} F_{j_{k-1}} & 0 & 0 & -G_{k-1}^T - G_{k-1} + \Omega_{k-1}^2 + \alpha P_{i_0} \end{matrix}\right] < 0$$

(37)

So we state the next theorem for the stabilization of the discrete time T-S parametric uncertain systems.

Theorem 2

Consider the discrete time uncertain Takagi-Sugeno system (6), the control law (20) and the $\Upsilon_{i_0 i_1, \ldots i_{k-1}, j_0 j_1, \ldots j_{k-1}}$ defined in (37). If exist a definite positive matrices P_i , matrices G_i, F_i , $i = \{1.....r\}$ and positives scalars $\tau_i, \ \mu_i$ and $\alpha > 0$ such that the conditions of lemma 2 are verified the system is globally asymptotic stable in closed loop.

The next work deals with the addition of more variables in the equation $\Upsilon_{i_0 i_1, \ldots i_{k-1}, j_0 j_1, \ldots j_{k-1}}$ to give a large field of solutions.

In this case a condition of stabilization is developed based on new Lyapunov functions and a new non-PDC control law (20).

B. *New Lyapunov Function : Second approach*

Consider the new non quadratic lyapunov function in equation (38) and the non-PDC control law in equation (20). In this new function, we associate for each Lyapunov matrices $P_{z(k)}$ a

scalar α .

$$V(x(k)) = x^T(k) G_{z(k)}^{-T} (\alpha_i P_{z(k)}) G_{z(k)}{}^{-1} x(k) \tag{38}$$

where $\alpha_i > 0$ with $i = \{1,..r\}$

Consider the same transformations and lemmas used to obtain equations (30), (37), theorems 1 and 2. The new form of equation for the stabilization of discrete time T-S fuzzy parametric uncertain systems with the use of the Lyapunov function (38) is under the following form in equation (39).

$$\Upsilon_{i_0 i_1,\ldots i_{k-1}, j_0 j_1,\ldots j_{k-1}} =$$

$$\left[\begin{array}{ccccc}
-\alpha_{i_0} P_{i_0} & (*) & (*) & (*) & 0 \\
E_{bi_0} F_{j_0} & -\mu_0 I & 0 & 0 & \ddots \\
E_{ai_0} G_{j_0} & 0 & -\tau_0 I & 0 & \ddots \\
A_{i_0} G_{j_0} - B_{i_0} F_{j_0} & 0 & 0 & -G_{j_0}^T - G_{j_0} + \Omega_0^2 & \ddots \\
0 & & & \ddots & \ddots \\
\vdots & & & & \ddots \\
\vdots & & & 0 & \ddots \\
\vdots & & & 0 & \ddots \\
0 & \cdots & \cdots & 0 & \ddots
\end{array}\right.$$

$$\left.\begin{array}{cccc}
\cdots & \cdots & \cdots & 0 \\
\ddots & \ddots & \ddots & \vdots \\
\ddots & & & \vdots \\
 & 0 & 0 & 0 \\
-G_{j_{k-1}}^T - G_{j_{k-1}} + \Omega_{k-2}^2 & (*) & (*) & (*) \\
E_{bi_{k-1}} F_{j_{k-1}} & -\mu_{k-1} I & 0 & 0 \\
E_{ai_{k-1}} G_{j_{k-1}} & 0 & -\tau_{k-1} I & 0 \\
A_{i_{k-1}} G_{j_{k-1}} - B_{i_{k-1}} F_{j_{k-1}} & 0 & 0 & -G_{j_{k-1}}^T - G_{j_{k-1}} + \Omega_{k-1}^2 + \alpha_{i_{k-1}} P_{i_{k-1}}
\end{array}\right] < 0 \tag{39}$$

With

$$\begin{cases} \Omega_{k-1}^2 = \tau_{k-1} H_a H_a^T + \mu_{k-1} H_b H_b^T \\ \Omega_{k-2}^2 = \tau_{k-2} H_a H_a^T + \mu_{k-2} H_b H_b^T \end{cases} \tag{40}$$

Using the equation (39) and the Lyapunov function (38) and the non-PDC controller (20), we propose the next theorem for the stabilization of the T-S parametric uncertain systems.

Theorem 3

Consider the discrete uncertain Takagi-Sugeno system (6), the control law (20) and the $\Upsilon_{i_0 i_1,\ldots i_{k-1}, j_0 j_1,\ldots j_{k-1}}$ defined in (39). If it exists a definite positive matrices P_i , matrices G_i, F_i , $i = \{1.....r\}$ and positives scalars τ_i, μ_i and positives scalars $\alpha_i > 0$ such that conditions of lemma 2 are verified, the system is globally asymptotic stable in the closed loop.

C. *New Lyapunov Function : Third approach*

The third Lyapunov function proposed in this paper represents an extention from the first Lyapunov function and the next Lyapunov function described bellow

The following Lyapunov function is used by [16,17].

$$V(x(k)) = x^T(k) \left(\sum_{i=1}^{r} h_i\left(z(k)\right) G_i \right)^{-T} \left(\sum_{i=1}^{r} h_i\left(z(k)\right)\left(P_i + \mu R\right) \right) \times \left(\sum_{i=1}^{r} h_i\left(z(k)\right) G_i \right)^{-1} x(k) \tag{41}$$

Which P_z is symmetric and definite positive matrix, and G_z is full rank matrix. The nonlinearities are expressed by the terms $h_i\left(z(k)\right) \geq 0$ with the convex sum property $\sum_{i=1}^{r} h_i\left(z(k)\right) = 1$.

The Lyapunov function used by [13, 24], is written under the following form.

$$V(x(t)) = \sum_{k=1}^{r} h_k(z(t)) V_k(x(t)) \tag{42}$$

$$V_k(x(t)) = x^T(t)(P_k + \mu R)x(t) \tag{43}$$

So the third proposed Lyapunov function is written under the following form in equation (44)

$$V(x(k)) = x^T(k) G_z^{-T} (\alpha P_z + \lambda R) G_z^{-1} x(k) \tag{44}$$

where $\alpha > 0$ and $0 \leq \lambda \leq 1$

The new form of equation for the stabilization of discrete time T-S fuzzy parametric uncertain systems with the use of the Lyapunov function (44) is under the following form in equation (45).

$$\Upsilon_{i_0 i_1,\ldots i_{k-1}, j_0 j_1,\ldots j_{k-1}, R} =$$

$$\left[\begin{array}{ccccc}
-\alpha P_i + \mu R & (*) & (*) & (*) & 0 \\
E_{bi_0} F_{j_0} & -\mu_0 I & 0 & 0 & \ddots \\
E_{ai_0} G_{j_0} & 0 & -\tau_0 I & 0 & \ddots \\
A_{i_0} G_{j_0} - B_{i_0} F_{j_0} & 0 & 0 & -G_{j_0}^T - G_{j_0} + \Omega_0^2 & \ddots \\
0 & & & \ddots & \ddots \\
\vdots & & & & \ddots \\
\vdots & & & 0 & \ddots \\
\vdots & & & 0 & \ddots \\
0 & \cdots & \cdots & 0 & \ddots
\end{array}\right.$$

$$\left.\begin{array}{cccc}
\cdots & \cdots & \cdots & 0 \\
\ddots & \ddots & \ddots & \vdots \\
\ddots & & & \vdots \\
 & 0 & 0 & 0 \\
-G_{j_{k-1}}^T - G_{j_{k-1}} + \Omega_{k-2}^2 & (*) & (*) & (*) \\
E_{bi_{k-1}} F_{j_{k-1}} & -\mu_{k-1} I & 0 & 0 \\
E_{ai_{k-1}} G_{j_{k-1}} & 0 & -\tau_{k-1} I & 0 \\
A_{i_{k-1}} G - B_{i_{k-1}} F_{j_{k-1}} & 0 & 0 & -G_{j_{k-1}}^T - G_{j_{k-1}} + \Omega_{k-1}^2 + \alpha P_i + \mu R
\end{array}\right] < 0$$

(45)

With the equation (45) and the Lyapunov function (44) and the non-PDC controller (20), we propose the next theorem for the stabilization of the T-S parametric uncertain systems.

Theorem 4

Consider the discrete uncertain Takagi-Sugeno system (6), the control law (20) and the $\Upsilon_{i_0 i_1,\dots i_{k-1},j_0 j_1,\dots j_{k-1},R}$ defined in (45). If exist a definite positive matrices P_i , matrices R, G_i, F_i , $i = \{1.....r\}$ and positives scalars τ_i, μ_i , positive scalar $\alpha > 0$ and $0 \le \lambda \le 1$ such that the conditions of lemma 2 are verified the system is globally asymptotic stable in closed loop.

In the next section, we add more values to the LMI in order to demonstrate their influence in stabilization region by affecting to each lyapunov matrices P_i , a positive scalar α_i .

D. New Lyapunov Function : Fourth approach

The fourth Lyapunov function used in this paper is written under the following form in equation (46)

$$V(x(k)) = x^T(k) G_z^{-T} (\alpha_i P_z + \lambda R) G_z^{-1} x(k) \qquad (46)$$

Under this Lyapunov function, the new condition of stabilization obtained in the next equation.

$$\Upsilon_{i_0 i_1,\dots i_{k-1},j_0 j_1,\dots j_{k-1},R} =$$

$$\left[\begin{array}{ccccc}
-\alpha_i P_i + \mu R & (*) & (*) & (*) & 0 \\
E_{bi_0} F_{j_0} & -\mu_0 I & 0 & 0 & \ddots \\
E_{ai_0} G & 0 & -\tau_0 I & 0 & \ddots \\
A_{i_0} G - B_{i_0} F_{j_0} & 0 & 0 & -G^T - G + \Omega_0^2 & \ddots \\
0 & & & \ddots & \ddots \\
\vdots & & & & \ddots \\
\vdots & & & 0 & \ddots \\
\vdots & & & 0 & \ddots \\
0 & \cdots & \cdots & 0 & \ddots
\end{array}\right.$$

$$\left.\begin{array}{cccc}
\cdots & \cdots & \cdots & 0 \\
\ddots & \ddots & \ddots & \vdots \\
\ddots & & & \vdots \\
 & 0 & 0 & 0 \\
-G^T - G + \Omega_{k-2}^2 & (*) & (*) & (*) \\
E_{bi_{k-1}} F_{j_{k-1}} & -\mu_{k-1} I & 0 & 0 \\
E_{ai_{k-1}} G & 0 & -\tau_{k-1} I & 0 \\
A_{i_{k-1}} G - B_{i_{k-1}} F_{j_{k-1}} & 0 & 0 & -G^T - G + \Omega_{k-1}^2 + \alpha_i P_i + \mu R
\end{array}\right] < 0$$

(47)

We state the following theorem for the stabilization of the discrete time T-S fuzzy parametric uncertain systems.

Theorem 5

Consider the discrete uncertain Takagi-Sugeno system (6), the control law (20) and the $\Upsilon_{i_0 i_1,\dots i_{k-1},j_0 j_1,\dots j_{k-1},R}$ defined in (47). If it exist a definite positive set of matrices P_i , matrices R, G_i, F_i , $i = \{1.....r\}$ and positives scalars τ_i, μ_i , $\alpha_i > 0$ and $0 \le \lambda \le 1$ such that the conditions of lemma 2 are verified, then the system is globally asymptotic stable in closed loop.

Four Lyapunov functions were proposed in this paper. They represent a direct extension from two other function in the literature. In the next section, we present their robustness by showing their influence on the stabilization region.

V. Simulation and Validation of Results

Consider TS discrete uncertain system with unstable open loop models. This system is modeled with two subsystems, so we have $r = 2$.

$$\mathrm{A}_1 = \begin{pmatrix} 0.2 & 0.27 & 0.1 \\ 0.4 & -0.6 & -0.3 \\ 0.1 & 0.5 & 0.8 \end{pmatrix} \quad , \quad \mathrm{A}_2 = \begin{pmatrix} -0.3 & 0.5 & 0.1 \\ 0.2 & 0.1 & -0.9 \\ 0.1 & -0.4 & 0.7 \end{pmatrix}$$

$$\mathrm{B}_1 = \begin{pmatrix} 1.1 \\ 0.8 \\ -0.9 \end{pmatrix}, \quad \mathrm{B}_2 = \begin{pmatrix} 1 \\ 0.5 \\ 0.73 \end{pmatrix} \quad \mathrm{H}_a = \begin{pmatrix} 1 \\ 0.45 \\ 0.45 \end{pmatrix}$$

$$\mathrm{H}_b = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathrm{E}_a = \begin{pmatrix} 0.2 & 1 & -0.4 \end{pmatrix}, \quad \mathrm{E}_b = \begin{pmatrix} 1 \\ 0 \\ 0.19 \end{pmatrix}$$

For the simulation the membership functions are choosen as follows:

$$h_1(z(k)) = \frac{1}{1 + (0.9 x_1(k))} \ , h_2(z(k)) = 1 - h_1(z(k))$$

With the application of theorem 2, with α=0.6 the results of LMI gives definite positive matrices P_1, P_2 and matrices $\mathrm{G}_1, \mathrm{G}_2, \mathrm{F}_1$ and F_2 :

$$P_1 = \begin{pmatrix} 5.4174 & 0.0220 & -0.0128 \\ 0.0220 & 5.6812 & 0.0189 \\ -0.0128 & 0.0189 & 5.5808 \end{pmatrix}, P_2 = \begin{pmatrix} 5.5756 & 0.0131 & -0.0251 \\ 0.0131 & 5.8226 & -0.0135 \\ -0.0251 & -0.0135 & 5.7746 \end{pmatrix}$$

$$G_1 = \begin{pmatrix} 0.6985 & -0.0527 & 0.1130 \\ -0.0160 & 0.3234 & 0.0015 \\ 0.0384 & 0.0242 & 0.5142 \end{pmatrix} \quad G_2 = \begin{pmatrix} 0.7715 & -0.0602 & 0.0654 \\ -0.0224 & 0.3343 & 0.0332 \\ 0.0342 & 0.0193 & 0.3660 \end{pmatrix}$$

$$F1 = (0.0997 \quad -0.1065 \quad -0.1300) \qquad F2 = (0.0777 \quad -0.0842 \quad -0.0964)$$

With the application of theorem 4, with α=0.6 , the results of LMI gives other definite positive matrices P_1, P_2 and matrices G_1, G_2, R, F_1 and F_2 .

$$P_1 = \begin{pmatrix} 2.6444 & 0.0149 & -0.0295 \\ 0.0149 & 2.6722 & -0.0537 \\ -0.0295 & -0.0537 & 2.6953 \end{pmatrix}, \; P_2 = \begin{pmatrix} 2.6291 & -0.0149 & 0.0295 \\ -0.0149 & 2.6012 & 0.0537 \\ 0.0295 & 0.0537 & 2.5782 \end{pmatrix},$$

$$G_1 = \begin{pmatrix} 0.3327 & 0.0229 & -0.0870 \\ 0.0811 & 0.2159 & -0.1457 \\ -0.0823 & -0.1033 & 0.2099 \end{pmatrix} \; G_2 = \begin{pmatrix} 0.3117 & 0.0126 & -0.0188 \\ 0.0093 & 0.1183 & 0.0265 \\ 0.0119 & 0.0194 & 0.0491 \end{pmatrix}$$

$$R = \begin{pmatrix} -4.7998 & -0.0984 & 0.2377 \\ 0.1461 & -4.8765 & 0.4018 \\ -0.2455 & -0.4761 & -4.9245 \end{pmatrix}$$

$$F_1 = (0.0399 \quad -0.0191 \quad -0.0272) \quad , \quad F_2 = (0.0248 \quad -0.0103 \quad -0.0176)$$

The figures 1, 2 and 3 show the convergence of state variables and the control signal to the equilibrium point zero with the application of the theorem 2.

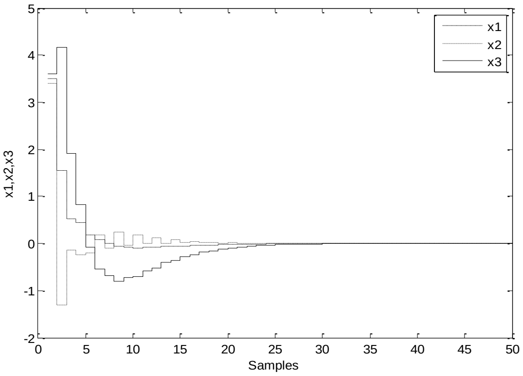


Fig. 1. Evolution of the state variables of sub-system 1

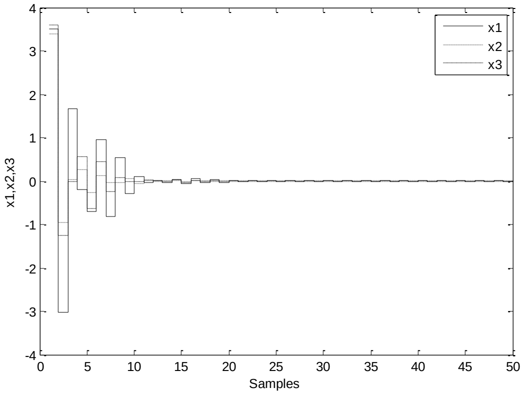


Fig. 2. Evolution of the state variables of sub-system 2

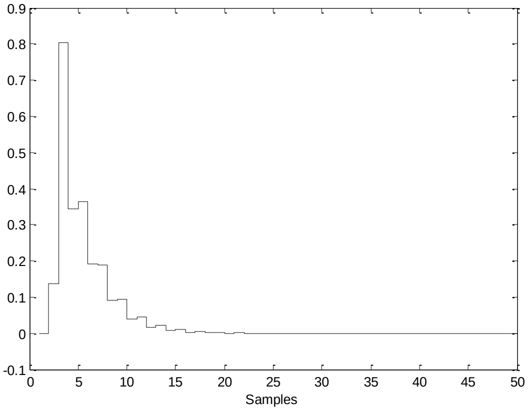


Fig. 3. Evolution of the non PDC controller signal

The next figure 4, show the feasible areas of stabilization for proposed theorems 2 and 3 and the effect of the choice of the parameters α and α_i to this areas. For the theorem 2 we choose (α=0.6) and for theorem 3 we choose

($\alpha_1 = 1,\ \alpha_2 = 1.6$)

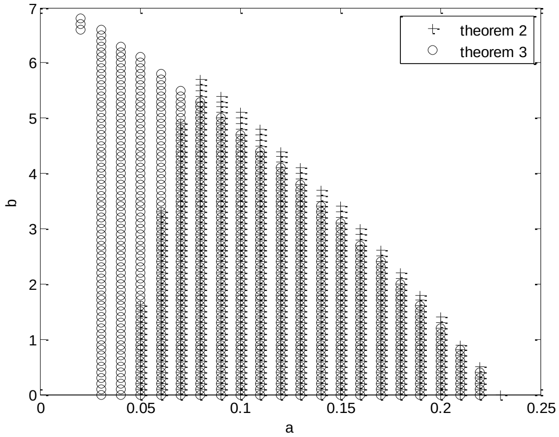


Fig. 4. Comparison between theorem 2 and 3

Theorem 3 gives a larger stabilization region than theorem 2. So by affecting for each Lyapunov matrices P a scalar α we obtain a large stabilization region than a single scalar α common to all Lyapunov matrices.

The next figures present the effect of increasing of number of parameters with $\alpha_i > 0$ in the stabilization region.

The figure 5 present the feasible area corresponding to theorem 3 with ($\alpha_1 = 0.01, \alpha_2 = 0.06$) presented by the mark (o) and ($\alpha_1 = 1,\ \alpha_2 = 1.6$) presented by the mark (+). So even we choose α_i near to zero, we obtain a larger stabilization region (feasible area of stabilization).

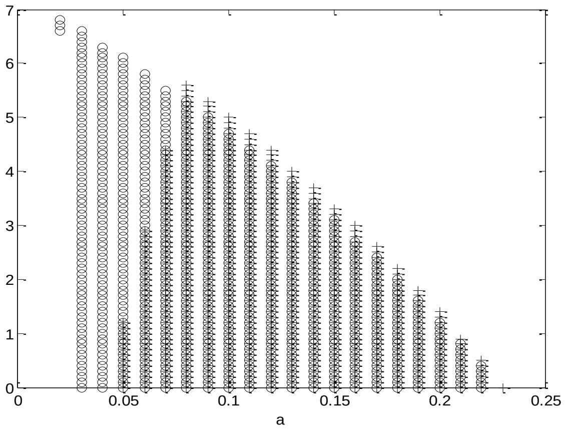


Fig. 5. Stabilization region of theorem 3

Figure 6 presents a comparison between theorem 4 and 5, it show the effect of the choice of parameters α and α_i

For the simulation, consider $(\alpha=0.6)$ for theorem 4 and $(\alpha_1 = 1,\ \alpha_2 = 1.6)$ for theorem 5. The use of theorem 5 give a large stabilization region then theorem 4.

The figure with the mark (+) represent the stabilization region of theorem 5 and the figure with the mark (o) represent the stabilization region of theorem 4.

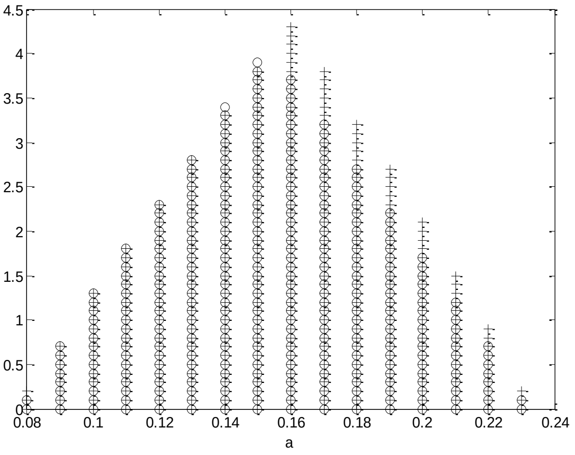


Fig. 6. Comparison between theorem 4 and 5

Figure 7 presents the effect of choice of parameter α_i near to zero with the use of theorem 5.

The figure (+) present the feasible area of stabilization for the values of ($\alpha_1 = 0.01, \alpha_2 = 0.06$) and (o) present the feasible area for the values ($\alpha_1 = 1,\ \alpha_2 = 1.6$). So to obtain a large stabilization region, α_i should be near to zero.

Figures 4,5,6 and 7 represent a comparison between the proposed theorems.

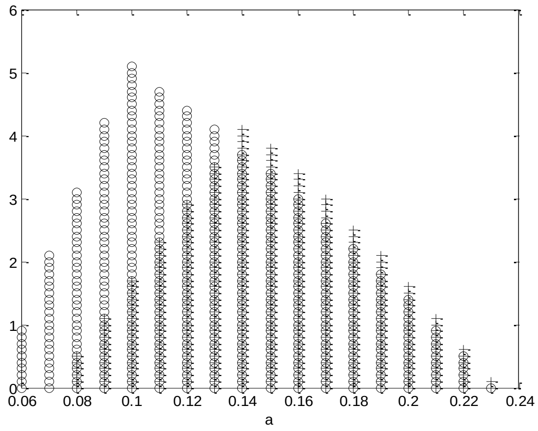


Fig. 7. Stabilization region of theorem 5

The conclusion obtained throught these figures demonstrates that the choice of a large number of parameters α_i affected to each Lyapunov matrices give a best stabilization region then one parameter and a smaller (near to zero) number also have a great effect to the stabilization region.

VI. Conclusion

This paper has developed a new fuzzy controller with state feedback for discrete time T-S parametric uncertain systems. The analysis of the stabilization problem is established by the use of Lyapunov function technique. In this case, four new Lyapunov functions are proposed. In This Lyapunov functions more parameters and slack matrix variables are introduced in order to facilitate and enrich the stabilization analysis. In the first Lyapunov function, a multiplication with a common scalar to each Lyapunov matrices is considered, In the second each Lyapunov matrices is multiplied with their own scalars. The use of the second function has a great influence to the stabilization region than the first. In the third and fourth functions, more parameters and slack matrix variables are introduced with common and single scalars for each Lyapunov matrices. Through the simulation results a single scalar for each Lyapunov matrices give a better stabilization region. The proposed theorems of stabilization was generalized between k and k+t samples time variation. Future research includes the development of design methods using these Lyapunov functions with fuzzy controller and observer in discrete time non parametric and mixed T-S uncertain systems.

Acknowledgements

The first author would like to thank the members of “Laboratoire de recherche en Automatique” (LARA – ENIT) for supporting him during the development of this work and for their advices and help.

References

[1] T. Takagi and M. Sugeno, “Fuzzy identification of systems and its application to modeling and control”, IEEE Transactions on System, Man and Cybernetics, vol 15, no1, pp.116–132. 1985.

[2] B. C Ding, “Quadratic boundedness via dynamic output feedback for constrained nonlinear systems in Takagi–Sugeno’s form”. *Automatica*, vol 45,no 5, pp 2093–2098. 2009

[3] X. Liu and and Q. Zhang, "New approaches to $H\infty$ controller designs based on fuzzy observers for T–S fuzzy systems via LMI". Automatica, vol 39 no 9, pp 1571–1582. 2003.

[4] H.J, Koo Lee J. B. Park and ,Y. H. Joo. "Robust fuzzy control of nonlinear systems with parametric Uncertainties", IEEE Transaction on Fuzzy Systems vol 9, no 2, pp 369–379. 2001

[5] Tanaka K and H.O. Wang, "Fuzzy Control Systems Design and Analysis", *John Wiley and Sons*, New York USA. 2001.

[6] G. B. Koo et al "Robust fuzzy controller for large-scale nonlinear systems using decentralized static output-feedback", International Journal of Control, Automation, and Systems, vol. 9, no. 4, pp. 649-658. 2011

[7] H.J. D Lee and W. Kim, "Robust Stabilization of T–S Fuzzy Systems: Fuzzy Static Output Feedback under Parametric Uncertainty". International Journal of Control, Automation, and Systems, vol. 7, no. 5, pp. 731-736 .2009

[8] D.H. Lee, M.H.Tak and Y.H. Joo, "A Lyapunov Functional Approach to Robust Stability Analysis of Continuous-Time Uncertain Linear Systems in Polytopic Domains", International Journal of Control, Automation,and Systems , vol 11. No. 3, pp. 460-469. 2013.

[9] Y.Y Cao and P.M Frank, "Stability analysis and synthesis of nonlinear time-delay systems via Takagi Sugeno fuzzy models", Fuzzy Sets and systems, Vol. 124 no 2, pp. 213-229. 2001

[10] L.A. Mozelli, R.M, Palhares F.O. Souza, and E.M. Mendes,"Reducing conservativeness in recent stability conditions of TS fuzzy systems", Automatica, Vol. 45, pp.1580–1583. 2009

[11] Lin C, Q.G.Wang and T.H. Lee , "Delay-dependent LMI conditions for stability and stabilization of T–S fuzzy systems with bounded timedelay", Fuzzy Sets and Systems,Vol. 157 no 9, pp. 1229-1247, 2006

[12] K. T. Tanaka, T. Hori and H.O. Wang, , "A multiple Lyapunov function approach to stabilization of fuzzy control systems", IEEE Transactions on Fuzzy Systems,Vol. 11 no 4, pp. 582–589.2003.

[13] Y Manai and M Benrejeb, " New Condition of Stabilisation for Continuous Takagi-Sugeno Fuzzy System based on Fuzzy Lyapunov Function", International Journal of Control and Automation vol. 4 no. 3, pp 51-64. 2011.

[14] A. Bouyahya. Y. Manai and J. Haggège "Observer design for Takagi-Sugeno discrete time uncertain systems", IEEE World Symposium on mechatronics engineering and applied physics WSMEAP 2015 Sousse,Tunisia.

[15] A. Bouyahya . Y. Manai and J. Haggège "New condition of stabilization for continuous Takagi-Sugeno fuzzy system based on fuzzy Lyapunov function" IEEE International Conference on Electrical Engineering and Software Applications (ICEESA), 2013 Hammamet Tunisia.

[16] A. Bouyahya . Y. Manai and J. Haggège "New condition of stabilization for non linear Takagi-Sugeno discrete time uncertain system"IEEE, 12th International Multi-Conference on Systems, Signals & Devices (SSD), 2015 Mahdia Tunisia

[17] M. Bernal, T. M. Guerra and A. Kruszewski "A membership-functiondependent approach for stability analysis and controller synthesis of Takagi–Sugeno models, " Fuzzy Sets and Systems., vol. 160, pp. 2776–2795, 2009.

[18] F. Delmotte, T.M. Guerra and A. Kruszewski, "DiscreteTakagi–Sugeno's fuzzy models: Reduction of the number of LMI in fuzzy control techniques," IEEE Trans. Syst.,Man, Cybern. B, vol. 38, no. 5, pp. 1423–1427, Oct. 2008.

[19] C. H. Fang, Y.S. Liu, S.W. Kau, L. Hong and C.H. Lee, "A new LMI based approach to relaxed quadratic stabilization of T–S fuzzy control systems," IEEE Trans. Fuzzy Syst., vol. 14, no. 3, pp. 386–397, Jun. 2006.

[20] A. Sala and C. Arino,, "Asymptotically necessary and sufficient conditions for stability and performance in fuzzy control: Applications of Polya's theorem," Fuzzy Sets and Systems., vol. 158, no. 24, pp. 2671–2686, 2007.

[21] H. Ning. Wu "An ILMI approach to robust H2 static output feedback fuzzy control for uncertain discrete-time nonlinear systems", Automatica, vol 44, pp 2333–2339, 2008.

[22] A. Kruszewski and T.M. Guerra "New Approaches for the Stabilization of Discrete Takagi-Sugeno Fuzzy Models". IEEE CDC/ECC, Séville, Espagne , 2005.

[23] Y. Wang, H. Zhang, J. Zhang and Y. Wang " An SOS-Based Observer Design for Discrete-Time Polynomial Fuzzy Systems" International journal of fuzzy systems, Vol 17, no 1, pp 94-104, March 2015.

[24] Y. Manai and M. Benrejeb, "New Condition of Stabilization of Uncertain Continuous Takagi-Sugeno Fuzzy System based on Fuzzy Lyapunov Function" *I.J.* Intelligent Systems and Applications, vol 4, pp 19-25. 2012

[25] A kruszewski , R Wang and T-M. Guerra" Nonquadratic stabilization conditions for a class of uncertain nonlinear discrete time TS fuzzy models: a new approach". IEEE Transactions on Automatic Control, vol 53, no 2,pp 606–611. 2008.

[26] J. Daafouz and J, Bernussou, "Parameter dependent Lyapunov functions for discrete time systems with time varying parameter uncertainties". System and Control Letters, vol 43, no 5, pp 355–359. 2001.

[27] M. Teixeira, R. Assunçao and E. Avellar, " On relaxed LMI-based design for fuzzy regulators and fuzzy observers". IEEE Trans. on Fuzzy Systems, vol.11 no 5, pp613-623. 2003.

[28] X. Du. and G.H.Yang, "Improved LMI Conditions for $H\infty$ Output Feedback Stabilization of Linear Discrete- time Systems", International Journal of Control, Automation, and Systems, vol. 8, no. 1, pp. 163-168, 2010

[29] M.C. Oliveira J. Bernussou, J.C. Geromel. " A new discrete-time robust stability condition". Systems & Control Letters, Vol. 37, pp261-265, 1999.

[30] S. Tong, S. Sui, and Y. Li "Fuzzy adaptive output feedback control of MIMO non linear systems with partial tracking errors constrained", IEEE Transactions on Fuzzy Systems, Vol 23, no 4, pp 729-742, August 2015.

[31] Y. Li, S. Tong, and T. Li "Observer-based adaptive fuzzy tracking control of MIMO stochastic nonlinear systems with unknow control direction and unknow deads-zones",IEEE Transactions on Fuzzy Systems, Vol. 23, no 4, pp 1228-1241, August 2015.

[32] X. Heng Chang « Robust Nonfragile H_∞ Filtering of Fuzzy Systems With Linear Fractional Parametric Uncertainties. IEEE T. Fuzzy Systems Vol 20, no 6, pp 1001-1011, 2012

[33] W. Jer Chang, C. Chieh Ku, P. Hwa Huang," Robust fuzzy Control for Uncertain Stochastic Time-Delay Takagi-SugenoFuzzy models for Achieving Passivity" Fuzzy Sets and Systems vol 161, no15, pp 2012 - 2032, 2010.

[34] A. Sarjaš, R. Svečko, and A. Chowdhury " An H_∞ Modified Robust Disturbance Observer Design for Mechanical-positioning Systems ". International Journal Of Control Automation and Systems, Vol 13no 3, pp 575-586, 2015.

[35] Q. Jia, H. Li, Y. Zhang, and X. Chen " Robust Observer-based Sensor Fault Reconstruction for Discrete-time Systems via a Descriptor System Approach" International Journal of Control Automation and Systems, Vol 13, no 2 pp 274-283,2015.

[36] D. Peaucelle, D. Arzelier, O. Bachelier and J. Bernussou, "A New Robust D-stability Condition for Real Convex Polytopic Uncertainty", Systems & Control Letters, Vol.40, pp21–30. 2000.

[37] A. Kruszewski and T. M. Guerra " New Approaches for the Stabilization of Discrete Takagi-SugenoFuzzy Models"44th IEEE Conference on Decision and Control, and the European Control Conference Seville, 2005, Spain.

[38] L.X. Wang and J.M. Mendel , "Fuzzy basis functions, universal approximation and irthogonal least-squares", IEEE T. on Neural Networks, Vol.3 no5, pp807-814. 1992.

[39] T. M. Guerra and L. Vermeiren, "LMI-based relaxed non quadratic stabilization conditions for nonlinear systems in the Takagi–Sugeno's form". Automatica, vol 40 no 5, pp 823–829. 2004

[40] A. Kruszewski, R Wang and T-M. Guerra, New approaches for stabilization of a class of nonlinear discrete time-delay models. Workshop IFAC TDS, 2006,L'Aquila, Italie.

Design of Intelligent Control System of Transformer Oil Temperature

Caijun Xu
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

Yuchen Chen
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

Liping Zhang*
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

Zhifeng Liu
School of Electronic and Electrical Engineering
Shanghai University of Engineering Science, SUES
Shanghai, China

***Abstract*—in working process of power transformer, which directly affects the safe operation of transformer oil temperature as well as the stability of the network, so vital to transformer oil temperature detection and control. Based on single chip and chip design of digital temperature measurement transformer oil temperature of an intelligent control system. The system uses a digital temperature sensor DS18B20 collection transformer oil temperature, improves the accuracy of the system. The low power consumption, strong anti-jamming ability of the SCM STC89C51 as the main controller to achieve control and real-time monitoring of transformer oil temperature, and input control module is designed for different transformer oil temperature preset control during normal operation, improving system usability and human-computer interaction.**

Keywords—transformer oil temperature; temperature control; STC89C51; DS18B20

I. Introduction

The core of transformer substation equipment, transformer repair and maintenance is a priority for maintenance personnel[1].Transformer oil temperature monitoring is maintaining and inspecting the top priority.Transformer transformer with high oil temperature will accelerate the ageing of the insulating materials, insulating materials combustion faults of transformer internal short-circuit caused by aging, rapid expansion after short circuit of transformer oil burning, transformers exploded, causing enormous economic loss and even cause accidents. In the usual maintenance work, we know transformer oil temperature gauge often fail, but poor accuracy, timeliness is not strong, and seriously affect the safe operation analysis of transformer[2,3].

In order to achieve real-time intelligent monitoring and control system of transformer oil temperature, presented by single-chip microcomputer controlled transformer cooling device the conventional control methods were improved, designed a new type of intelligent control system for temperature of transformer oil. This article gives detailed design of intelligent control system of transformer oil temperature, system design, enhanced human-computer interaction, the system is more stable and reliable[4]. The transformer oil temperature control system meets the requirements for stable and secure operation of transformer substation, for intelligent substation and laid a good foundation. Can be widely used in a variety of transformer oil temperature detection and control.

II. Control System Design

As shown in Fig.1, the master controller, keyboard module, the temperature acquisition module, temperature control module, display module, alarm module, which is made up of the transformer oil temperature intelligent control system.

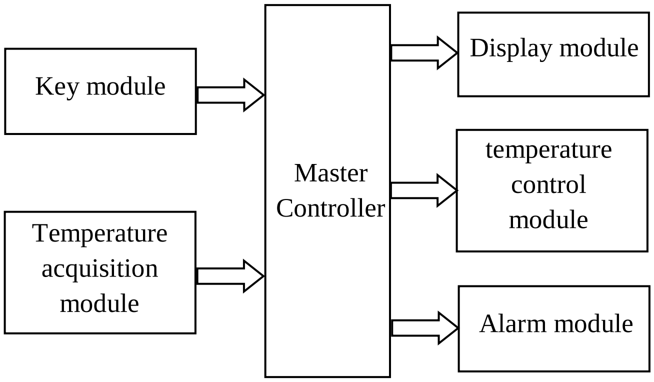


Fig. 1. block diagram of control system design

From Fig.1, we can know when the system is powered, user can through the keyboard module input preset temperature, then transmitted to the main controller processing. When the system real-time collection of temperature and the preset temperature do not match, the alarm module sends an alarm signal. At the same time, start the output control module, which makes the temperature control equipment work, adjust the system's temperature and displayed the real time temperature of the collected temperature value. When real-time acquisition of the temperature exceeds the preset temperature, control system start alarm circuit and sends out an alarm signal. At the same time to start cooling device to reduce system temperature. Repetition the transformer oil temperature is stable within the expected range.

III. Hardware Design of Control System

Control system structure is determined, each module needs to be designed.

1) The main controller module

Main controller is the core of the whole system of intelligent control. According to the system to achieve the function, main controller with input and output port is used to control the input and output module, also the main controller have programming control, considering the needs of all aspects, can be used in high reliability, strong processing ability, high speed, control function of single chip STC89C51 as the main controller[5,6].

2) The temperature acquisition module

Considering transformer oil temperature control system to realize function of high precision temperature measurement and can be used small, low power consumption, high performance digital temperature sensor DS18B20 acquisition temperature[7,8]. In the system the module is mainly the transformer oil temperature signal acquisition, and conversion to digital signal. It's can achieve an accuracy of ±0.5℃ oil temperature detection. Maintenance personnel can display real-time data direct access to current transformer oil temperature. The circuit diagram is shown in Figure 2.

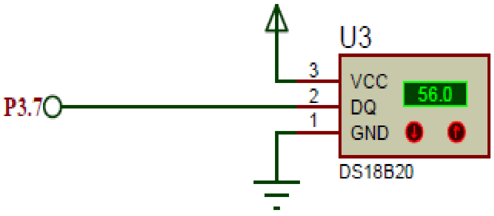


Fig. 2. The temperature acquisition module

3) The temperature display module

Considering the need to implement real-time temperature monitoring function, you can select LCD1602 with practical, cost-effective as the display module, because the LCD1602 has high display quality, digital interface, simple and comparatively small size, light weight and so on. Temperature display module in the system's main function is: displays the preset temperature value and the acquisition value, so as to achieve the purpose of real-time monitoring of transformer oil temperature. The circuit diagram is shown in Figure 3.

4) The temperature control module

In order to realize intelligent control of the temperature of transformer oil in this design system, the temperature control module mainly consists of the relay and the cooling device, which is the control of the external load of the relay[9]. By the key circuit input setting value, while the temperature acquisition circuit to obtain the current temperature signal, and then the temperature signal sent to the MCU processing. If real-time acquisition of transformer oil temperature higher than the temperature preset limit value, the micro-controller will drive the alarm module, and sends out the alarm signal, and driving cooling equipment work, when the temperature dropped to normal range and cooling equipment to stop working. To realize temperature intelligent control, to ensure that the transformer can work normally. The circuit is shown in figure 4.

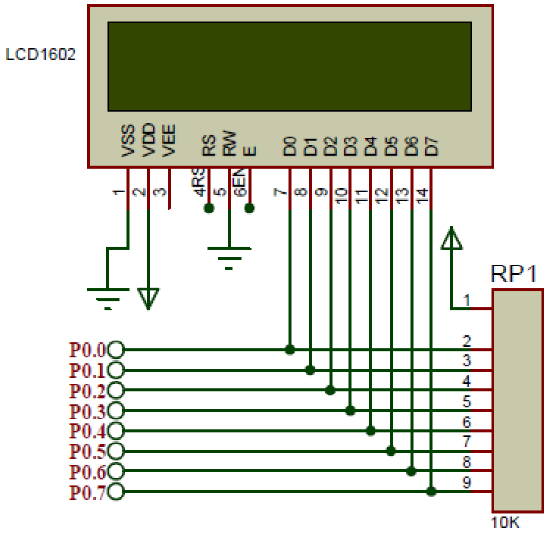


Fig. 3. The temperature display module

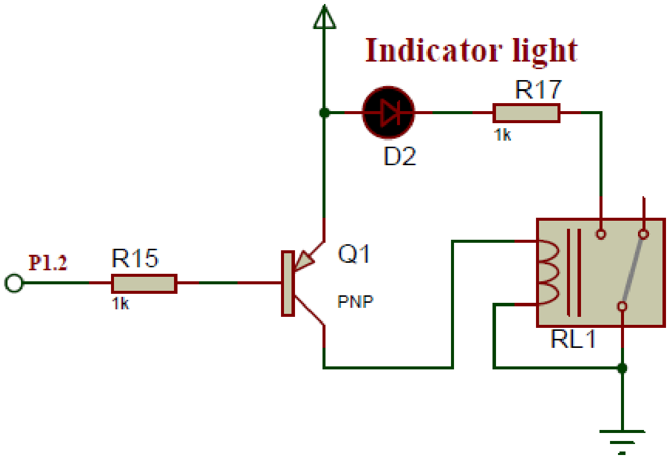


Fig. 4. The temperature control module

5) The alarm module

The core component of the alarm module is a buzzer. When the oil temperature of the real-time acquisition value exceeds the preset value, the micro-controller will drive the alarm module, light-emitting diode light, buzzer alarm signal, attract the attention of maintenance personnel. The circuit is shown in figure 5.

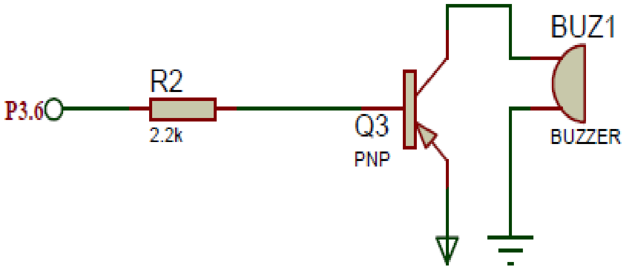
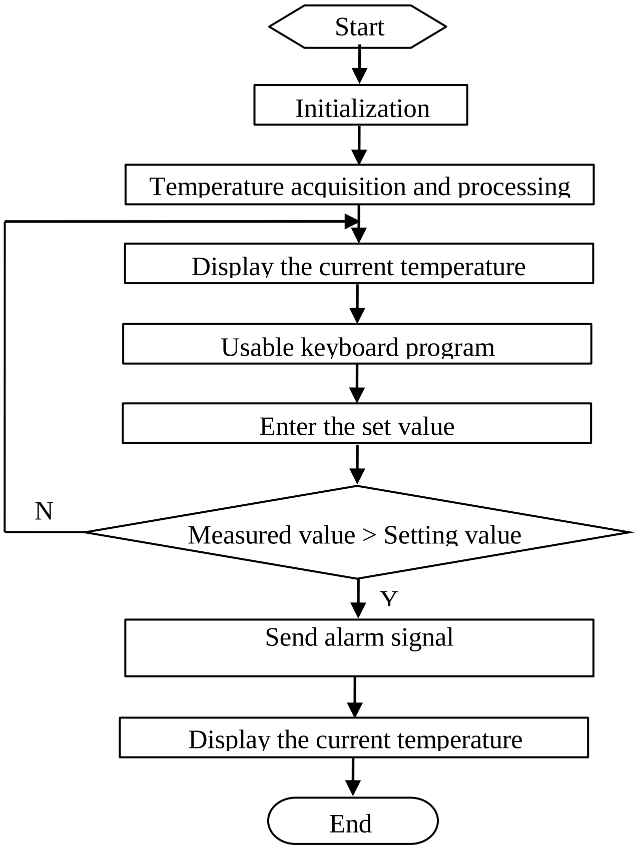


Fig. 5. The alarm module

IV. Software Design and Simulation of Control System


Start
Initialization
Temperature acquisition and processing
Display the current temperature
Usable keyboard program
Enter the set value
N
Measured value > Setting value
Y
Send alarm signal
Display the current temperature
End

A. Main program flow diagrams of control system

Software design of control system

According to the functions of the control system, the system can be divided into several modules of software design program. Then analyze the module program algorithm to finally write the procedure meets mission requirements[10]. Key handling subroutine is mainly input data to the MCU.Temperature acquisition program is the temperatures of the collecting system, and convert analog signals into digital signals.

Display subroutine is mainly display the temperature signal. By more than a few function subprogram, temperature sensors, controllers, integrated with the function of temperature control device, transformer oil temperature intelligent control, control systems of the main program flow is shown in Figure 6.

B. Simulation of control system

After the completion of the hardware and software design, it is also required by the software simulation to verify the function of the system to achieve. In this design will use the Proteus software to simulate and verify, Proteus software not only has the function of simulation, but also can simulate the MCU and peripheral devices, is the best tool for microcontroller and peripheral device simulation. Proteus is one of the most famous simulation software in the world[11]. It is the whole process from the concept to the product.

When transformer oil temperature is higher than the preset value, the MCU P1.1 output low level, buzzer alarm signal, cooling temperature indicator lights, controlled cooling equipment of relay, drive the load to cool the system, the simulation results as shown in Figure 7.

When the transformer oil temperature in the system temperature setting value range, temperature sensor continuous acquisition temperature signal, transfer to the microcontroller processing, single chip processor, the system maintained at a constant temperature and the simulation result as shown in Figure 8.

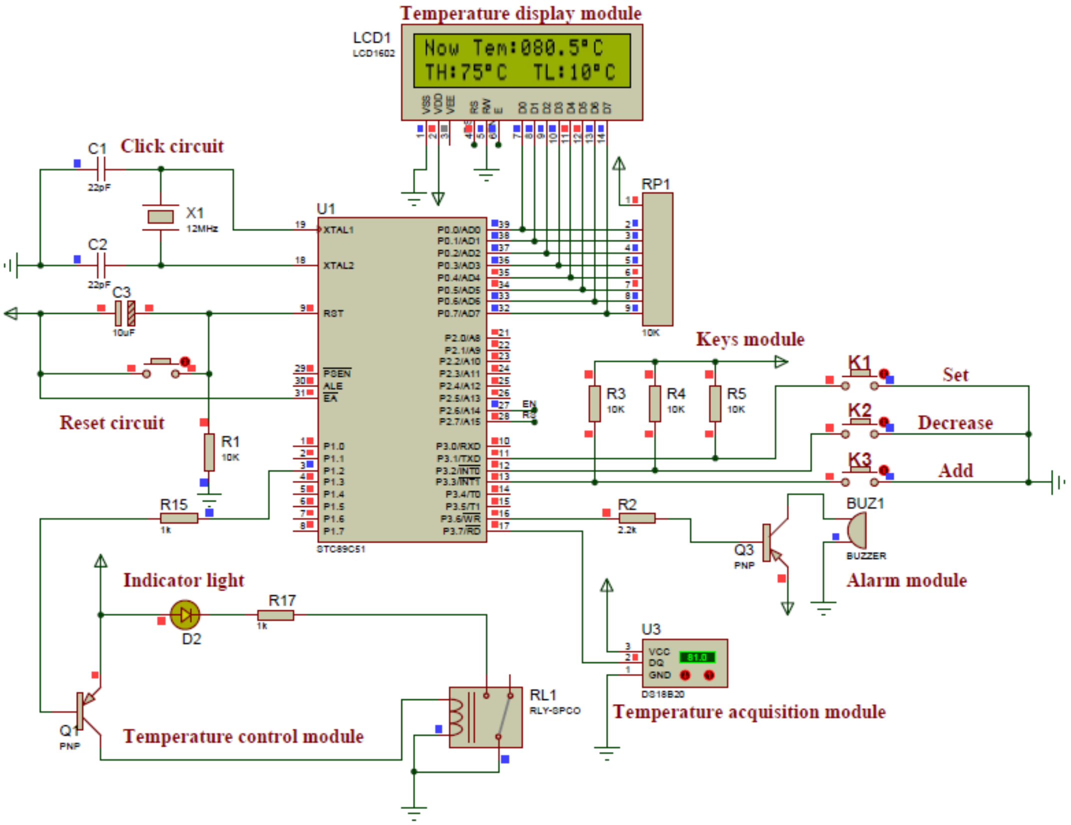


Fig. 6. Simulation of transformer oil temperature exceeds the set temperature the system value

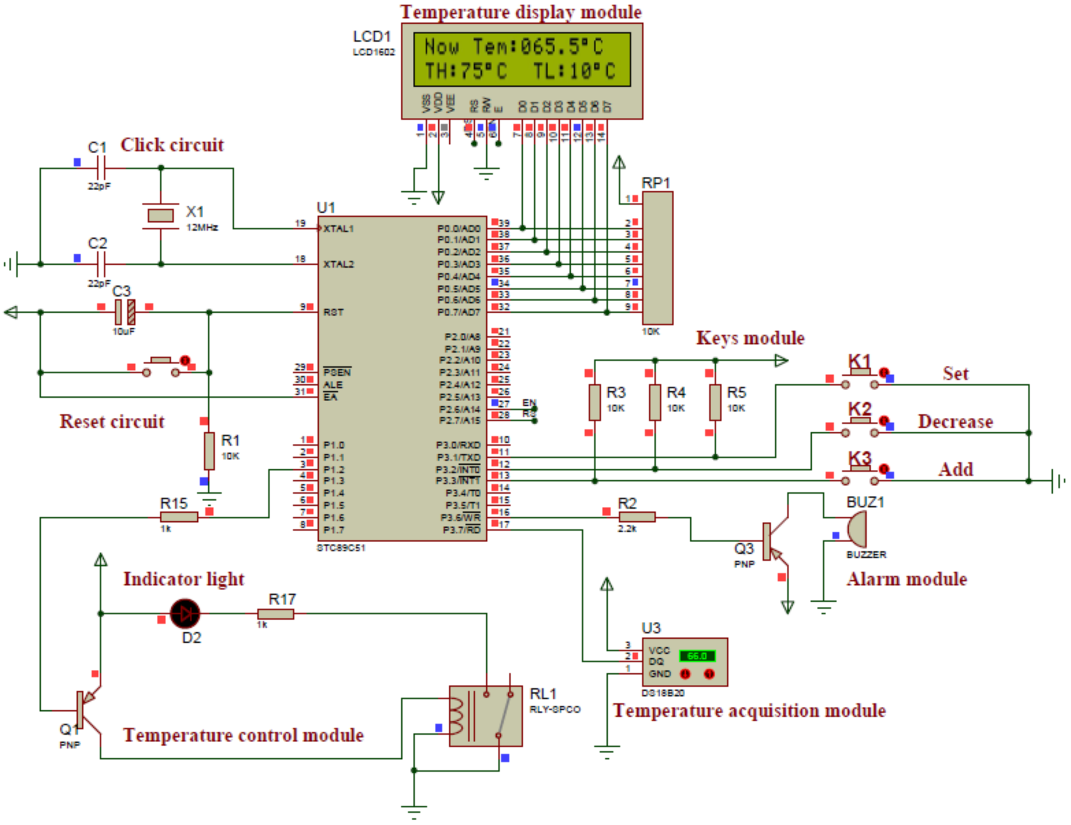


Fig. 7. Simulation of transformer oil temperature setting temperature range

V. Test Results and Analysis of Control System

In order to verify the feasibility of the control system in Shanghai University of Engineering Science, 1th substation for the actual test, chose a 10kV oil-immersed transformer, equipped with a cooling fan. Control system can control cooling fan works, control of transformer oil temperature has achieved good results, while achieving fuel temperature monitor and display. Control system to run repeatedly after a period of time, real time measurements of temperature and oil temperature gauge shows the temperature contrasts, as shown in table 1.

TABLE I. The Transformer Oil Temperature Measurement Data

Display values/℃	32	38	42	53	56	60	68
measured values/℃	32.1	38.3	41.8	53.1	56.1	59.9	68.4

Table 1 shows the transformer oil temperature control system with high accuracy and high stability. In the actual measurement process, due to the presence of magnetic field in the transformer tank, the temperature value part will be a deviation, but the data from multiple measurements can know that the system fully meet the design requirements.

VI. Conclusion

In this design mainly to the transformer oil temperature for detection and control of the object, from hardware and software two parts design, will eventually hardware and software are combined to form the transformer oil temperature intelligent control system, realization of transformer oil temperature detection and control. The real-time temperature detection function provides the accurate data for the analysis of the safe operation of the transformer. Strong alarm and temperature control function, reduce the incidence of safety accidents. The system is a new type of digital transformer temperature detection and control system can be widely used in electric power substation, applicable to the detection and control of many kinds of transformer oil temperature. The intelligent temperature control system can be extended to other fields, and it has a very wide range of applications.

References

[1] Li Pu, Xiang Xuejun, He Tilong. Based on CC1010 transformer oil temperature wireless measurement design [J]. Journal of Three Gorges University Journal (NATURAL SCIENCE EDITION) 2006 (04).

[2] Zhang Qing. Study on the temperature control system of single chip microcomputer [J]. Journal of Shanghai Jiao Tong University, 2007.

[3] Chen Jie, Huang Hong. Sensor and detection technology. Higher Education Press, 2002.

[4] Lai Shouhong. Micro computer control technology [M]. Beijing: Mechanical Industry Press, 2006.

[5] Chen Jing, Zhang Xiaoxi. Design of small constant temperature box based on single chip microcomputer [J].Modern electronic technology, 2014.

[6] LI Chang-Hua, QI Xiang-Dong. Design of Intelligent Control and Monitoring Systems for Power Transformer Oil Temperature[J]. Journal of Taiyuan University of Science & Technology, 2014.

[7] Lin, Yong Sheng. "Analysis and Countermeasures of a Transformer Oil Temperature Protection Malfunction." Zhejiang Electric Power (2010).

[8] Lesieutre, B. C., W. H. Hagman, and J. L. Kirtley. "An improved transformer top oil temperature model for use in an on-line monitoring and diagnostic system." IEEE Transactions on Power Delivery12.1(1997).

[9] Allan R N, Hizal E M. Prebreakdown phenomena in transformer oil subjected to nonuniform fields[J]. Proceedings of the Institution of Electrical Engineers, 1974.

[10] Zhao, Ailing, and J. Huang. "Temperature Control System Based On Single Chip Microcomputer." Journal of Anyang Institute of Technology(2008).

[11] Ou Yajun. The application of Proteus software in the single chip microcomputer experiment [J]. science and technology information, 2006.

Design of Wireless Temperature Measuring System Based on the nRF24l01

Song Liu
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Zhiqiang Yuan
Shanghai Electric Power Design Institute Co. Ltd
Shanghai Electric Power Design Institute
Huangpu District, Shanghai 200025, China

Yuchen Chen
College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

***Abstract*—Wireless data transmission system which composed of wireless data transmission device nRF24L01, temperature sensor [DS18B20,] and STC89C52. The system can collect and transmit temperature information and display it on LED, when the temperature excess the set value, the system will alarm by the buzzer. The hardware and software of the design are explained in detail. Finally, the application of this system in wireless temperature collection system is discussed.**

***Keywords*—nRF24L01; wireless data transmission; DS18B20; STC89C52**

I. Introduction

To avoid difficulties, easy to maintain, and improve system reliability, Compared with the previous RS485, CAN bus communication mode to collect the temperature of the temperature acquisition system. The wireless transmission chip nRF24L01 has the function of sending and receiving[1]. The hardware link layer protocol of the wireless transmission chip is very reliable, and can complete the sending, receiving, displaying and alarming of the signal[5].

II. Overall System Design

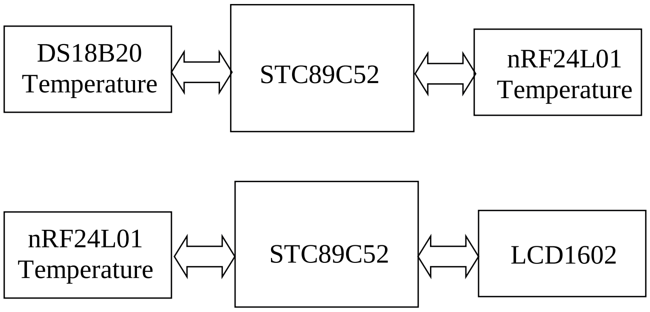


Fig. 1. System architecture diagram

In order to make the whole design idea clear, As show in the figure 1, the whole system is divided into two parts, which are sent and received. One is the acquisition and transmission part, the main use of STC89C52 as the main control chip[13], through the wireless transmission module DS18B20 temperature acquisition module nRF24L01 acquisition of temperature. Two is to receive the display part, mainly for the acquisition of temperature data processing, through the wireless transmission module nRF24L01 to handle the good data passed to the LCD screen.

III. System Design

nRF24L01 are produced by NORDIC, working on 2.4GHz~2.5GHz ISM band[10]. As a wireless transceiver chip, its built-in frequency generator, enhanced "Shock Burst" mode controller, power amplifiers, oscillators, modulators and demodulators, can be directly connected to the microcontroller I/O[2]. nRF24L01 built-in data link layer protocol, and four work modes can be configured through the configuration registers. The chip's biggest feature is the improved measurement of cable in the past abuses, and relatively accurate, reliable measurement values. By combining without A/D converter DS18B20 temperature acquisition modules[11], more convenient and efficient to collect and measure the temperature.

System hardware design is mainly composed of two parts of the collection and transmission and display[15]. Figure 2 for the collection and transmission circuit diagram, the circuit mainly by temperature sensor DS18B20, microcontroller STC89C52 and nRF24L01 composition.

Collection sent circuit 5V DC power supply, in order to strengthen the DS18B20 temperature measurement accuracy, data ports can be connected to the 4.7K pullup resistor connected with the monolithic integrated circuit system[12]. nRF24L01 needs is 3.3V voltage, this can transform a regulator AMS1117[4]. Throughout the entire acquisition circuit, schematic simplicity and easy to understand, and still achieve real-time acquisition of wireless temperature, reflects the nRF24L01 wireless transmission module and the DS18B20's efficient, convenient and practical.

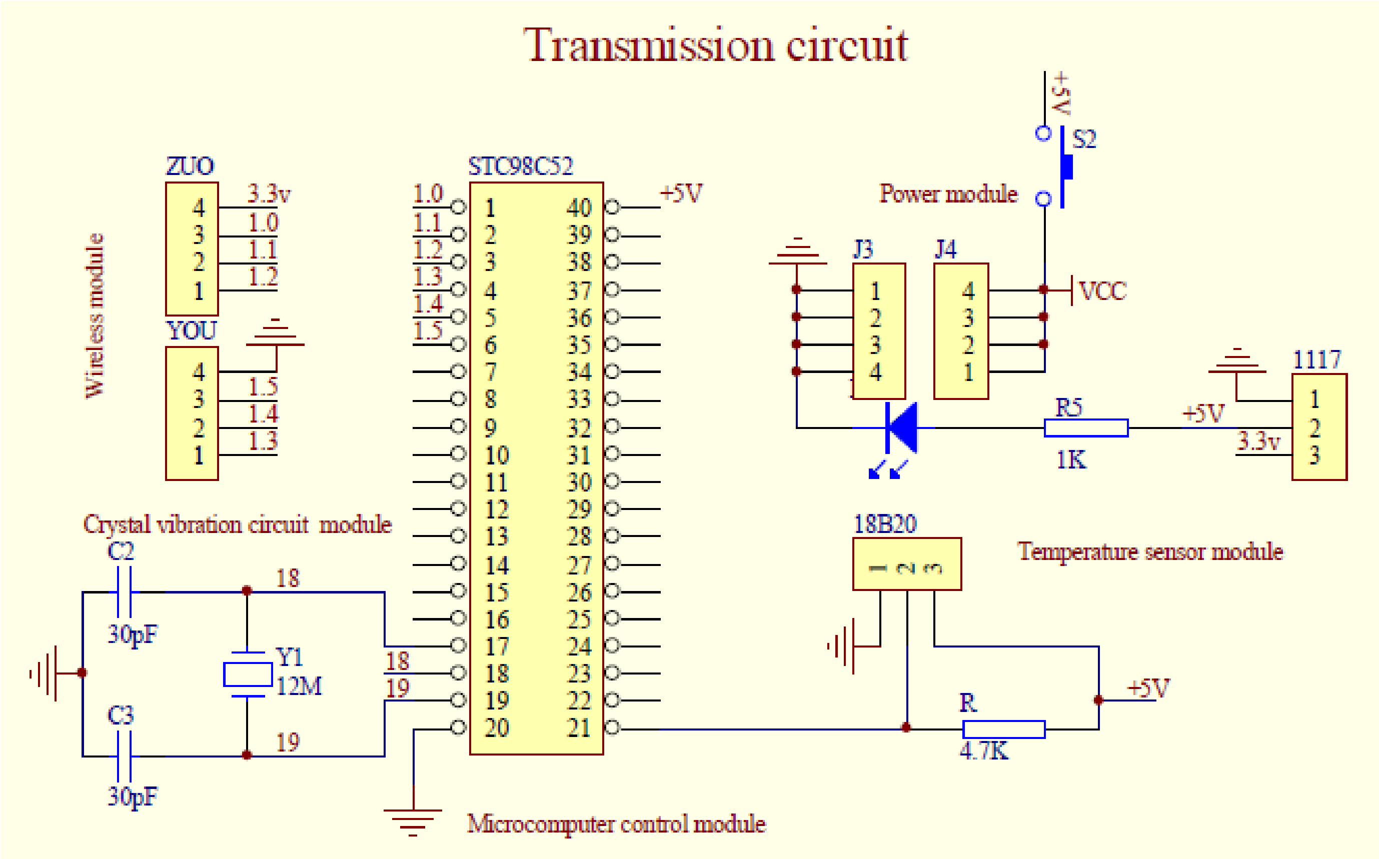


Fig. 2. Outgoing circuit

IV. System Software Design

1、Data acquisition on the part of first, and then registers the nRF24L01 wireless sensor chip configuration, bring it to a launching State and DS18B20 temperature acquisition module reset and then sends signals to the DS18B20, enable it to transform the data. Since the DS18B20 without A/D converter, it can directly read the temperature value, the end result sent by nRF24L01, concrete flow chart as shown in Figure 3.

Here special attention is single bus temperature sensor DS18B20 chip, its hardware interface is relatively simple, but relatively complex software programming of data acquisition, transmission and, therefore, strictly in accordance with the nRF24L01 configuration storage, DS18B20 and MCU interface protocol is realized by the strict timing[8]. Although software is relatively complicated, but STC89C52 speed can compensate for other deficiencies. In addition, DS18B20 programming when you want to initialize, and according to the specific requirements of the ROM operation command, memory operations, according to a certain order for data processing. If hanging on the bus one DS18B20, does not need to match ROM, skip ROM command is executed after initialization, and then send temperature conversion commands. After the temperature conversion is complete, the temperature value staged the send buffer in tx_buf, and then sent through the nRF24L01.

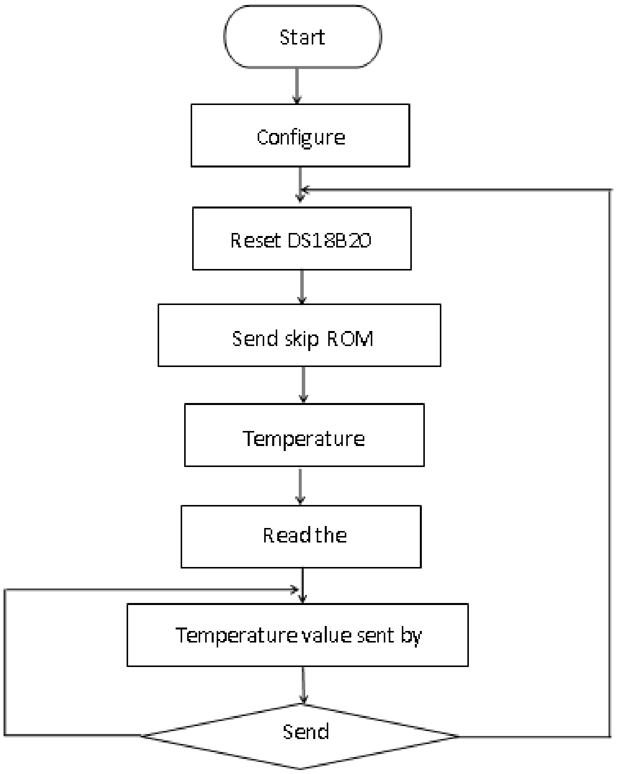


Fig. 3. Data acquisition flow

2、Because of the STC89C52 SPI bus interface, the software simulation is needed to implement the SPI bus[16]. Therefore, should be strictly in accordance with the timing of the SPI requirements, otherwise it will lead to the failure of the operation of nRF24L01[3]. All the commands of the nRF24L01 are only one byte, which is divided into reading, writing, reading, data receiving buffer, writing and sending data buffer. At the same time, the contents of the STATUS register of the MISO output[7]. Read and write program code for nRF24l01 is as follows:

```
Unit SPI_RW (unit uchar)
{
Unit bit_ctr；
for (bit_ctr=0；bit_ctr<8；bit_ctr++)
{MOSI=(unhar&0x80)；
uchar=(uchar<<1)；
SCK=1；
Uchar|=MISO；
SCK=0；}
Return (uchar)；
}
```

nRF24L01 transceiver devices requires programming to both sides, was the predominant use is receiver and sender configuration register problems[10]. This debugging method of the request, if two debugging communication extremely difficult and are not easy to find the problem. Verified in practice, the correct approach is to debug the sender, accurate and complete to be sent, go to debug the receiver, so that you can successfully complete the sending and receiving of data.

Program for the sender are as follows:

```
Void nRF24L01_TxPacket(unsigned char * tx_buf)
{
CE=0;
SPI_Write_Buf(WRITE_REG
RX_ADDR_P0,TX_ADDRESS,TX_ADR_WIDTH);
SPI_Write_Buf(WR_TX_PLOAD,tx_buf,TX_PLOAD_
WIDTH);
SPI_RW_Reg(WRITE_REG + CONFIG,0x0e);
CE=1;
iner Delay_us(10);
}
```

NRF24L01 at the receiving end is configured to receive mode, RX_AW is the address, load the data width is TX_PL_W, so that you can receive the data interrupt the CRC checksum is 2 bytes, nRF24L01 is in a P0WER_UP State[4]. the procedures are as follows:

```
Void init_nRF24L01_receive(void)

{

iner Delay_us(100);

CE=0;

CSN=1;

SCK=0;

SPI_Write_Buf(WRITE_REG+TX_ADDR,TX_ADDRES
S,TX_ADR_WIDTH);

SPI_Write_Buf(WRITE_REG+RX_ADDR_P0,RX_ADD
RESS,RX_ADR_WIDTH);

SPI_RW_Reg(WRITE_REG+EN_AA,0x01);SPI_RW_Re
g(WRITE_REG + EN_RXADDR,0x01);

SPI_RW_Reg(WRITE_REG + RF_CH,0);

SPI_RW_Reg(WRITE_REG+RX_PW_P0,RX_PLOAD_
WIDTH);

SPI_RW_Reg(WRITE_REG + RF_SETUP,0x07);

}
```

3、Results of experiments collected measurements, analysis as follows. This system uses real-time temperature acquisition is a wireless transceiver module on the environment, its main feature is the wireless measurement[14]. Distance is a major factor influencing the results of the experiment, so in the measurement process will send and receive a distance between two modules for a substantive examination and data analysis. In the measuring process, statistics on the measuring range and measuring values and sampled according to the error function and error calculation method, the final statistics are shown in table 1.

TABLE I. TEMPERATURE DATA AND STATISTICS

Measure distance/m	Ambient temperature/℃	Measurement/ ℃
3	24.2	24.3
6	24.2	22.1
9	24.2	19.7
More than 12	24.2	0

According to experimental data and related literature review and related information on the chip, 5m within the most accurate measurement results, if the premise of the whole system without any improvement under the optimized measurement scope is only limited within 12m. Considering the measurement range, usually take the external PA and LAN chip nRF24L01 chip integrated in common.[10] The wireless module to transmit power with PA and LAN more, launch distance farther, the signal is more stable, the measurement distance can reach 1000m.[3]

V. CONCLUDING REMARKS

Now the rapid development of science and technology, many areas were gradually to the development, which includes the intelligent temperature measuring and testing as we know it. Knowledge and intelligence gathering disciplines, including automatic control technology and power electronics, and for temperature acquisition we use is extremely widespread. nRF24L01, DS18B20 and STC89C52 components are introduced in this paper the wireless temperature collecting system, this system is a reflection of intelligence[9]. nRF24L01 device with low cost, high performance, and high-performance compensation to the field wiring is difficult, and increases the reliability of the system[3]. STC89C52 using single-chip microcomputer control chip, combined with no a/d chip DS18B20 temperature acquisition[13]. Practice proves that the system stability, reliability and accuracy of high practical value.

REFERENCES

[1] Shi Zhiyun, Gai Jianping, Wang Daihua, Zhang Zhijie. new high-speed nRF24L01 RF devices and applications [j]. electronic devices overseas, 2007.38 (3): 358-361

[2] Song Haidong, He Yingjie, Ma Lingling. design of wireless temperature measuring system based on the nRF24L01 [j]. today e-2010 (8): 12-14

[3] Ju Wei, Xu Liang, Diao Xiu MU. nRF2401 transceiver chip based wireless temperature and humidity collection system [j]. Ocean University of China School of information, in 2007.17 (3): 5-7

[4] Wang Zhen, Hu Qing, Huang Jie. design of wireless temperature measuring system based on the nRF24L01 [j]. 2009,17 electronic design engineering (12): 10-11

[5] Ya Wang ,Yi Jia, Qiushui Chen,Yan Yun Wang A Passive Wireless Temperature Sensor for Harsh Environment Applications[J]. Sensors,2008,8:7982-7995

[6] Wan Guangyi, Sun Jiuan, Cai Jianping.SOC microcontroller experiments, practice and application design-based on C8051F series [M]. Beijing: Beihang University press, 2006.79-80

[7] Chen A, Huang Y.Research on Medical Wireless Frequency Hopping Communication by nRF24L01[M]//Mechanical Engineering and Technology. Springer Berlin Heidelberg, 2012: 735-740.

[8] Guiyun Tian.Foundation and Application of Microcontroller [M].Higher Education Press.2001.11-35-36

[9] Bentley,R.E.Temperature and humidity measurement.In Handbook of Temperature measurement[J].Springer:New York,1998,233:5-7

[10] nRF24L01 Single Chip 2.4Gz Transceiver Product Specification[R], Nordic SEMICONDUCTOR,2007.

[11] Dallas Semiconductor Corporation. DS18B20 Programmable Resolution 1-Wire Digital Thermometer[P].Product Datasheet.2002

[12] DS18B20 Programmable Resolution 1-Wire Digital Thermometer. DALLAS SEMICONDUCTOR.

[13] Atmel Corporation.8-bit Microcontroller with 8K Byte In-System Programmable Flash AT89C52 Preliminary[S].2001.

[14] Cheol Hee Park, Min-Chil Ju.Coexistence mechanism based on adaptive frequency hopping for interference-limited wpan applications. Signal Processing and its applications,Proceedings, 2003(I):269-272.

[15] Mohamed T, Baday W.Integrated hardware-software platform for image processing applications. System-on-chip for Real-time Applications,4 IEEE International workshop, 2004.

[16] Li Zhaoqing. The principle and interface technology of single chip microcomputer [M]. Beihang University press, 2005.35-38

Risk Diffusion Modeling and Vulnerability Quantification on Japanese Human Mobility Network from Complex Network Analysis Point of View

Kiyotaka Ide, Hiroshi Sato, Tran Quang Hoang Anh, and Akira Namatame
Department of Computer Science
National Defense Academy of Japan
Yokosuka, Japan

Abstract—**The human mobility networks are vital infrastructure in recent social systems. Many efforts have been made to keep the healthy human mobility flows to maintain sustainable development of recent well-connected society. However, the inter-connectivity sometimes raises unintended diffusion and amplification of the intrinsic risks, which is difficult to forecast because of the complexity of the underlying networks. Therefore, it is believed that modeling and simulation of the risk diffusion in the human mobility networks are suggestive and meaningful. Also, recent improvement of usability of individual-level human mobility data and capabilities of high-performance computing technologies enable us to employ the data-driven approaches. In this paper, the risk diffusion dynamics is modeled based on the SIS epidemic model and the vulnerability index is defined to quantify the node-level easiness of suffering risks. We also conduct the link removal test to find the better risk mitigation methods.**

Keywords—Human mobility network; Risk analysis; Complex network analysis; SIS model

I. Introduction

Analyzing the human mobility networks is essential for making economically efficient and socially resilient social systems. It can be applied to various fields such as urban planning, public policy, and epidemic control. Maintaining the healthy flows on human mobility networks is also vital for sustaining modern society. At the same time, the complex connectivity and mobility in our society bring intrinsic risks. For example, the outbreak of Ebola virus disease from 2014 in West Africa has spread through the global human mobility networks to five other countries in the region and a few cases were found even outside the continent including several countries in Europe and the U.S. Because of the various factors, such as increment of the number of travelers, longer distances of their trips, diversification of the means of transportations, urbanization, and uneven distribution of population, the analysis of the actual dynamics on the human mobility networks becomes more difficult. Therefore, modeling and simulation of the risk diffusion on the human mobility networks are the practical options and meaningful approaches. Also, recent improvement of usability of individual-level human mobility data renders researchers' interests toward the data-driven approach, which is also promoted by the dramatical improvement of capabilities of high-performance computing technologies. Recent efforts to open the governmental statistical data to public enable us to utilize highly credible and authorized dataset.

So far, there exist a large number of previous works which analyze regional human mobility networks based on real data. For example, about ground transportation networks such as railway network and subway network, Latora and Marchiori (2002) analyze the subway network in Boston [1], Sen et. al. (2003) studied Indian railway network [2], Sienkiewicz et al. (200) analyze the public transportation networks in Poland [3], De Montis et al. (2007) examined the inter-urban commuting network among the 37 municipalities in the Sardinia region, Italy [4], Li et al. (2007) investigated Chinese railway network [5], and Soh et al. (2010) studied the bus and railway network in Singapore [6]. About the airline networks, many researchers have analyzed the world airline network (WAN) from the complex network analysis point of view [7-10] as well as for regional or domestic airport networks [11-13]. The analysis of the human mobility networks show several commonly shared characteristics. For example, as shown in most of the complex network analysis results, the weight distribution as well as degree distribution in large-scale networks tends to show scale-free property. Also, some studies reported that the networks show small world property, such as in the studies on airport network [7, 9].

In this paper, the analytical results of the risk diffusion dynamics on Japanese domestic human mobility network (JDHMN), Japanese airline network (JDAN), and world airline network (WAN) using the real human mobility data are stated. Our analysis can be applied to various fields such as diffusion of disease, rumors, and political disturbance which are spread along with the human mobility. The human mobility networks are often applied to the researches of meta-population epidemic diffusion model in which the human mobility networks, especially WAN, are utilized as paths connecting with the areas where the epidemic happens [14-23]. Moreover, K.J. Mizgier et al. (2013) recently applied the probabilistic approach to model the diffusion of the influence caused by the disruption in the supply chain and conducted the vulnerability assessment in the supply chain analysis [24].

In this paper, the link removal test in JDHMN, JDAN, and WAN for efficient control of risk diffusion are implemented, which imitates restricting the human flows on certain routes instead of restricting traffics at certain nodes. This approach can assess the influences to the risk diffusion dynamics when shutting some routes in the human mobility networks. There exist some related works about the link removal tests. Marcelino et al. (2012) evaluate the effect of the link removal in the top 500 international airports network when they simulate the SEIR meta-population epidemic model using the actual data of the outbreaks of influenza A (H1N1) in 2009 [2]. Hossain et al. (2013) investigate the resilience of the Australian Airports Network from a complex network analysis point of view [13]. They examined the resilience measures including the size of the largest component and reachability when closing the links following the certain ranking criteria of the links. Also, Verma et al. (2014) investigated the connectivity in the WAN as removing links following the ranking based on their weight. They found that the WAN is resilient when the links having large weight are cutting down since these links tend to connect between high degree nodes and there are many alternative paths between them. Conversely, they found that the closing the links having low weight damages the connectivity of the network [26].

In the rest of this paper, the chapter II introduces the datasets for the JDHMN, JDAN, and WAN. Also, the networks created from the datasets are analyzed from the stand point of the complex network analysis. In the chapter III, assuming the risks are propagate probabilistically through the links of the networks, the risk analysis on these human mobility networks are discussed. Also, the link removals for the risk mitigation are investigated to propose the effective risk mitigation methods. Finally, the chapter IV concludes this paper.

II. Human Mobility Network

A. *Japan domestic human mobility network*

In this work, the openly available public datasets which are published by the Ministry of Land, Infrastructure, Transport and tourism (MLIT) of Japan are utilized. The datasets include the amounts of passengers passing the borders between prefectures in Japan with several means of transportations, such as cars, trains, busses, airlines, and ships (http://www.mlit.go.jp/sogoseisaku/soukou/sogoseisaku_soukou_fr_000008.html). The datasets include the information of the passengers' purposes of travels, genders, and ages. The values of the human mobility data are estimated based on the results of the questionary based random sampling surveys to the passengers and the activity reports provided from the operators of the transportation services. In order to consider the overlapping when integrating the number of passengers of multiple means of transportation, the dataset is adjusted based on the credibility of the data. Also, the dataset does not include the human mobility data between the prefectures within the Kanto region (Tokyo, Kanagawa, Chiba, and Saitama), Kinki region (Osaka, Kyoto, Hyogo, and Nara,), and Chukyo region (Aichi, Gifu, and Mie). These three are largest regions in Japan which are the critical areas for our analysis. Therefore, the lacking data for these regions are complemented by another public dataset (http://www.mlit.go.jp/k-toukei/cgi-bin/search.cg). This dataset only includes the passengers' travel data counted by the transportation service operators. Fig.1 shows the visualization of the JDHMN. The locations of the nodes indicate the locations of the local governmental offices of each prefecture. The connections between these nodes represent the human mobility flows between the connecting prefectures. The color, width, and transparency of the links represent the relative significance of the volume of human mobility.

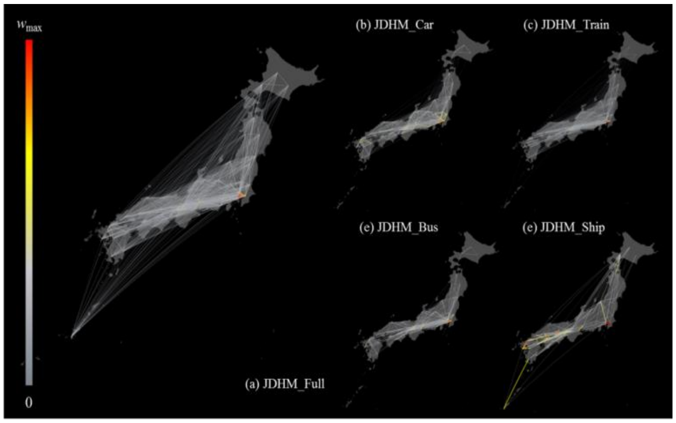


Fig. 1. Visualization of the weighted network of Japan domestic human mobility (JDHM) derived from the openly available public data. The locations of the nodes in these networks are corresponding with the locations of the prefectural governmental offices. The link between the nodes represents the human mobility between the prefectures. The color, width, and transparency of the links represent the relative amounts of the human mobility on the link

B. *Airline networks; Japanese domestic and world airline network*

In this work, the public open datasets of the JDAN in 2014 are used. The datasets are published by the MLIT annually and they contain the information of monthly and yearly flight number, the number of passengers, and weights of freights. We also analyze the dataset of the WAN whose data is provided by openflights.org (http://openflights.org/data.htm) as open data. The datasets include the data of the world-wide airports network and the information of the airline company operating the routes. The largest strongly connected cluster from the WAN which consists of 3,34 airports are extracted.

Fig.2 (a) shows the network structure of JDAN where the relative number of passengers is used to weight the links. Fig.2 (b) is the structure of WAN from openflight.org.

C. *Airline networks; Japanese domestic and world airline network*

Firstly, the JDHMN are analyzed from the complex network analysis point of view. Fig.3 shows the distributions of the link's weights (i.e. amounts of people) of JDHMN. Since, if we do not consider the weighted on the links, the network structures are almost complete network, the distributions of the links' weights in each network are investigated. In these figures, the red plots represent the frequency for every 10,000 travelers. The fitting lines in the figures were computed utilizing the data between the links' weights from 0 to 10^6 people which show the power law distributions with the power exponents within the range of 1.3 ± 0.1.

Next, the community detection which finds the hidden community structures (i.e. the clusters of densely connected components) in the networks are conducted.

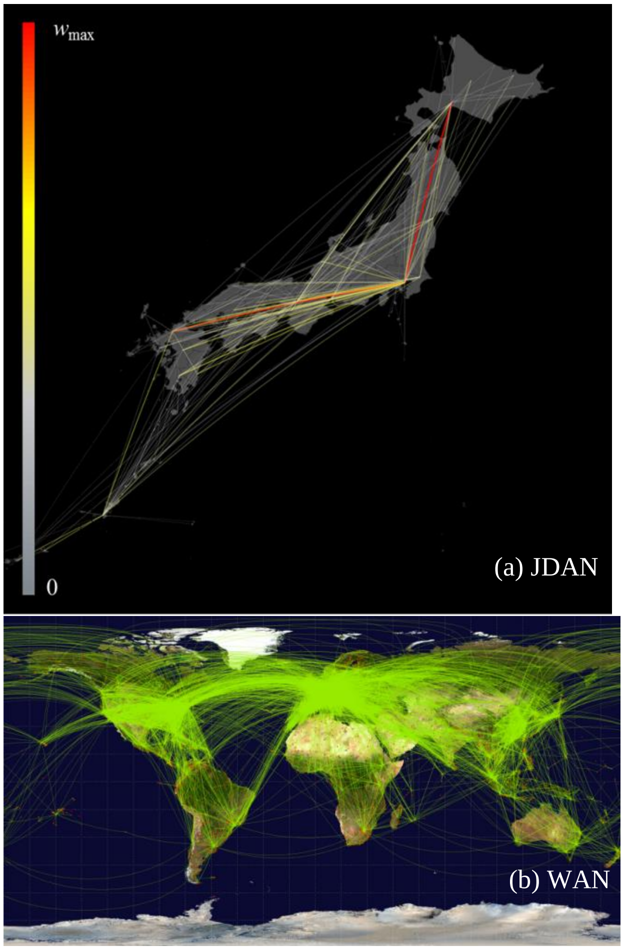


Fig. 2. Visualization of the airline networks; (a) Japan domestic airline network (JDAN) in 2014 in which the color, width, and transparency represents the relative number of passengers and (b) world airline network (WAN) from the website of openflights.org (http://openflights.org/data.html)

We utilized the infoMAP algorithm [27] for the community detection. The infoMAP algorithm solves the problem of optimally compressing the information of a random walk occurs on the network. The optimally compressed information can be recovered to the original information as closely as possible when the compressed information is decoded, which can be considered as the problem of finding the optimal partition of the clustered structures. The infoMAP algorithm is applicable into detecting community structures in the directed and weighted networks. Also, it is reported that the best performance of community detection is attained in the comparison tests on the benchmark networks comparing in the several well-known community detection algorithms [28].

Table.1 shows the results of the community detection on JDHMN, JDAN and WAN. The first row of the table shows the numbers of the detected communities in each network when the weighted modularity Q_w for each network is maximized. In the community detection algorithms, the partition of the network is optimized so that the value of Q_w is maximized. Therefore, higher the value of the optimal modularity Q_w, better the partition of the network is. The weighted modularity Q_w can be computed by the following equation [27],

$$Q_w = \sum_{i=1}^{m} \left(\frac{w_{ii}}{w} - \frac{w_i^{\text{in}} w_i^{\text{out}}}{w^2} \right) \qquad (1)$$

where w indicates the total weights of the links in the network, w_{ii} represents the total weights of the links in a detected module i, w_i^{in} is the total weights of the inward links coming into the module i, and w_i^{out} is the total weights of the outward links departing from the module i, and m denotes the number of the detected modules. It is generally known that, when the optimal modularity $Q_w > 0.3$, the network can be considered having community structures. Therefore, as can be seen in Table.1, the JDHMN and WAN have the community structure, meanwhile, JDAN shows almost zero optimal modularity Q_w, which means JDAN does not have community structure.

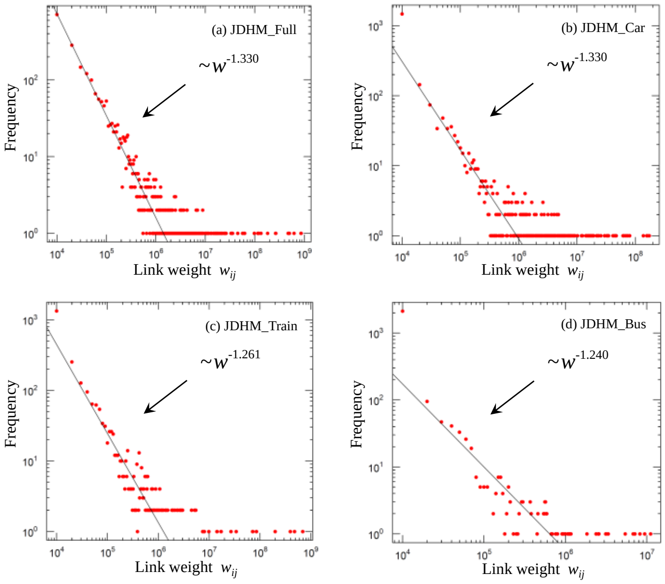


Fig. 3. Distribution of the links' weights of several JDHMN; (a) for full means of transportation, (b) cars, (c) trains, and (e) bus. The red plots represent the frequency for every 10,000 travelers. The linear regression curves show the links' weight distributions follow power law with the power exponent within the range of 1.3±0.1

TABLE I. RESULTS OF THE COMMUNITY DETECTION OF JDHMN AND JDAN AND WAN BY INFOMAP ALGORITHM

Network	JDHMN					JDAN	WAN
	Full	Car	Bus	Train	Ship		
# of Communities N_c	10	9	8	7	10	4	163
Optimal Modularity Q_w	0.446	0.597	0.525	0.320	0.693	-0.004	0.606

Fig.4 shows location of the communities which was detected from the networks of the full means of transportation in the JDHMN. The each colored region on map corresponding with 10 communities which are detected by the community detection shown in Table.1 This separation is well fit with the traditionally used regions of Japan, such as Tohoku region, Kanto region and Kinki region etc., which is another evidence that supports the network can be divided into the community structures and the infoMAP algorithm shows excellent performance.

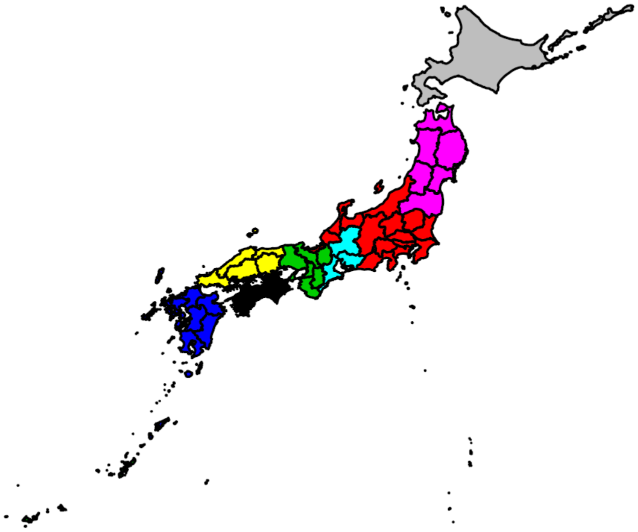


Fig. 4. The ten community structures detected from the network of the full means of transportation of JDHM by the infoMAP algorithm. Each colored region corresponds with the detected communities of prefectures. The black solid separation line on the map represents the borders between the prefectures

The network assortativity is one of the conventionally used network properties to quantify the tendency that each node in the network tend to connect to the nodes which have similar degree, which can be computed by a correlation function of degrees of nodes connected each other [29]. When the network assortativity is positive, the nodes in the network tend to connect with the nodes with similar degree, meanwhile, when it is negative, the nodes in the network tend to connect with the nodes having different degree.

The following equation is the definition of the network assortativity r [29],

$$r = \frac{M^{-1}\sum_{\phi}(\prod_{i\in F(\phi)}k_i) - \left[\frac{M^{-1}}{2}\sum_{\phi}(\sum_{i\in F(\phi)}k_i)\right]^2}{\frac{M^{-1}}{2}\sum_{\phi}(\sum_{i\in F(\phi)}k_i^2) - \left[\frac{M^{-1}}{2}\sum_{\phi}(\sum_{i\in F(\phi)}k_i)\right]^2} \qquad (2)$$

where $F(\phi)$ represents the set of connecting pair of two nodes, ϕ denotes the identification number of links connecting the pairs, M denotes the number of links in the network, k_i denotes the degree of the node i. Then, this concept was extended to the weighted assortativity [30] as defined below,

$$r^w = \frac{H^{-1}\sum_{\phi}(\varpi_\phi \prod_{i\in F(\phi)}k_i) - \left[\frac{H^{-1}}{2}\sum_{\phi}(\varpi_\phi \sum_{i\in F(\phi)}k_i)\right]^2}{\frac{H^{-1}}{2}\sum_{\phi}(\varpi_\phi \sum_{i\in F(\phi)}k_i^2) - \left[\frac{H^{-1}}{2}\sum_{\phi}(\varpi_\phi \sum_{i\in F(\phi)}k_i)\right]^2} \qquad (3)$$

$$\bar{\delta}_i = \frac{1}{d_i}\sum_{j=1}^{d_i}\left|d_j - d_i\right| \Big/ \sum_{i=1}^{N}\frac{1}{d_i}\sum_{j=1}^{d_i}\left|d_j - d_i\right| \qquad (7)$$

where ϖ_ϕ denotes the weight of ϕth link, H represents the sum of the weights in the network.

Table.2 shows the weighted assortativity r^w of the JDHMN, JDAN, and WAN.

TABLE II. WEIGHTED ASSORTATIVITY OF THE NETWORKS

Network	JDHMN					JDAN	WAN
	Full	Car	Bus	Train	Ship		
Weighted Assortativity r^w	-0.0221	-0.0624	-0.114	-0.0395	-0.0803	-0.252	-0.0185

Local assortativity was firstly proposed by Piraveenan et al. (2008) [31], and improved by themselves by removing a bias towards low-degree nodes as unbiased local assortativity [32]. The concept of the unbiased local assortativity is to calculate the contribution from each node in network assortativity defined by Eq.(2) so that the sum of the values of the local assortativity of each node is equal to the network assortativity.

$$\hat{\rho}_i = \frac{j(j+1)(\bar{k} - \mu_q)}{2M\sigma_q^2} \qquad (4)$$

Eq.(4) shows the definition of the unbiased local assortativity of node i [32],

where M represents the number of links in the network, the excess degree j means the number of links except the link which is used to reach the nodes, therefore it can be calculate $d_i - 1$ when d_i denotes the degree of node i, $\bar{k}$ denotes the means of the excess degree, σ_q represents the standard deviation of the network's excess degree distribution $q(k)$, and μ_q denotes the expectation of the excess degree distribution $q(k)$.

However, the unbiased local assortativity $\hat{\rho}_i$ is a relative measure, and it can quantify the relative contributions of each node to the network assortativity. Therefore, it works only when comparing within the focal network. Only considering the local value itself on a node, we cannot judge the node is assortative or disassortative. Therefore, recently, Thedchanamorthy and Piraveenan et al. (2014) proposed

$$\delta_i = \lambda - \bar{\delta}_i \qquad (5)$$

alternative approach to compute the local assortativity [33].

$$\lambda = \frac{r+1}{N} \qquad (6)$$

Their new definition of the local assortativity δ_i is as follows,

where λ is a scaling factor computed as follows,

where r denotes network assortativity derived from Eq.(1), and N denotes the number of nodes.

$\bar{\delta}_i$ in Eq.(7) is the relative average neighbor difference which computed as follows, where d_i denotes the degree of node i. d_j represents the degree of neighbor node j of node i. The relative average neighbor difference means the relative values that quantify the relative dis-assortativity of each node so that the sum of $\bar{\delta}_i$ = 1.

This concept of the local assortativity δ_i is extended to the weighted network and proposed the local weighted assortativity δ_i^w. The proposed definition of the local weighted assortativity is as follows,

$$\delta_i^w = \lambda^w - \bar{\delta}_i^w \tag{8}$$

where the scaling factor λ^w can be calculated by just using weighted network assortativity r^w instead of the normal network assortativity r as follows,

$$\lambda^w = \frac{r^w+1}{N}, \tag{9}$$

and the relative average neighbor difference in terms of weight, $\bar{\delta}_i^w$, can be computed as follows,

$$\bar{\delta}_i^w = \frac{\frac{1}{d_i}\sum_{j=1}^{d_i}\left|\sum_{k=1}^{d_j} w_{j,k}^{in} - \sum_{j=1}^{d_i} w_{i,j}^{in}\right|}{\sum_{i=1}^{N}\frac{1}{d_i}\sum_{j=1}^{d_i}\left|\sum_{k=1}^{d_j} w_{j,k}^{in} - \sum_{j=1}^{d_i} w_{i,j}^{in}\right|} \tag{10}$$

where $w_{i,j}^{in}$ represents weight of the inward link from the neighbor nodes j of the node i to the target node i. The node k represents the neighbor nodes of the node j

Fig.5 shows the distribution of the local weighted assortativity δ_i^w for the JDHMN and JDAN, and Fig.6 shows distribution of δ_i^w for WAN. The blue circles indicated the nodes which have the negative local weight assortativity, meanwhile, the red circles show the nodes with the positive local weight assortativity. The size of these circles indicated the significance of the node strength (i.e. weighted indegree) and the transparency represents the significance of the absolute values of the weighted local assortativity δ_i^w.

The locations of the nodes in Fig.5 are corresponding with the locations of the local government offices in each prefecture in Japan. As can be seen in these figures, in the JDHMN of the full means, car, bus, and train and JDAN, remarkably large and thick blue circle locates at Tokyo, which indicates that Tokyo is the hub in which a large number of travelers' mobility concentrated on and the average difference of travelers' flows between the neighbor prefectures' is extremely large.

In the JDHMN of the full means, car, bus, and train, Osaka and Aichi are the second and the third largest hubs. In ship networks, the hub-like nodes are located at several prefectures in west part of Japan as well as around Tokyo.

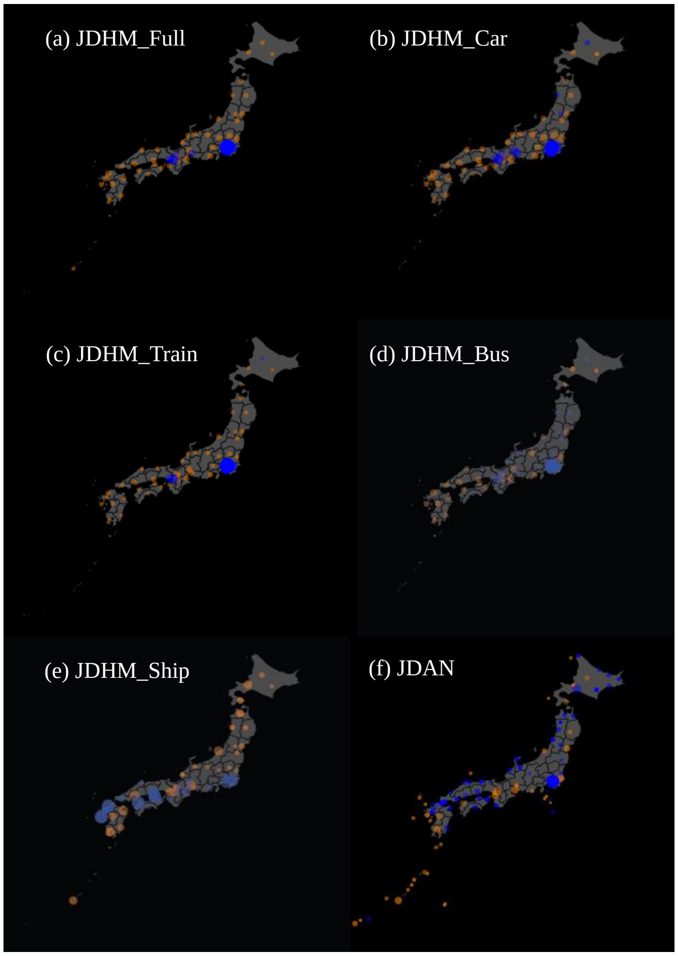


Fig. 5. Distribution of the local weight assortativity δ_i^w for JDHMN for several means of transportation and JDAN

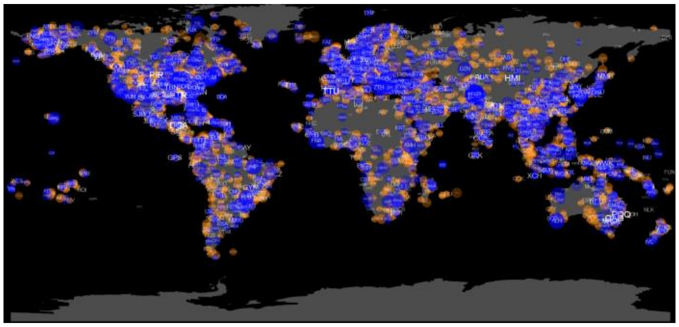


Fig. 6. Distribution of the local weight assortativity δ_i^w for WAN

III. Risk Analysis of Human Mobility Networks

A. Risk diffusion model

In this section, the risk diffusion in the human mobility networks is analyzed. To model the risk diffusion in the networks, the epidemic model (the susceptible-infected-susceptible (SIS) model) [34-38] is applied.

It is assumed that risk propagates through the links, and the diffusion probability on each link is proportional with the relative weights of each links.

The risk diffusion probability matrix **M** is defined as follows,

$$\mathbf{M} = \beta\mathbf{W} + (1\text{-}\delta)\mathbf{I} \tag{11}$$

where the (i, j) element in the relative propagation probability matrix **W** can be computed by the link weight $w_{i,j}$ from node i to node j divided by the maximum link weight $w_{\max}$ as follows,

$$\mathbf{W} = \begin{pmatrix} w_{1,1}/w_{\max} & w_{1,2}/w_{\max} & \cdots & w_{1,n}/w_{\max} \\ w_{2,1}/w_{\max} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ w_{n,1}/w_{\max} & \cdots & \cdots & w_{n,n}/w_{\max} \end{pmatrix} \tag{12}$$

Then, the risk diffusion probability from node i to node j is given by the $\beta \times w_{i,j} / w_{\max}$. Here, β is the weighting parameter to control the significance of risk transition, and δ represents the risk removal probability which is set to a fixed value.

The risk propagation dynamics over time is them given as follows,

$$\mathbf{p}(t) = \mathbf{p}(0)\mathbf{M}^{t-1}, t = 1,2,3,\cdots, \tag{13}$$

where $\mathbf{p}(t)$ is the risk probability vector in which the i^{th} element represents the risk probability of node i at time t, and $\mathbf{p}(0)$ denotes the initial risk probability vector.

B. Vulnerability index

The vulnerability index (VI) which can quantify the node-level easiness of being suffered from the risk is proposed. The VI is defined by the accumulative probability of being suffered from the risks when each node is selected as the initially infected node in turn. The following equation shows the equation to compute the VI vector $\mathbf{v}_{\mathrm{VI}}$,

$$\begin{aligned} \mathbf{v}_{\mathrm{VI}} &= \sum_{\mathrm{t}=1}^{\infty} \mathbf{p}(\mathrm{t}) \\ &= \mathbf{e}\left(\mathbf{I} + \mathbf{M} + \mathbf{M}^2 + \mathbf{M}^3 + \cdots + \mathbf{M}^t + \cdots\right) \\ &\approx \mathbf{e}\left(\mathbf{I} - \mathbf{M}\right)^{-1}, \end{aligned} \tag{14}$$

where **e** denotes the all-one vector which represents the initial attack assigned to all nodes in the network, namely $\mathbf{p}(0) = \mathbf{e}$. Then, to remove the influences from the initial attack to all nodes and consider the differences of the size of the network, the normalized VI vector $\mathbf{v}'_{\mathrm{VI}}$ is suggested as follows,

$$\mathbf{v}'_{\mathrm{VI}} = \frac{\mathbf{v}_{\mathrm{VI}} - \mathbf{e}}{N} \tag{15}$$

However, Eq.(14) can work only when the risk diffusion probability matrix **M** are sufficiently small and converge to zero matrix at the infinite time. Therefore, the relative vulnerability index $\hat{\mathbf{v}}_{\mathrm{VI}}$ is proposed as follows,

$$\hat{\mathbf{v}}_{\mathrm{VI}} = \mathbf{e}\left(\hat{\mathbf{M}} + \hat{\mathbf{M}}_2 + \cdots + \hat{\mathbf{M}}_T\right)\Big/T \tag{16}$$

where T is a sufficient large positive integer and $T = 100$ is used for our computations. $\widehat{\mathbf{M}}_x$ is a matrix in which each element of $\mathbf{M}^x$ is divided by the sum of the elements $m_{i,j}$ of $\mathbf{M}^x$ as follows,

$$\hat{\mathbf{M}}_x = \frac{\mathbf{M}^x}{\sum_{i,j} m_{i,j}} \tag{17}$$

We compare the relationships between the relative vulnerability index $\hat{\mathbf{v}}_{\mathrm{VI}}$ and the weighted in-degree (i.e. node strength) vector $\mathbf{w}_{\mathrm{in}}$, weighted out-degree vector $\mathbf{w}_{\mathrm{out}}$, weighted eigenvector centrality vector $\mathbf{c}^{w}_{\mathrm{EVC}}$, weighted closeness centrality vector $\mathbf{c}^{w}_{\mathrm{CC}}$, weighted betweenness centrality vector $\mathbf{c}^{w}_{\mathrm{BC}}$, and weighted PageRank vector $\mathbf{c}^{w}_{\mathrm{PR}}$. $\mathbf{w}_{\mathrm{in}}$ is a vector in which the i^{th} element consists of the sum of the weights of the inward links to the i^{th} node. $\mathbf{w}_{\mathrm{out}}$ is a vector in which the i^{th} element consists of the sum of the weights of the outward links directing from to the i^{th} node. $\mathbf{c}^{w}_{\mathrm{EVC}}$ is the largest eigenvector of the weighted matrix. $\mathbf{c}^{w}_{\mathrm{CC}}$ considers the path with the minimal weights as shortest path, and computes the inverse of the sum of the weights of the weighted shortest path. c$\mathbf{c}^{w}_{\mathrm{BC}}$ is a vector in which the i^{th} element considers the path with the minimal weights as a shortest path and computes the relative value of the number of the weighted shortest paths which pass through the node i. $\mathbf{c}^{w}_{\mathrm{PR}}$ uses the weighted matrix to compute the transition probability matrix **M**, and the weight on a link between node i and node j, w_{ij}, is used instead of a_{ij} of the adjacency matrix. Table.3 shows the correlation coefficients between the six weighted centrality measures and relative vulnerability index $\mathrm{v}\hat{\ }_{\mathrm{VI}}$ for each human mobility networks, JDHMN, JDAN, and WAN. As can be seen in this table, $\mathbf{c}^{w}_{\mathrm{EVC}}$ shows very high correlation with $\hat{\mathbf{v}}_{\mathrm{VI}}$.

Fig.7 shows the correlation diagram between the relative vulnerability index $\hat{\mathbf{v}}_{\mathrm{VI}}$ and the weighted eigenvector centrality $\mathbf{c}^{w}_{\mathrm{EVC}}$ in the seven human mobility network when β=0.5 and δ=1. The red circles represent the node having the positive local weighted assortativity and the blue triangles represent the nodes having the negative local weighted assortativity. The size of the plots indicates the absolute significance of the local weighted assortativity of each node. Since the weights of each nodes are basically different (except the special case like the homogeneously weighted network), the positive local weighted assortativity means that the averaged absolute difference between the weight of the target node and the weights of its neighbor nodes is relatively small, which is achieved when the target node is connected to both of higher and lower weight neighbor nodes evenly, or when the weights with connected neighbor nodes are similar, which is sometimes occur at the peripheral nodes. As can been seen in Fig.7, $\hat{\mathbf{v}}_{\mathrm{VI}}$ shows positive correlation with the weighted eigenvector centrality $\mathbf{c}_{\mathrm{EVC}}^{w}$.

As can be seen in these figures, the nodes having high relative vulnerability index and high $\mathbf{c}_{\mathrm{EVC}}^{w}$ have very large negative local weighted assortativity.

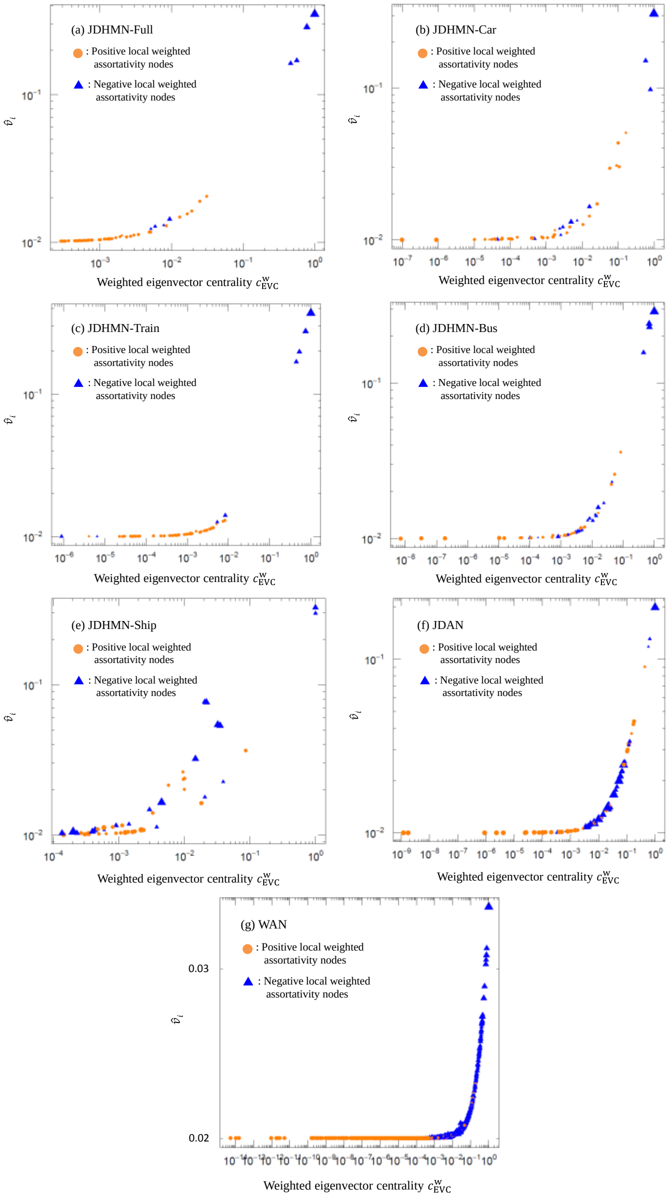


Fig. 7. Relationship between the elative vulnerability index $\hat{\mathbf{v}}_{\mathrm{VI}}$ and $\mathbf{c}_{\mathrm{EVC}}^{\mathrm{w}}$. in the JDHMN of (a) Full, (b) Train, (c) Car, (d) Bus, (e) Ship, (f) JDAN and (g) WAN when β = 0.5 and δ = 1. The red circles represent the node having the positive local weighted assortativity and the blue triangles represent the nodes having the negative local weighted assortativity. The size of the plots indicates the absolute significance of the local weighted assortativity of each node

C. *Mitigation of risk diffusion*

How to control risk diffusion in real networks is important concern. In this section, the effects of the link removals as the methods to mitigate the risk diffusion are investigated. The effects when links in the human mobility networks are cut based on a specific ranking strategy. Some approaches have been proposed to control the risk diffusion dynamics in networks. The node removal approach is straightforward to apply the real world situation, for example shutting down an airport in the airline network, but this approach needs to close all the routes connecting to the removed node, which might not be a cost-effective manner. Therefore, the link removal approach is employed. In this section, the four strategies to rank the candidates for link removal, namely i) link-betweenness-based removal, ii) link-weights-based removal, iii) weighted link-salience-based removal and iv) random-choice-based removal are employed. Then, the links are cut as following one of these four ranking strategies and observe the evolution of the average of the normalized VI in each human mobility network as increasing the fraction of the link removal.

The link salience proposed by Grady et al. (2012) is one of the measures to quantify the relative importance of each link [39]. The link salience is computed as follows, at the first step, the effective distance of each link which is an inverse of the link's weight is computed. Then, we can search the shortest pass from an arbitral node to another node as minimizing the sum of the effective distance on the path. This enables us to compute the shortest path tree from an arbitral node to all the other nodes. In this step, even if any links were chosen several times in a shortest path trees, we do not consider the overlaps of the links (i.e. every shortest path tree can be represented by the N×N adjacency matrix). After we computed the shortest path trees from every node and summing up the shortest path tree matrix, the link salience s on each link is obtained by dividing by the number of nodes. Fig. 8 shows the distribution (histogram) of the link salience in the JDHMN, JDAN and WAN. As can be seen, about 50% of links in JDAN, WAN, and JDHMN of ship are found to be no salience with 0, and about 20% of links have salience of 1. This means that, in these networks, the 20% of links are always chosen in the shortest path trees, meanwhile 0% of links are not used at all. In the networks of JDHMN-Full, JDHMN-Car, JDHMN-Train, and JDHMN-Bus, about 90% of links has the link salience of 0 and only a few percent of links shows the link salience of 1, which means that only a few nodes with high salient links are very critical to attain the effective connection in the networks. Fig.9 shows the distribution of link salience on a map. The color gradation of the links shows the significance. Then, the evolution of the averaged normalized vulnerability index $\mathbf{v}'_{\mathrm{ave}}$ with each link removal strategies is examined as changing the link removal ratio from 0.01 to 0.09 by 0.01. Fig. 10 shows the comparison results of the link removal tests for JDHMN and JDAN when β=0.5 and δ=1.

TABLE III. THE CORRELATION COEFFICIENTS BETWEEN THE SIX WEIGHTED CENTRALITY MEASURES AND RELATIVE VULNERABILITY INDEX V'VI FOR EACH HUMAN MOBILITY NETWORKS

Weighted centrality	Correlation coefficient	JDHMN					JDAN	WAN
		Full	Train	Bus	Ship	Car		
$\mathbf{w}^{in}$	Pearson	0.9425	0.9428	0.9111	0.5920	0.8973	0.9432	0.9501
	Spearman	0.8036	0.8976	0.6630	0.8892	0.8191	0.9603	0.8100
$\mathbf{w}^{out}$	Pearson	0.9346	0.9430	0.8444	0.5960	0.7016	0.9445	0.9500
	Spearman	0.8047	0.8983	0.6649	0.8922	0.8229	0.9579	0.8078
$\mathbf{c}_{\mathrm{EVC}}^{\mathrm{w}}$	Pearson	0.9975	0.9999	0.9965	0.9738	0.9532	0.9999	0.9997
	Spearman	0.9963	0.9978	0.9972	0.8722	0.9776	0.9991	0.9937
$\mathbf{c}_{\mathrm{CC}}^{\mathrm{w}}$	Pearson	0.9429	0.9685	0.9282	0.4465	0.7012	0.8320	0.6261
	Spearman	0.7818	0.9123	0.8289	0.9034	0.8439	0.9770	0.9202
$\mathbf{c}_{\mathrm{BC}}^{\mathrm{w}}$	Pearson	0.5193	0.4846	0.3289	0.0796	0.2875	0.8755	0.6799
	Spearman	0.4712	0.0872	0.1459	0.5880	0.4701	0.2441	0.2456
$\mathbf{c}_{\mathrm{PR}}^{\mathrm{w}}$	Pearson	0.8337	0.8432	0.6993	0.5483	0.6405	0.9304	0.8896
	Spearman	0.4873	0.4626	0.2905	0.8470	0.3055	0.7579	0.4349

The plots in these figures show the relative averaged v'_{ave} to the maximum averaged v'_{ave} for each fraction of link removal. As shown in these figures, in the JDHMN, the strategy ranking with the link weight is the most effectively reduce the averaged v'_{ave}. Only the 1%-removal of the links reduced the averaged v'_{ave} by 74.3% in average.

The strategy of random removal is the worst strategy which reduces the v'_{ave} almost linearly to the fraction of link removal. However, in JDHMN_Car and JDHMN_Ship, the effectiveness of the 1%-link removal are only 72.1% and 28.6%, which can be considered because these two networks have comparatively higher optimal modularity as shown in Table.1.

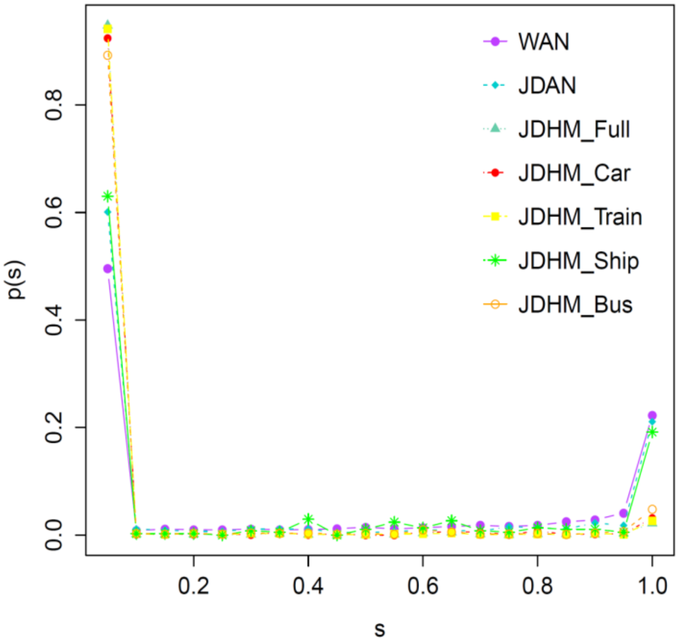


Fig. 8. Distribution of the local weight assortativity δ_i^w for WAN.of the link salience

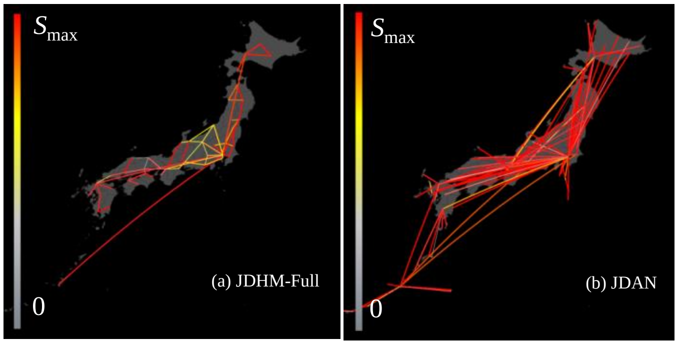
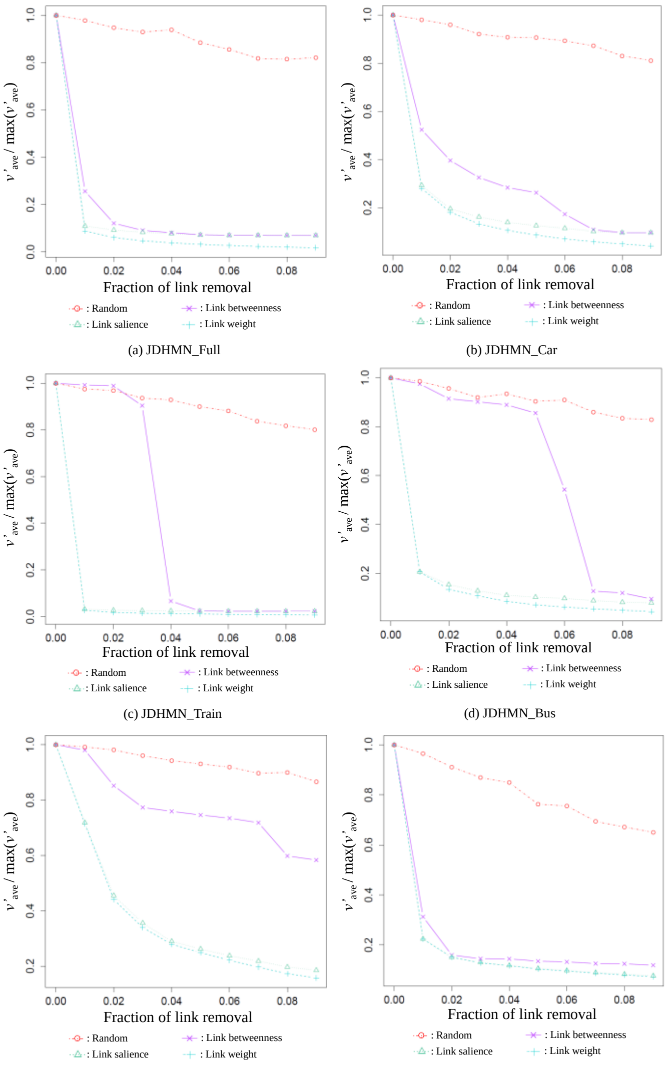


Fig. 9. Distribution of the link salience on a map for (a) JDHM-Full and (b) JDAN. The color gradation in (a) and (b) represents the significance of the link salience on the links

IV. CONCLUSION

In this paper, the network for each mean of the JDHMN as well as JDAN and WAN are analyzed firstly. The link weight distribution shows the power low with the power exponent of 1.3±0.1 in the JDHMN. Also, the community detection results that the community structures found in the human mobility networks except the JDAN. Moreover, the local weighted assortativity are defined and computed for each human transportation network. We also model the risk diffusion dynamics based on the SIS epidemic model and the simulation results show that the vulnerability index on each node has the strong positive correlation with the weighted eigenvector centrality. Furthermore, the link removal tests are implemented with comparing the link removal strategy. The comparison results of the effects of 1%-link removal show that, when we employ the link removal strategy based on the links' weights, the average risk can be reduced by 74.3% in average in the JDHM, but only 72.1% and 28.6% in JDHMN_Car and JDHMN_Ship both of which have strong modular structures. For future works, we need to carefully consider the cost of link removal to model more practical scenarios. Also, the human mobility networks are not static network but the dynamically evolving networks. Therefore, it will be necessarily to care about the influence from the temporal evolution of networks.



(a) JDHMN_Full

(b) JDHMN_Car

(c) JDHMN_Train

(d) JDHMN_Bus

(e) JDHMN_Ship

(f) JDAN

Fig. 10. Comparison results of the link removal tests when β=0. and δ=1 for (a) JDHM-Full, (b) JDHM-Car, (c) JDHM-Train, (d) JDHM-Bus, (e) JDHM-Ship, and (f) JDAN. The plots show the averaged v'_{ave} divided by the maximum averaged v'_{ave} as changing the fraction of link removal from 0.01 to 0.09 by 0.01

References

[1] V. Latora, M. Marchiori, "Is the Boston subway a small-world network?," Physica A, vol.314, no.1–4, pp.109–113, 2002.

[2] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, S. S. Manna, "Small-world properties of the Indian railway network," Phys. Rev. E, vol.67, no.3, pp.036106, 2003.

[3] J. Sienkiewicz, J. Hołyst, "Statistical analysis of 22 public transport networks in Poland," Phys. Rev. E, vol.72, no.4, pp.1–11, 2005.

[4] A. De Montis, M. Barthélémy, A. Chessa, A. Vespignani, "A The structure of interurban traffic: a weighted network analysis," Environment and Planning B: Planning and Design. vol.34, no.5, pp.905–924, 2007.

[5] W. Li, X. Cai, "Empirical analysis of a scale-free railway network in China," Physica A, vol.382, no.2, pp.693–703, 2007.

[6] H. Soh, S. Lim, T. Zhang, X. Fu, G. K. K. Lee, T. G. G. Hung, P. Di, S. Prakasam, L. Wong, "Weighted complex network analysis of travel routes on the Singapore public transportation system," Physica A, vol.389, pp.5852–5863, 2010.

[7] L. A. N. Amaral, A. Scala et al., "Classes of small-world networks," Proc. Natl. Acad. Sci. USA, vol.97, no.21, pp.11149-11152, 2000.

[8] R. Guimera, L. A. N. Amaral, "Modeling the world-wide airport network," Eur. Phys. J. B, vol.38, no.2, pp.381-385, 2004.

[9] R. Guimera, S. Mossa et al., "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," Proc. Natl. Acad. Sci. USA, vol.102, no.22, pp.7794-7799, 2005.

[10] O. Woolley-Meza1, C. Thiemann, D. Grady, J. J. Lee, H. Seebens, B. Blasius, and D. Brockmann, "Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements," Eur. Phys. J. B, vol.84, pp.589–600, 2011.

[11] G. Bagler, "Analysis of the airport network of India as a complex weighted network," Physica A, vol.387, no.12, pp. 2972-2980, 2008.

[12] J. Wang, H. Mo et al., "Exploring the network structure and nodal centrality of China's air transport network: A complex network approach," Journal of Transport Geography, vol.19, no.4, pp.712-721, 2011.

[13] M. Hossain, S. Alam, T. Rees, H. Abbass, "Australian Airport Network Robustness Analysis: A Complex Network Approach," Proceeding of Australasian Transport Research Forum 2013, Brisbane, Australia, 2013.

[14] L. Hufnagel, D. Brockmann, T. Geisel, "Forecast and control of epidemics in a globalized world," Proc. Natl. Acad. Sci. USA, vol.101, no.42, pp.15124-15129, 2003.

[15] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco et al., "Multiscale mobility networks and the spatial spreading of infectious diseases," Proc. Natl. Acad. Sci. U S A, vol.106, pp.21484–21489, 2009.

[16] V. Colizza, R. Pastor-Satorras, A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," Nat. Phys., vol.3, no.4, pp.276–282, 2007.

[17] W. V. D. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, "The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale," BMC Infectious Diseases, vol.11, no.37, 2011.

[18] D. Balcan et al., "Modeling the spatial spread of infectious diseases: the global epidemic and mobility computational model," J. Comput. Sci., vol.1, pp.132–145, 2010.

[19] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley et al., "Strategies for mitigating an influenza pandemic," Nature, vol.442, pp.448–452, 2006.

[20] L. A. Rvachev, I. M. Longini, "A mathematical model for the global spread of influenza," Mathematical biosciences, vol.75, no.1, pp.3-22, 1985.

[21] A. J. Tatem, D. J. Rogers, S. I. Hay, "Global Transport Networks and Infectious Disease Spread," In: S. I. Hay, A. Graham and D. J. Rogers, Editor(s), Advances in Parasitology, Academic Press, vol. 62, Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications, pp. 293–343, 2006

[22] D. Brockman, D. Helbing, "The hidden geometry of complex, network-driven contagion phenomena," Science, vol.342, no.6164, pp.1337–1342, 2013.

[23] S. Eubank et al., "Modelling disease outbreaks in realistic urban social networks," Nature, vol.429, pp.180–184, 2004.

[24] K. J. Mizgier, P. J. Matthias, S. M. Wagner, "Bottleneck identification in supply chain networks," International Journal of Production Research, vol.51, no.5, pp.1-14, 2013.

[25] K. Marcelino, M. Kaiser, "Critical paths in a metapopulation model of H1N1: Efficiently delaying influenza spreading through flight cancellation," PLoS Currents Influenza 2012, Apr 23, 2012.

[26] T. Verma, N. H. H. Araujo, "Revealing the structure of the world airline network," Sci. Rep, vol.4, p. 5638, 2011.

[27] M. Rosvall, C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," Proc. Natl. Acad. Sci. USA, vol.105, p.1118, 2008.

[28] A. Lancichinetti, S. Fortunato, "Community detection algorithms: a comparative analysis," Phys. Rev. E, vol.80, p.056117, 2009.

[29] M. E. J. Newman, "Assortative mixing in networks," Phys Rev Lett, vol.89, no. 20, p.208701, 2002.

[30] C. C. Leung, H. F. Chau, “Weighted assortative and disassortative networks model,” Physica A, vol.378, no.2, pp.1591–602, 2007.

[31] M. Piraveenan, M. Prokopenko, A. Y. Zomaya, “Local assortativeness in scale-free networks,” Europhysics Letters, vol.84, no.2, p.28002, 2008.

[32] M. Piraveenan, M. Prokopenko, A. Y. Zomaya, “Classifying complex networks using unbiased local assortativity,” Proc. of the Alife XII Conference, Odense, Denmark, 2010.

[33] G. Thedchanamoorthy, M. Piraveenan, D. Kasthuriratna, U. Senanayake, “Node assortativity in complex networks: an alternative approach,” Procedia Computer Science, vol.29, pp.2449–2461, 2014.

[34] R. Albert, A. L. Barabasi, “Statistical mechanics of complex networks,” Rev. Mod Phys., vol.74, pp.47-94, 2002.

[35] S. N. Dorogovtsev, J. F. F. Mendes, “Evolution of networks,” Advances in Physics, vol.51, pp.1079-1187, 2002.

[36] M. E. J. Newman, “The structure and function of complex networks,” SIAM Review, vol.45, pp.167-256, 2003.

[37] S. N. Dorogovtsev, A. V. GoltsevV, J. F. F. Mendes, “Critical phenomena in complex networks,” Rev. Mod. Phy., vol.80, p.1275, 2008.

[38] R. Pastor-Satorras, A. Vespignani, “Epidemic dynamics and endemic states in complex networks,” Phys. Rev., vol. 63, p.066117, 2001.

[39] D. Grady, C. Thiemann, D. Brockmann, “Robust classification of salient links in complex networks,” Nature Communication, vol.3, p.864, 2012.

Detection and Defense Against Packet Drop Attack in MANET

Tariq Ahamad
College of Computer Engineering & Sciences
Prince Sattam Bin Abdulaziz University, Saudi Arabia

***Abstract*—MANET is a temporary network for a specified work and with the enormous growth MANETs it is becoming important and simultaneously challenging to protect this network from attacks and other threats. Packet drop attack or gray hole attack is the easiest way to make a denial of service in these dynamic networks. In this attack the malicious node reflects itself as the shortest path and receives all the packets and drops the selected packets in order to give the user the service that that is not correct. It is a specific kind of attack and protects the network and user from detecting this malicious activity. In this article I have proposed an efficient for step technique that confirms that this attack can be detected and defended with least efforts and resource consumption.**

Keywords—MANET; gray hole; DoS; packet drop; security

I. Introduction

MANET (Mobile Adhoc Network) is a dynamic mobile network that can exist and can be formed without any predefined and preexisting network and communication network1 .The concept of Ad Hoc network depends on the availability of the devices that are to be connected to each other to form the network2. Thus, unlike other existing and traditional networks, these networks do not depend on any pre-existing network or infrastructure to carry out their operations and their this dynamic character reduces their cost and implementation time.

The backbone of the Ad Hoc Network is the routing protocols that enable multi – hop data transfer or communication in these networks[3]. Since the topology of these dynamic networks keep on changing so changes the attacks on these networks and in order to deal with these malicious attacks these routing protocols must be robustic[4]. The pre-existing routing protocols easily deal with changing topologies but the malicious attacks always remain the issue to be fixed. In this article I have evaluated the robustness of existing routing protocols against the malicious attacks and assess the quality and impact of security improvements.

II. Threats in Ad Hoc Networks

Reliability of the devices or nodes that are to be used to form an Ad Hoc network is most important concept to be kept in mind as devices or nodes act both as computers and routers. Since the topology keep on changing due to dynamic behavior of the network, this change is supported by routing protocols so as to establish the dynamic routes[5] . Since routing information is very sensitive and can be targeted by the attackers in order to harm the network or the applications running in the network as illustrated in the figure 1.

Since all the Ad Hoc Networks thoroughly depends on Routing protocols, there are many sources that make use of this idea and attack them and the two major sources are:

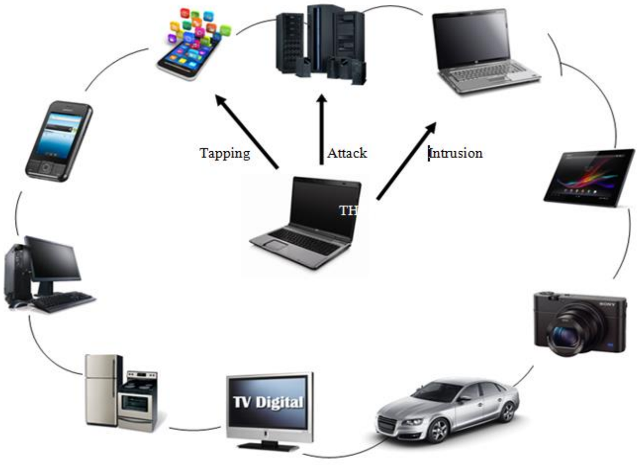


Fig. 1. Attack in Ad Hoc Network

1) As per the basic cyber attack practice, the first comes from explicit attackers. By inserting a new large pool of routes or using old routing information or distracting the current routing pool, an intruder can divide the network or delay the traffic and can cause inefficient routing and affect the quality of service (QoS).

2) The most dangerous and that can cause severe effect to the dynamism and reliability of Ad Hoc Network comes from inside the network by gray nodes or gray hole (compromised nodes) and can exploit the routing information of the other nodes and can affect the service as they are the part of the network.

III. Packet Drop Attack (Gray Hole)

Packet drop attack (gray hole) is a denial of service (DoS) attack in which a router relays or drops data packets instead of discarding for a specific network destination at specific time – a packet after every n number of packets or after every t number of seconds[6]. It is slightly different from black hole as black hole is a general denial of service (DoS) attack that drops packets as its key constraints are very specific. It is an active attack that leads dropping of packets[7]. The attacking node at first agrees to forward data packet or messages then fails to do so and starts behaving like a malicious node[8]. At first the attacker node behaves normally and replies true route replies(RREP) messages to other nodes to invoke route request (RREQ) messages and accepts or takes the sending packets and finally drops few or all packets to launch denial of service (DoS) attack[9]. If nodes in the neighborhood try to send data packets over attacking or victim nodes lose connection to target or destination node or network and may want to discover or rebuild a route again by broadcasting route request (RREQ) messages. Attacking node send route reply (RREP) messages to establish route[10]. This process doesn't stop until attacking node achieves its goal like battery power consumption, bandwidth consumption etc.

IV. Proposed Mechanism

We will start by making some assumptions and are going to be considered for formulating network model and later present the complete details of the proposed system.

A. *Basic network model*

The first thing that we are going to consider is assuming that a MANET (Mobile adhoc network) consists of almost similar types of devices. Every device may travel aimlessly or stay immobile in a specific location for a temporary slot of time. Also every device may leave or join the network or even fail at any instance of time. The MANET (mobile adhoc network) follows peer to peer networking principals over fixed shared bandwidth and multihop wireless nodes. Assuming a non-zero ID for each node to differentiate between them and all the channels and links in the MANET (mobile ad hic network) to be bidirectional. The proposed technique doesn't make any assumption malicious mode operations of the wireless nodes interface as compared to current security frameworks. The malicious node may not only experience or face extra computation and power consumption in processing the moving data packets, but also will not be effective where devices have equipped directional antennas. The number of packet drop nodes may vary at different instances of time in the MANET (mobile adhoc network) and may disturb or decline the MANET communication by cooperating with each other.

B. *Modules of the proposed mechanism*

My Suggested technique will use two detection procedures i.e., local and cooperative detection models to recognize malevolent node (grayhole) in MANET (Mobile Ad Hoc Network. The moment malevolent node is recognized and confirmed the mechanism has a notification procedure added that sends a message to all the nodes , so as to identify the malevolent node and isolate the malevolent node and make sure that it is not allowed any access to anypart of the MANET and its resources.

My mechanism is a four step scheme and all the four steps are invoked sequentially. Following are the four steps.

1) *Multihop Data Collection (MDC)*
2) *Local Anomaly Detection (LAD)*
3) *Collective Anomaly Detection (CAD)*
4) *MANET Alarm*

C. *Multihop Data Collection (MDC)*

Every node in the MANET gathers packet forwarding data in its surrounding multihop zone and saves that in the Data Routing Information Table (DRIT). Figure 2, shows DRIT of node 5 and the numbers used in DRIT shows that node 5 maintains data routing information of neighboring nodes 4, 6, 7, 8, 9. As per table 1, in row one column "from" indicates as node 5 has sent the packet received from corresponding one and column "thru" from same row indicates that node 5 has sent the data packet to that node. So, node 5 neither received nor sent any packet towards node 4 as mentioned in the Table 1. However, node 5 forwarded and collected data packet from and to node 6. So, following this approach each node creates and maintains a DRIT. After a fixed time interval, every node recognizes its hop nodes with which it hasn't been involved for data communication and calls on a detection procedure to investigate them further. This investigate is done on those nodes which have 0 (zero) entries in both the columns i.e., "from" and "thru" in DRIT. Thus as per table 1, node 5 invokes local detection scheme for node 4. In row one of DRIT with column "RTS/CTS", the ration "RTS/CTS" gives an approximate idea regarding the amount of entreaty approaching for communication and amount of data packets transmission that the selected node is executing in real time. The importance and use of "checkbit" in Data Routing Table is explained in the next step.

D. *Local Anomaly Detection (LAD)*

This method or mechanism is initiated by a node after recognizing and confirming a node as suspicious by inspecting DRIT. Initiator node initiates or invokes the LAD procedure and chooses cooperative node in the neighboring nodes after checking its DRIT and cybercasts a RREQ (route request) packet to its first-hop nodes, seeks route thru cooperative node. The initiator node will receive a good amount of RREP (route reply) packets from its multihop area to its RREQ (route request) message and surely will receive and experience a RREP (route reply) from the doubtable one as well, if it really is a packet drop node. Once RREP is sent by the suspected node and the moment it is received from the suspected node (SN), the initiator node sends enquiry or probe data packet to the cooperative node (CN) thru the suspected node (SN) and enquires the cooperative node whether it received the enquiry packet or not , right after given interval of time i.e. time to live (TTL). The initiator node updates its data routing information table by using adding 1 (under "chekbit" against suspected node's NID), right after it receives confirmation that the cooperative node has received the enquiry data packet. In case , if the enquiry or probe packet does not reach the cooperative node then the initiator node rises its degree of intuition about suspected node and invokes the (CAD) co-operative anomaly detection scheme.

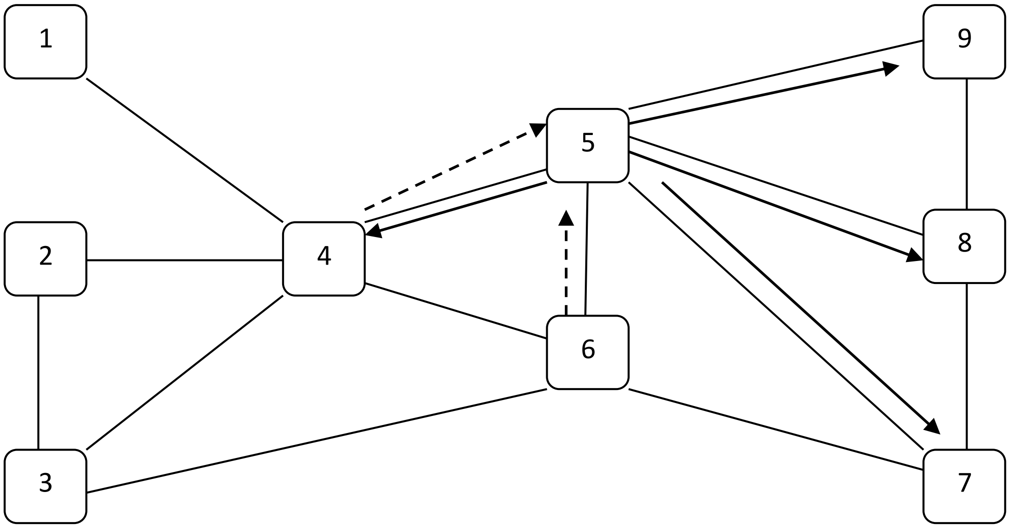


Fig. 2. MANET Topology

TABLE I. DATA ROUTING INFORMATION TABLE (DRIT)

NID	From	Thru	RTS/CTS	Check Bit
4	0	0	15	0
6	1	1	5	1
7	0	1	3	0
8	1	0	6	1
9	0	1	4	0

In figure 1, initiator node, Node 5 invokes LAD procedure for the suspected node (SN) 4 and selects node 6 as the cooperative node because both the entries of node 6 are 1 under "from" and "thru" columns and becomes most reliable and trustworthy node for node 5. Node 5 cybercasts a RREQ (route request) packet to all its 1-Hop nodes i.e. 4, 6, 7, 8, 9 and request for a route to cooperative node Node-6. After receiving a route reply (RREP) from suspected node 4, node 5 sends an enquiry packet to the node 6 via node 4 and confirms from node6 about probe packet. If node 6 confirms that it received the enquiry packet then node 5 makes an entry and adds 1 under "checkbit" column corresponding to node 4 in DRIT. And in case node 6 doesn’t confirm on the arrival of enquiry packet then node 5 initiates the cooperative detection anomaly scheme.

E. Cooperative Anomaly Detection (CAD)

This mechanism increases the detection influence by the decreasing chance of false and fake identification of local anomaly detection (LAD) scheme. The CAD mechanism is initiated whenever an initiator node notice that the enquiry data packet didn’t reach the cooperative node via suspected node. The initiator node initiates cooperative identification scheme (process) and broadcast a CAD request packet to all the 1-hop nodes of the suspected node. When the neighboring nodes of the malicious suspected node accepts the cooperative identification request packet then each of the neighboring nodes sends route request (RREQ) to the suspected node seeking a route to initiator node. Once the suspected node reacts with a Route reply (RREP) message, every node forwards a "further-enquiry-packet" to the initiator node including the same route. This route definitely will include suspected node because suspected node is 1-Hope (neighbor) of every requesting node even the initiator node. Each neighboring node of suspected node (except initiator node) now informs the initiator node that one more packet called "further-enquiry-packet" has already been forwarded to it and this alerting packet from every neighboring node is forwarded towards the initiator node thru the routes that do not involve node. This step is extremely important to assure that suspected node is not aware of this ongoing process of cross check. The initiator node will receive a lot of further-enquiry-packets and alerting packet. The initiator node prepares a Probe-Check-Table and will have only two fields i.e. NID (Node ID) and PS (ProbeStatus). NID field will have identifiers of nodes from which it have received the notification message. Entry "1" is put under the PS (ProbeStatus) communicating to the nodes which sent Further-Enquiry-Packet to initiator node as shown in Table II.

TABLE II. NID=NODEID, PS=PROBE STATUS

NID	PS
6	0
7	1
8	1
9	1

If the suspected node behaves like a malicious node or packet drop node, it is kept away and secluded from MANET by initiating the global-alarm-detection procedure. The frequency of invoking the detection and identification procedure is key factor for assuring the expected output in the MANET because packet drop node can shift its state from good to bad frequently. The frequency of the invoke calls of the identification procedure must be prepared on highest number of data packet drops that the MANET app tolerates. In the worst case scenario, malicious node shifts its state from nice to worst right after the invoked wound of detection algorithm is done and can return back to nice state before the next invoke call. Although these situations are rare, the invoke call frequency must be calculated on the approximation of the amount of data packets dropped by packet drop node during that time slot and the highest value of packet drops that is applicable to maintain the expected and sought QoS (Quality of Service).

F. MANET Alarm

This scheme is initiated to form or create a "network-wide-notification-system" in order to send alarm packet to all the nodes and devices in Mobile Adhoc Network (MANET) about the (malicious node) packet drop node that has been detected by CAD scheme. It also certifies that none of the network resources or services to be allowed to these malicious nodes and are kept isolated from the rest of MANET (Mobile Adhoc Network).

A security problem arises after the identification and isolation procedures of suspected nodes. A set of malicious nodes (gray hole) can collaborate to hurl a malign attack by falsely incriminating legal node and segregate it from the MANET (simple isolates a legitimate node). To prevent this in the MANET. I suggest a procedure that is somehow similar to existing thresh hold cryptography. In my proposed procedure, when a cooperative-detection-procedure identifies and confirms a SN to be gray hole initiated by a node, broadcast an alarm message digitally signed using its private key. The complete sign is created only when at least “n” number of nodes put their signs into the alarm message. The suspected node (malicious node) is kept away from the MANET after the alarm message is verified and authenticated with full signature. Thus our proposed mechanism is strong and feasible against collusion that involves maximum n-1 malicious nodes (gray hole) in an area inside MANET. Once the node is identified and confirmed as malicious node, its NID (node ID) is entered into “Malicious_node_list” a global list file of malicious node. This Malicious_node_list is broadcasted in the MANET periodically whenever an update is made to it. The Malicious_node_list can be adjoined with the routing message RREQ and RREP. So that there must not be any extra overhead. On the other hand, every node may keep a partial record of faulty nodes which are in its 1-Hop neighborhood. This existing partial record must change and update whenever its neighborhood changes. Since the nodes require to know the whereabouts of its multihop nodes for routing only, this procedure will be best fit for protocols, AODV in particular.

V. Conclusion

In my research article, I have proposed a reliable and efficient mechanism for detecting packet drop attack in Mobile ad hoc networks (MANET). Due to their dynamic phase shifting character, it is difficult to detect them. My proposed technique will boost the reliability by presciently initiating a cooperative scheme that involves neighboring nodes of malicious node. Suspect and detection decision are done with the help of consensus algorithm that is based on thresh hold cryptography. The proposed mechanism is efficient and effective with controlled overhead and great detection rate.

Acknowledgement

This research was funded and conducted at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia during the academic year 2015/2016 under research number 2015/01/3850.

References

[1] Ahamad T, Aljumah A. "Detection and Defense Mechanism against DDoS in MANET", Indian Journal of Science and Technology. 2015 Dec Vol 8(33).

[2] Aljumah A, " Detecting Distributed Denial Of Service (Ddos) Attack Using TTLv Constraint In Mobile Adhoc Networks (MANET) ", Science Internationals, 2015 Dec Vol 27(6),5037-5040.

[3] Ahamad T, Aljumah A. ,"Ad Hoc Network & Black Hole - Threat and Solution". American Journal of Scientific Research , Issue 104 , Nov, 2014.

[4] Uddin M, Alsaqour R, Abdelhaq M. Intrusion detection system to detect DDoS attack in gnutella hybrid P2P Network. Indian Journal of Science and Technology. 2013 Feb; 6(2):71–83.

[5] Abdelhaq M, Hassan R, Ismail M. A study on the vulnerability of AODV routing protocol to resource consumption attack. Indian Journal of Science and Technology. 2012 Nov; 5(11):3573–7.

[6] Ahamad T, Aljumah A," Hybrid Approach Using Intrusion Detection System", International Journal of Engineering Research & Technology, Vol. 3 Issue 2, February - 2014

[7] Tsou PC, Chang JM, Lin YH, Chao HC, Chen JL. Developing a BDSR scheme to avoid black hole attack based on proactive and reactive architecture in MANETs. 13th International Conference on Advanced Communication Technology: Seoul; 2011 Feb 13-16. p. 755–60.

[8] Baadache A, Belmehdi A. Avoiding black hole and cooperative black hole attacks in Wireless Ad hoc Networks, International Journal of Computer Science and Information Security. 2010; 7(1):10–6.

[9] Arunmozhi S.A., Venkataramani Y."A Flow Monitoring Scheme to Defend Reduction-of- Quality (RoQ) Attacks in Mobile Ad-hoc Networks", Information Security Journal: A Global Perspective, Vol.19, No.5, 2010, pp.263- 272.

[10] Hyojin K, Ramachandra B. C., JooSeok S,"Novel Defense Mechanism against Data Flooding Attacks in Wireless Ad Hoc Networks", IEEE Transactions on Consumer Electronics, Vol. 56, No. 2, May 2010, pp. 579-582.

[11] Mistry N, Jinwala D. C., Zaveri M,"Improving AODV Protocol against Blackhole Attacks", Proceedings of the International Multiconference of Engineers and Computer Scientist, Hong Kong, Vol. II, 2010.

The Enhanced Arabchat: An Arabic Conversational Agent

Mohammad Hijjawi
Faculty of Information Technology
Applied Science Private University
Amman, Jordan

Zuhair Bandar
School of Computing
Manchester Metropolitan University
Manchester, UK

Keeley Crockett
School of Computing
Manchester Metropolitan University
Manchester, UK

***Abstract*—The Enhanced ArabChat is a complement of the previous version of ArabChat. This paper details an enhancements development of a novel and practical Conversational Agent for the Arabic language called the “Enhanced ArabChat”. A conversational Agent is a computer program that attempts to simulate conversations between machine and human. Some of lessons was learned by evaluating the previous work of ArabChat . These lessons revealed that two major issues affected the ArabChat’s performance negatively. Firstly, the need for a technique to distinguish between question and non-question utterances to reply with a more suitable response depending on the utterance’s type (question and non-question based utterances). Secondly, the need for a technique to handle an utterance targeting many topics that require firing many rules at the same time. Therefore, in this paper, the “Enhanced ArabChat” will cover these enhancements to improve the ArabChat’s performance. A real experiment has been done in Applied Science University in Jordan as an information point advisor for their native Arabic students to evaluate the Enhanced ArabChat.**

Keyword—Artificial Intelligence; Conversational Agents and Arabic

I. Introduction

From Turing test (imitation game) time[1], which he was tried to solve his test by answering his question “if a computer could think, how could we tell?”, number of researches tries to solve his test by developing a conversational agent. A Conversational Agent(CA) is a computer program that attempts to simulate conversations between machine and human [2]. Since that time, number of CAs types has been raised due to the diversity of applications that’s could CAs applied in. This is including Embodied CA, Linguistic CA and mixed approach between them [3].

The Embodied type has a humanoid or animated character which shows a body reactions such as facial expressions, eyes movement and the character sounds [3]. Linguistic CAs deals with spoken or/and written conversations without to embed the embodied abilities. Finally, the mixed approach which can share the features of both types [3].

This paper is interested to build a linguistic CA. Therefore, the main approaches that be used to build such types of CAs will be introduced which are Natural Language Processing (NLP), Semantic Sentence Similarity (SSS) measures and Pattern Matching (PM) [4].

The NLP which is defined in computing as “the computational processing of textual materials in natural human languages” [5] is based on understanding a sentence. Technically, NLP-based CAs uses grammar rules and a list of attribute/value pairs to extract the conversation’s speech act type from the sentence [6]. Then, it uses these extracted information to fill a template-based response [6]. However, extraction such information is not easy at all as it depends on many linguistic factors [6]. In a rich language especially the sematic languages such as Arabic, this extraction will be harder to process [4, 6].

The SSS approach is based on checking the similarity level in semantic between two sentences [3]; the first sentence is the conversation itself and the second is a scripted pattern inside the CA. The most closed pattern in semantic (meaning), its response will be replied as an answer to the conversation. The SSS approach is based on computational semantic based manual built databases such as WordNet [3]. However, such database established in 2006 [3], and the research in SSS, in general, is still a young research area in the Arabic language [4].

The PM approach is the most common used approach for its simplicity and because it is language independent [4]. It does not need complex pre-processing stages like the previous approaches, so it is not expensive computationally. Consequently, a number of CAs such as [4, 7-9] used this approach to handle conversations for applications deal with large numbers of users in a real-time environment like the Internet [10]. Basically, this approach based on matching a conversation with a pre-structured patterns to find the suitable one. Then, the response that related to the best matched pattern will be replied [4]

All the three approaches have advantages and disadvantages that can be cleared in different references such as [3, 4, 8, 11-14]. However, as mentioned above, this paper is considered as a complement work for the previous edition of ArabChat [4] to enhance its performance and called the “Enhanced ArabChat”. The rest of this paper describes the enhancements details. The next sections describe the Enhanced ArabChat framework and the conducted experiment and its results. Finally, a general evaluation has been conducted to evaluate the proposed CA.

II. The Enhanced ArabChat

The previous work of ArabChat [4] considered the first phase of ArabChat development. In this paper, number of novel modules has been integrated into the ArabChat to lunch the second phase of it to be the “Enhanced ArabChat”.

Before proceeding with the enhancement explanation, it is important to summarise the first edition of ArabChat framework. The ArabChat is PM approach which based on pattern matching technique to handle the Arabic textual user conversations.

The ArabChat is a rule-based CA modelled into three main modules which are scripting language, engine and brain [4]. The scripting language is a predefined language used to script an application’s domain in order to represent it. The scripting language is structured as a rule-based language that contains contexts (main domain topics) which each context has several rules (sub-domain topic) and each rule has number of patterns (the simulated user sentence) and responses. While, the brain is a structured knowledge base that is used to store the domain’s scripts. Finally, the engine handles user’s utterances (conversations) by matching them with the scripted domain and replying with a suitable response. The conversation remains ongoing until one of the conversation’s parties (user and ArabChat) terminates it. The ArabChat has the right to terminate the conversation and close a session for many reasons described in [4]

As discussed before, number of lessons have been learned by evaluating the previous work of ArabChat [4]. These lessons revealed that two major issues affected the ArabChat performance negatively. Firstly, the need for a technique to distinguish between question and non question utterances in order to reply with a more suitable response depending on the utterance’s type. Secondly, the need for a technique to handle an utterance targeting many topics that require firing many rules at the same time. For instance, ArabChat has two rules to deal with two different topics which are “Accommodation” and “Transportation” in Jordan, and the user targets the two topics in the same utterance like “how much is the cost for the student accommodation in Jordan and how much is the average cost for the transportation as well”. ArabChat was unable to reply for both topics (the rule that has the best matched pattern will be fired). Therefore, these issues have been taken into consideration in order to improve ArabChat‘s performance and continue developing to generate the Enhanced ArabChat. These two issues are related to the ArabChat engine’s work but also they need amendments on the rule’s structure (scripting language and brain) in order to meet these engine-based improvements. Therefore, it can be summarised that all the new required amendments can be classified as engine-based amendments and scripting language-based amendments.

The engine-based amendments deal with developing the two required modules. Firstly, the need for a technique to distinguish between question and non question utterances. Therefore, in the Enhanced ArabChat, the module “Utterance Classification” which deals with this issue will be developed. The second module is to handle an utterance requiring the firing of many rules at the same time. Therefore, the module “Hybrid Rule” which deals with this issue will be developed in the Enhanced ArabChat.

The scripting language-based and brain amendments summarised by the need of adding some new features to the rule’s structure in order to meet the requirements of the new amendments of the Enhanced ArabChat engine which are the “Utterance Classification” module and the “Hybrid Rule” module. In addition, other features of the rule’s structure will be added in order to facilitate evaluating the Enhanced ArabChat.

III. The Enhanced ArabChat Framework

The Enhanced ArabChat framework is a complement of the first version of ArabChat framework. Consequently, The Enhanced ArabChat includes all of the developed modules in the first version of ArabChat and the new integrated developed modules (“Utterance Classification” and “Hybrid Rule”) as Figure 1 describe. The new two developed modules ((“Utterance Classification” and “Hybrid Rule”) had been published in two different papers in [15, 16] that related to the same research work for building the Enhanced ArabChat.

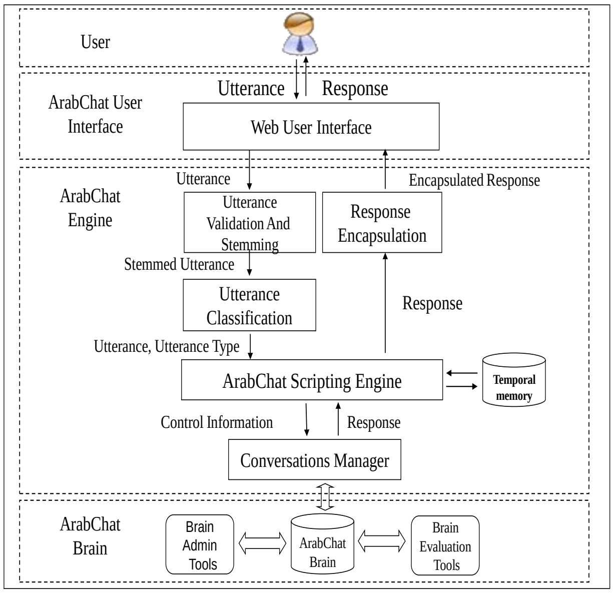


Fig. 1. The Framework of the Enhanced ArabChat

A. The Enhanced ArabChat Scripting Language

The scripting language of the Enhanced ArabChat is a complement to the scripting language of the previous version of ArabChat [4]. The new modules in the Enhanced ArabChat require amending the rule’s structure. Therefore, only the new amendments on the rule’s structure will be described in this section. In the Enhanced ArabChat, a rule has three new parameters, which are:

- “نوع القاعدة” “Rule type”.
- “تتعارض مع” “Conflict With”.
- “المعلومات المطلوبة” “Information Requirements”.

The first parameter (“Rule Type”) has been added due to the new module “Utterance Classification”. This module classifies the utterance into question and non question utterances. The question utterance requires firing question-based rule. While, the non question utterance requires firing non question-based rule. As a result, each rule in the Enhanced ArabChat has its own type either a question-based or non question-based rule.

The second parameter (“Conflict With”) has been added due to the new module “Hybrid Rule”. This module deals with an utterance requesting different information that requires firing many rules. In reality, some rules (topics) may be conflicting with each other. These conflicting rules should not be fired together for the same utterance. If so, the rule that has the highest strength will fire. The parameter “Conflict With” is used to alert the engine that the rule has raised a conflict with other rules, which prevents the engine from firing them together.

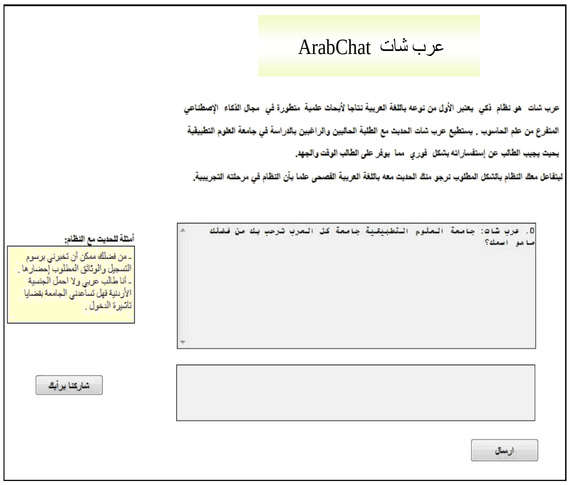
The third parameter (“Information Requirements”) has been added to the rule’s structure for the new ArabChat evaluation purposes. A rule’s “Information Requirements” is a collection of words that should exist in an utterance in order to fire the rule. This parameter will be described in details later.

B. The Enhanced ArabChat framework components

In this section, only the new modules in the Enhanced ArabChat will be described. The new module “Utterance Classification” embedded in the Enhanced ArabChat engine where the module “Hybrid Rule” embedded in the Enhanced ArabChat scripting engine. In addition, a little amendment has been done on the web user’s interface in order to let users to evaluate the Enhanced ArabChat by filling a questionnaire as described in the next section.

1) Web User Interface

The Web User Interface (WUI) for the Enhanced ArabChat has the same functions as the previous version of ArabChat WUI. In addition, in this WUI (for the Enhanced ArabChat), an online user questionnaire is added and linked by the button “شاركنا برأيك” “share your opinion” as shown in Figure 2. This questionnaire will be used to evaluate the Enhanced ArabChat. Moreover, there is a box to show the user examples of how they can communicate with the system.



ArabChat عرب شات

عرب شات هو نظام ذكي يعتبر الأول من نوعه باللغة العربية نتاجا لأبحاث علمية متطورة في مجال الذكاء الإصطناعي
المتفرع من علم الحاسوب . يستطيع عرب شات الحديث مع الطلبة الحاليين والراغبين بالدراسة في جامعة العلوم التطبيقية
بحيث يجيب الطالب عن إستفساراته بشكل فوري مما يوفر على الطالب الوقت والجهد.
ليتفاعل معك النظام بالشكل المطلوب نرجو منك الحديث معه باللغة العربية الفصحى علما بأن النظام في مرحلته التجريبية.

أمثلة للحديث مع النظام:
- من فضلك ممكن أن تخبرني برسوم التسجيل والوثائق المطلوب إحضارها .
- أنا طالب عربي ولا احمل الجنسية الأردنية فهل تساعدني الجامعة بقضايا تأشيرة الدخول .

0. عرب شات: جامعة العلوم التطبيقية جامعة كل العرب ترحب بك من فضلك ما هو اسمك؟

Fig. 2. The Enhanced ArabChat WUI

2) The Enhanced ArabChat engine

The Enhanced ArabChat has a new integrated two modules which are “Utterance Classification” and “Hybrid Rule”. As discussed above, only the new modules in the Enhanced ArabChat will be described in this paper.

a) Utterance classification

If the validation process detects a valid utterance, then the utterance classification process will start. The Enhanced ArabChat will first classify the utterance as either a question or non question utterance. This classification is based on a set of rules generated from a top-down decision tree induction technique as described in [15].

The “Utterance Classification” module is fully described in a related research work for ArabChat in [15]. In this section, the need of classifying an utterance will be only described. Consider the following rule “How-travel-to-Amman” which is designed to answer utterances asking about the ways of travelling to Amman:

< How-travel-to-Amman >

a: 0.1

p: * الذهاب إلى عمان *   * go to Amman *

p: الطرق * إلى عمان *   * ways * to Amman

p: الذهاب إلى عمان *   * go * Amman

r: بإمكانك الذهاب إلى عمان من خلال الحافلات العمومية المتواجدة في مجمع الحافلات

(You can go to Amman by public buses that stationed in the buses station)

Consider the two utterances “ كيف أستطيع الذهاب إلى عمان من الزرقاء؟” “How I can get to Amman from Zarqa?” and “ أنا أفضل الذهاب إلى عمان في الصيف” “I like to visit Amman in the summer”. The two utterances matched the same pattern “ * الذهاب إلى عمان *” because they shared the same keywords as the pattern. However, the first utterance is considered as question while the second is considered as non question. The two different utterances require different responses. Therefore, scripting two different types of rules (question and non question) for the same topic is important to deal with question and non question utterances. However, scripting the two rules will not solve the problem properly as their patterns may still share the same keywords. Consequently, the rule that has the highest strength (best matched) will fire [4]. Therefore, adding extra keywords to the question based rules’ patterns is important. These extra keywords might be some interrogative words such as “كيف” “How”. For instance, the pattern might be “* كيف * الذهاب إلى عمان” in order to match the mentioned question utterance.

Scripting all of the expected interrogative words for each question-based rule will increase the number of patterns. Alternatively, the “Utterance Classification” module has been developed. In addition, this module might increase the Enhanced ArabChat performance by replying to an utterance depending on its type (question or non question). For a specific topic, a question-based utterance might need an accurate answer for the question, while a non-question-based utterance might need agreement or disagreement with the user's thoughts.

The "Utterance Classification" methodology itself is already explained in details in a related research work for ArabChat in [15]. However, in this paper [15], a novel technique has been proposed and developed to classify the Arabic sentences into questions and non-questions based sentences. This classification was based on structural information contained in Arabic function words. The developed technique extracts the function words features by replacing them with numeric tokens and replacing each word with a standard numeric token. The extraction process or the classification rules based on building a decision tree and it provides a high effective classification results.

After determining the utterance's type, the utterance and its type are sent to the scripting engine in order to deal with the utterance depending on its type.

b) The Enhanced ArabChat scripting engine

The "Enhanced ArabChat" scripting engine is a complement of the previous version of ArabChat scripting engine. The integration of the "Utterance Classification" module will affect the methodology of the Enhanced ArabChat scripting engine methodology. After integration, the Enhanced ArabChat scripting engine works depending on the classified utterance's type as depicted in Figure 3. When the utterance is classified as question, the engine explores the question-based rules as depicted in Figure 4. Contrarily, if the utterance is non-question, the engine deals with non question-based rules. The Figure 4 represents the scripting engine methodology of the Enhanced ArabChat after classifying the utterance.

When the utterance is classified as a question but the engine cannot match it with any question-based rules, it will keep the generated utterance's type as it is. Then, it will switch to explore the non-question-based rules (as Figure 3 displays). This switching process (in case no matching occurs in question-based rules) has been adopted to give the utterance another chance to match non question-based rules. If there is no matching in the non-question-based rules as well, the engine checks the previous processed context's patterns if the previous processed utterance's type is the same type as the current processed utterance's type. If so, the scripting engine starts matching the previous context's patterns with the processed utterance. If matching occurs, the scripting engine checks if the utterance tries to target many rules (the utterance has many requested information that require firing many rules, see the "Hybrid Rule" as depicted in Figure 4.

Usually, the user's utterance has one topic (one topic means the utterance targets one rule) to deal with. Other utterances might have many topics to deal with which requires firing many rules for the same utterance. The Enhanced ArabChat scripting engine was designed and developed to deal with this issue.

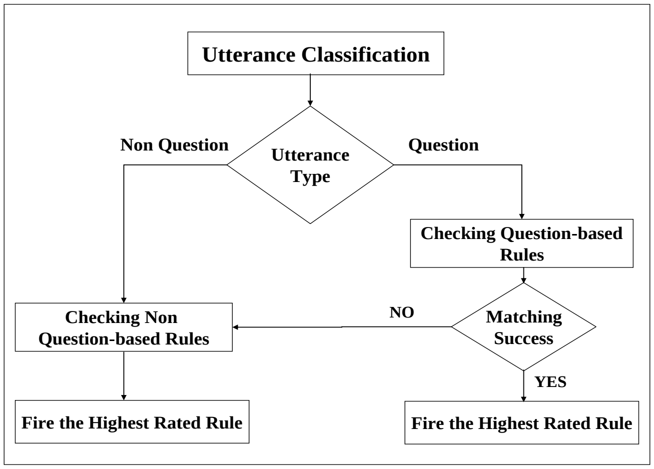

Utterance Classification
Non Question
Utterance Type
Question
Checking Question-based Rules
NO
Checking Non Question-based Rules
Matching Success
YES
Fire the Highest Rated Rule
Fire the Highest Rated Rule

Fig. 3. The Enhanced ArabChat scripting engine methodology based on utterance type

When the scripting engine detects that the utterance has multiple requested topics that require firing many rules, it will deal with all of them in one rule known as the "Hybrid Rule". For instance, an utterance is requesting information about the documents that are needed to register in a university and the fees of registration. Assuming the Enhanced ArabChat has two different rules to deal with these different topics (registration's documents and registration fees). If so (the utterance has the two topics), the scripting engine will deal with them by start creating the "Hybrid Rule" in temporal memory as described later.

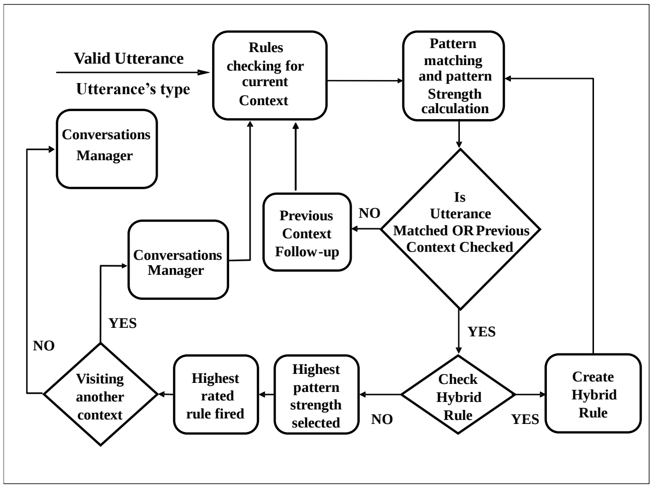

Valid Utterance
Utterance's type
Rules checking for current Context
Pattern matching and pattern Strength calculation
Conversations Manager
Is Utterance Matched OR Previous Context Checked
NO
Previous Context Follow-up
Conversations Manager
YES
NO
YES
Visiting another context
Highest rated rule fired
Highest pattern strength selected
NO
Check Hybrid Rule
YES
Create Hybrid Rule

Fig. 4. The Enhanced ArabChat scripting engine methodology

The Hybrid Rule

Before developing the "Hybrid Rule", when two or more patterns belonging to different rules match against a user's utterance, only the rule that has the highest pattern strength will fire. Matching different patterns for different rules means that the utterance contains (requests) different topics, and the ArabChat replies with just one topic which could be considered a weak point in any CA.

Therefore, the “Hybrid Rule” has been proposed and developed in the Enhanced ArabChat in order reply to an utterance requesting a number of topics.

A “Hybrid Rule” is a hybridization process used by Enhanced ArabChat scripting engine to hybridize all rules that their patterns matched the processed utterance in one rule called the hybridized rule or “Hybrid Rule”. After this hybridization, the “Hybrid Rule” will have the ability to let the engine to reply to all targeted topics in the utterance.

When the Enhanced ArabChat scripting engine detects that many patterns belong to different rules match the same utterance, it starts the hybridization process for the matched rules by accumulating their information (as described in [16]) and creates the “Hybrid Rule” in a temporal memory. A Hybrid Rule’s structure is like any rule’s structure in terms of having an activation level and all other components. However, a Hybrid Rule differs from other rules with the number of patterns and responses. A “Hybrid Rule” will only have one hybridized pattern to match with the utterance that contains many targeting topics and one hybridized response to reply to the targeting topics.

Sometimes a user might merge a large number of topics inside the same utterance, which could lead to generating a very long response. Therefore, in order to avoid a very long response, the Enhanced ArabChat enables the scripter to determine the maximum number of rules to be fired for the same utterance, depending on the highest rules’ strengths priority. Some rules might conflict with each other if they are already designed to deal with opposite topics. When a user targets such conflicting rules in the same utterance, the engine will not fire these conflicting rules for the same utterance in terms of naturalness response. Instead, the Enhanced ArabChat will fire the rule that has the highest strength among them.

As discussed earlier in this section, the Hybrid Rule was created in temporary memory and is kept until the next user’s utterance is processed, to ensure that the consecutive utterance does not target the same topics that the Hybrid Rule handled. If so, the Enhanced ArabChat will re-send the Hybrid Rule’s response to the user in order to avoid recreating the Hybrid rule and thus reducing the processing time.

IV. Experiments and Evaluation

As discussed earlier in this paper and in the published related work in [4], there are number of lessons that have been learned by analysing the first version of ArabChat logs. As a result, these lessons led to continue development to generate the Enhanced ArabChat. In this paper, a number of experiments will be conducted in order to test the full Enhanced ArabChat. In addition, the evaluation methodology (RMUT) that was adopted to evaluate the first version of ArabChat [4] does not give a precise result. Therefore, a new comprehensive evaluation methodology which gives more precise results will be conducted in this paper to evaluate the Enhanced ArabChat. The Enhanced ArabChat evaluation methodology is comprised of two main approaches: namely, objective approach and subjective approach. The objective approach will be applied through developed automatic evaluation measures and logs. The subjective approach will be performed with recourse to human judgment using the user’s questionnaire.

The Enhanced ArabChat’s applied domain is the same domain as the first version of ArabChat. However, the improvements that were conducted in the Enhanced ArabChat’s engine led to modify the Enhanced ArabChat applied domain scripts’ structure (contexts and rules) to meet the new modifications. Consequently, the Enhanced ArabChat’s applied domain contexts and rules were structured to consist of both question-based and non question-based contexts and rules.

Table 1 represents the first version of the ArabChat applied domain’s contexts. In the first version of ArabChat, the contexts numbers are from 1 to 5 (see Table 1) and their rules are regarded as non question-based in the Enhanced ArabChat scripts because they deal with non question-based topics. Context numbers 6 to 32, apart from 30 and 31 (the contexts numbers 30 and 31 regarded as non question-based rules), may be targeted by question-based and non-question-based topics. These contexts (from 6 to 32 apart from 30 and 31) and their rules are regarded as question-based in the Enhanced ArabChat in order to deal with question-based utterances as their scripts were already scripted to deal with this type of utterances. Figure 5 represents the first version of ArabChat applied domain diagram. Then, in the Enhanced ArabChat, a new 25 contexts and their rules have been scripted and regarded as non question-based in order to deal with non question-based topics. Subsequently, patterns scripting process is done for these rules (non question-based rules) in order to match non question-based utterances. In total, the Enhanced ArabChat applied domain consists of 57 contexts, 907 rules and 20944 patterns as depicted in Figure 5.

TABLE I. THE FIRST VERSION OF ARABCHAT APPLIED DOMAIN'S CONTEXTS

#	Context Name	Context Name in English
1	البداية	Home Context
2	اسم المستخدم	User Name Capturing
3	فحص الكتابة بغير اللغة العربية	Test input language
4	الخروج	User exit (session terminating)
5	ألفاظ سيئة	Bad words
6	معلومات عامه عن الأردن	General Information about Jordan
7	معلومات عامه عن جامعة العلوم التطبيقية	General Information about ASU
8	المواصلات	Transportation
9	السكن	Accommodation
10	الإقامه /تأشيرة الدخول	Visa/residency
11	القبول والتسجيل	Acceptance and registration
12	عدد الساعات المعتمده للتخصص	Courses credit hours
13	رسوم الساعه المعتمده للتخصص	Courses hour's fee
14	الحد الأدنى لمعدل القبول في التخصص	Minimum average of a course acceptance
15	التحويل	Transfer
16	التأجيل	Postpone
17	دليل الهاتف	Phone book
18	دليل البريد الالكتروني	E-mail book
19	دفع الرسوم	Fees paying
20	خصومات الطلبة	Fees discounts
21	الدورات التدريبية	Training courses
22	التأمين الصحي	Health insurance
23	التجسير	Bridging
24	توفر التخصص والكلية	Courses availability
25	خدمات الجامعه	University services
26	الدراسات العليا	Postgraduate studies
27	المرشد الأكاديمي للطلبه	Academic advisors
28	متابعة الطلبة الخريجين	Graduated students' follow-up
29	أسئلة شخصية قد تسأل للنظام	Questions to ArabChat
30	لمتابعة الحديث	To continue conversations
31	معالجة الجمل الناقصة(تخصصات)	Dealing with utterances with insufficient explanation
32	مدن الأردن	Jordan cities

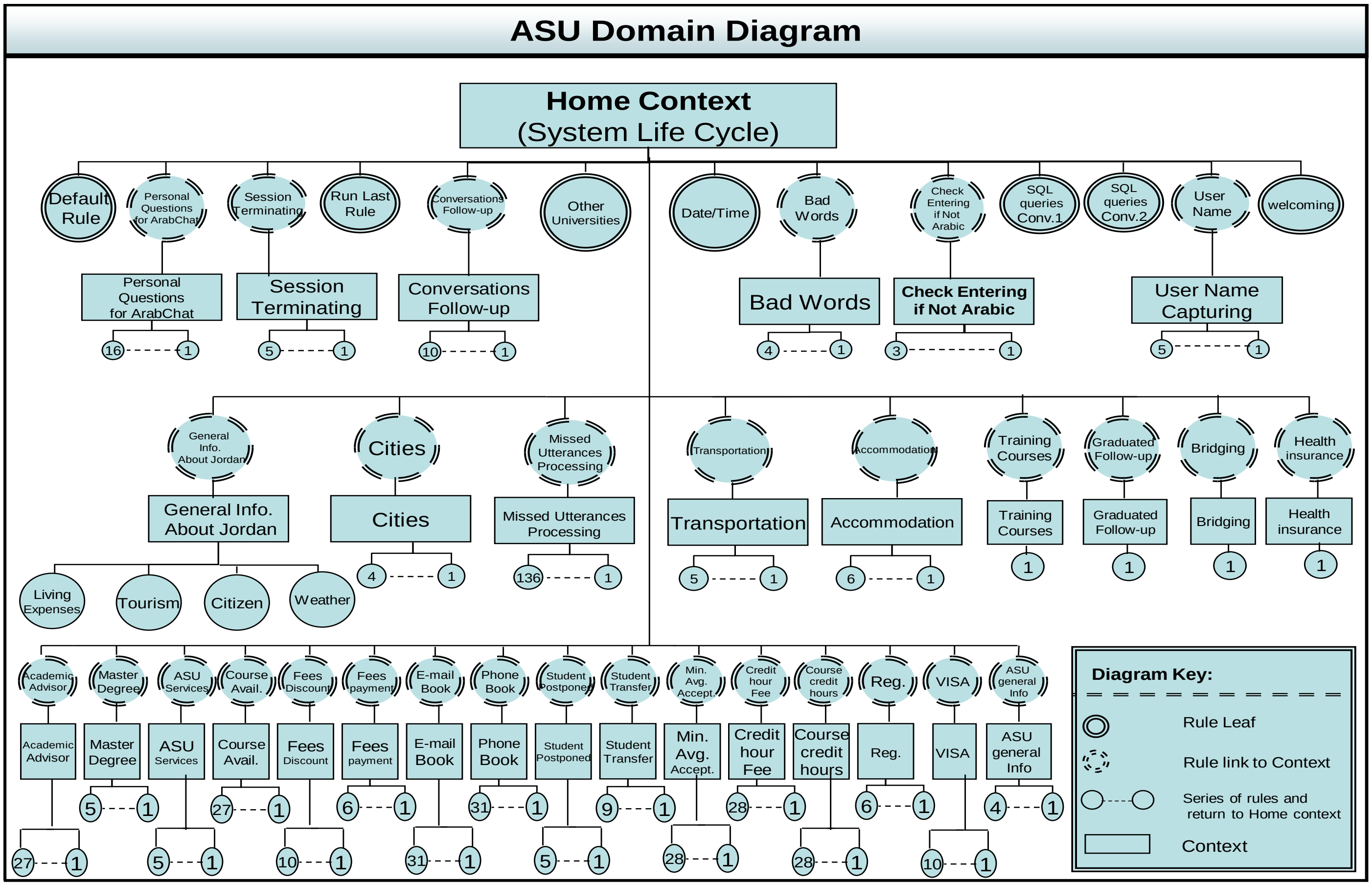


Fig. 5. The first version of ArabChat applied domain diagram

A. Experiment 1

Experiment 1 was conducted to test the full Enhanced ArabChat capabilities from different aspects. Firstly, the experiment tested the Enhanced ArabChat's scripting engine in terms of its ability to recognise patterns' wildcards and matching utterances properly. In addition, test the "Hybrid rule" feature in the scripting engine. Secondly, experiment 1 examined the Enhanced ArabChat classification methodology (the Enhanced ArabChat classifier). Moreover, through analysing the un-matched utterances log, the experiment investigated the applied domain if it is meet the users' needs.

Experiment 1 methodology

The Enhanced ArabChat was deployed on the ASU (Applied Science University) website [17] and accessed by all qualified users such as registered students, non registered students, and employees. The Enhanced ArabChat was available online and in use for 23 days.

Experiment 1 results

The Enhanced ArabChat handled 1766 utterances from 203 users, an average of 8.699 utterances per user. The most accessed contexts are presented in Table 2; these contexts were reported through using the automatic "Domain statistics" report (see [4]).

Table 3 represents the targeting distribution for experiment 1 users based on core domain contexts and general domain contexts. The core domain refers to contexts related to ASU students' issues, while the general domain represents the remaining contexts. Table 3 contents was manually collected through classifying the domain's contexts into "General Domain" and "Core Domain" contexts and then summation the number of targeting for each class's contexts (number of targeting for a context can be reported using the automatic "Domain statistics" report, see [4]) after classifying it whether it is related to core domain or general domain. The results of the Enhanced ArabChat classifier are presented in Table 4. In this table, the number of classified utterances, divided based upon utterance type are presented. All of experiment 1's results will be discussed with the evaluation of the Enhanced ArabChat in the next section.

TABLE II. THE MOST 5 TARGETED CONTEXTS

#	Context Name	Times been targeted (Percentage)
1	Courses Fees (Question-based)	239 (13.533 %)
2	Admission/Registration (Question-based)	193 (10.928 %)
3	Bad words (Non question-based)	146 (8.267 %)
4	Continuing Conversation (Non question-based)	103 (5.832 %)
5	Accommodation (Question-based)	61 (3.454 %)
Total		**742 (42.015 %)**

Utterance classified Type	Number of classified utterances (Percentage)
Question-based Utterance	1005 (56.908%)
Non Question-based Utterance	761 (43.07%)
Total	(100%)

TABLE III. CORE DOMAIN VS. GENERAL DOMAIN; DISTRIBUTION TARGETING BY USERS

Scripted Domain Type	Times been targeted (Percentage)
General domain	302 (25.209 %)
Core domain (Information Point Adviser)	896 (74.791 %)
Total	**1198 (100 %)**

TABLE IV. QUESTION VS. NON-QUESTION CLASSIFICATION RESULTS BY ENHANCED ARABCHAT

The Enhanced ArabChat evaluation based on experiment 1 results

The Enhanced ArabChat will be evaluated using a comprehensive evaluation methodology depending on the results of experiment 1. This evaluation methodology aims to check whether the components of the Enhanced ArabChat are doing their tasks properly (“Glass box approach”). Moreover, the opinions of the Enhanced ArabChat users in experiment 1 will be taken into consideration in this evaluation methodology.

Therefore, this evaluation methodology consists of two main parts, which are objective and subjective evaluations, to cover all previously mentioned aims. The objective evaluation will be used to examine the Enhanced ArabChat as one component and the Enhanced ArabChat individual components using automatic techniques or logs manual checking, as will be discussed in the next section. On the other hand, the subjective evaluation will be done with recourse to users’ judgment. Therefore, a questionnaire has been developed in order to ask the users about their opinion of using the Enhanced ArabChat.

1) The Enhanced ArabChat objective evaluation

The objective evaluation has been done based on the “Glass box approach”. The “Glass box approach” evaluates the main components of the Enhanced ArabChat, namely, the scripting engine, the “Hybrid Rule”, the applied domain coverage, and the “Utterance Classifier”. In addition, a manual check will be conducted to examine the Enhanced ArabChat interaction speed performance and to determine the most Arabic interrogative used in users’ utterances.

a) The Enhanced ArabChat scripting engine Evaluation Aim

The evaluation aim is to test the functionality of the main component of the Enhanced ArabChat: the scripting engine. This evaluation will determine whether or not the scripting engine is doing its tasks properly such as recognising patterns’ wildcard, matching utterances successfully and navigates among the scripted contexts.

Evaluation Methodology

This evaluation will be done by determining the RMUT (Ratio of Matched Utterances to the Total) of the Enhanced ArabChat users. The RMUT (see [4]) is automatically calculated per user by the Enhanced ArabChat once a user session is closed and it can give a general overview of scripting engine’s performance.

Evaluation results and discussion

The evaluation results show that the average RMUT for the 203 users of the Enhanced ArabChat is 67.836%. According to Table 2, the third most targeted context by users was the “Bad words” context which means existing of unserious users and by checking the “Unmatched” log, it has been noticed that number of utterances were unmatched due targeting this context with uncovered keywords or colloquial words. However, by checking the “Unmatched” log, it is possible to notice that the unmatched utterances were due to the use of colloquial words such as using the colloquial phrase “شلونك” instead of “كيف حالك” “How are you”. In addition, there were a number of unmatched utterances due to misspelled keywords such as “مندق” instead of “فندق” “Hotel”. Moreover, there were a number of unmatched utterances due to missing patterns or missing keywords in the scripting patterns. Finally, it has been noted that the fewer amount of unmatched utterances was due to targeting uncovered topics such as requesting more accurate information about tourism in Jordan or asking about courses teachers names, which is outside the main scope of the Enhanced ArabChat domain. Given this, it is possible to conclude that the Enhanced ArabChat scripting engine achieved a reasonable performance (67.836%.of all utterances) in terms of its ability to handle conversations successfully.

b) The “Hybrid Rule” module evaluation

Evaluation aim

The aim is to evaluate the Enhanced ArabChat scripting engine performance in terms of its ability to fire more than one rule for the same utterance at the request of that utterance.

Evaluation methodology

The evaluation methodology is based on manually analysing the Enhanced ArabChat logs in order to determine the utterances that targets more than one topic in them which requires firing more than one rule. Then, analysing the Enhanced ArabChat responses for such utterances and determining manually whether or not the number of replied topics has been performed.

Evaluation results and discussion

Analysing the Enhanced ArabChat logs revealed that there are 121 utterances targeting more than one topic. From those utterances, 85 utterances targeting two topics for the same utterance such as " ما هو سعر الساعة لتخصصي المحاسبة ونظم المعلومات الادارية" "How much is the credited hour for accounting and management information systems courses" and " ما سعر تخصص التمريض وكم عدد ساعاته" "How much is the nursing course and how many credited hours it have". In the first utterance, the user targeted two different rules in the same context, which are "Accounting fees" and "Management Information systems fees", while the second utterance targeted two different rules related to two different contexts, which are "Nursing fees" and "Nursing credit hours". In contrast, 36 utterances targeted 3 topics for the same utterance.

For utterances that targeted two topics, the Enhanced ArabChat replied successfully to 82.354% of them (70 utterances) with the two targeted topics, while the Enhanced ArabChat replied successfully to 7.058% of utterances (6 utterances) with one topic only. The Enhanced ArabChat failed to reply to 10.588% of those (9 utterances) with any topic. Instead, it fired a default rule for the current processed context. The reasons for un-replying to some topics was due to either missing patterns or topics being outside the domain, such as "متى يبدأ التسجيل في الجامعة وما اسم مدير التسجيل" "When the registration in the university will begin and what is the registration manager name". In this utterance the second part of it asking about the registration department manager's name is not covered in the scripted domain. Consequently, the Enhanced ArabChat will reply to only the first part of the utterance asking about the first date of registration, which it already covered in the scripted domain. For utterances that targeted three topics, the Enhanced ArabChat replied successfully to 75% of them (27 utterances) with the three targeted topics. In contrast, the Enhanced ArabChat replied successfully to 13.888% of those (5 utterances) including two topics only. The Enhanced ArabChat failed to reply to 11.111% of those (4 utterances) with any topic. The reasons of un-replying to some topics caused by either missing patterns or topics outside the scripted domain. Given this, the "Hybrid Rule" implementation performance is good enough to fire more than one rule for the same utterance, which means replying to more than one topic at the same time. This might increase the Enhanced ArabChat performance.

c) Domain coverage evaluation

Evaluation aim

The aim is to evaluate aspects related to the scripted domain. Firstly, whether or not the scripted domain coverage was adequate and covered the user needs. This metric will show if the scripted contexts and rules are sufficient to answer all ASU students concerns. Secondly, the evaluation will discuss the most targeted contexts reported in Table 2. The most targeted contexts were measured to identify the topics that users show most interest in and then giving them more priority for future scripting. Thirdly, the evaluation will discuss the users' targeting distribution for the core domain contexts and the general domain contexts, which is reported in Table 3. This will help finding which type of domain contexts the user is most interested in.

Evaluation methodology

The scripted domain coverage was determined manually by checking the content of the "Unmatched" log (see [4]). This log contains all unmatched utterances by the engine. These were mainly because; either the unmatched utterance's topic was not covered in the domain, or the topic is covered, but there is no pattern that matched the utterance. For each context, the Enhanced ArabChat automatically accumulates the number of times it is targeted using the "Domain Statistics" tool (see [4]). Given this, the most targeted contexts are easy to report. The targeting distribution by users between the core domain contexts (contexts numbers from 6 to 57 apart from 29 and 30) and the general domain contexts (contexts numbers from 2 to 5) was manually determined by counting the number of times each context has been targeted, which was already reported using the "Domain Statistics" tool [4].

Evaluation results and discussion

As discussed earlier, it has been revealed that the fewer amounts of unmatched utterances was due to uncovered targeted topics, such as requesting more accurate information about tourism in Jordan or asking about courses teachers names. In addition, some users converse about more detailed topics such as fees, discounts, and student transferral issues require asking the ASU employees. All of these uncovered topics are outside the main scope of the Enhanced ArabChat domain. As a result, Enhanced ArabChat domain coverage is enough to handle ASU's students' utterances. The most accessed contexts are presented in Table 2 and Table 3 presents the accessed distribution by users between the core domain and the general domain. According to Table 2, three of the five most highly accessed contexts are related to the core domain, namely course fees, admission/registration, and accommodation. Although most contexts were students related, there were general contexts in the Enhanced ArabChat which addressed other issues such as "general information about Jordan". Most users concentrated on issues related to their concerns (student issues) rather than general issues. This was expected because the Enhanced ArabChat was employed as an information point advisor for ASU students. The second highly accessed context is "Admission/Registration". This context deals with issues related to admission, registration fees and procedures, indicating that a large number of users were unregistered students. The third most targeted context is "Bad words" that includes rude utterances containing impolite words. This context, which has a high accessed rate, is negatively affecting the results because many of these utterances went unmatched and affected the RMUT ratio. The fourth most accessed context is "Continuing Conversation". This context deals with utterances usually used by users in order to continue the conversation, such as "ok", "great", and "that's fine". Table 3 shows results that confirm that ArabChat's users' focus was on contexts related to their concerns (student issues) rather than those dealing with general issues not directly related to the ASU. As mentioned earlier, this is expected as the Enhanced ArabChat was employed as an information point advisor for ASU students.

As a result, the coverage of general issues (general domain contexts) is reasonable for the nature of work of the Enhanced ArabChat as the most interested domain is the core one.

d) The Utterance Classification module Evaluation

Evaluation Aim

The evaluation aim is to test the performance of the “Utterance Classification” module in the Enhanced ArabChat. As discussed earlier, the “Utterance Classification” module was developed to classify the processed utterances into question-based and non-question-based utterances.

Evaluation Methodology

The Enhanced ArabChat classifies the utterance and stores its type (question-based or non-question-based) in the “Brief log” [4]. A manual checking of the classified utterances in the “Brief log” was conducted in order to find out the real correct classification accuracy.

Evaluation Results

The Enhanced ArabChat “Utterance Classification” module results are presented in TABLE IV. , while the results of the manual check of the classified utterances are presented in Table 5.

TABLE V. ACCURATE (QUESTION VS. NON-QUESTION) CLASSIFICATION RESULTS (MANUAL CHECKING)

Utterance classified Type	**Number of classified utterances (Percentage)**
Question-based Utterance	1312 (74.292%)
Non Question-based Utterance	454 (25.708%)
Total	(100%)

Discussion

Table 4 shows that the Enhanced ArabChat users entered 56.908% of their utterances as question-based utterances. The manual checking of the logged utterances revealed the real percentage of the question-based utterances is 74.292% and not 56.908% as the “Utterance Classification” module generated. The manual checking showed that misclassified question-based utterances were due to the following reasons:

- Conversing using Colloquial Arabic, containing stop words written the Colloquial way such as using “شو” instead of “ماذا” “What” and “هسا” instead of “الان” “Now”. The adopted stop words list by the Enhanced ArabChat’s classifier does not contain such words as they differ among people and thus it is very hard to enumerate them.

- Attaching some interrogative words with other words not applicable in Arabic such as “ ماتكلفة المواصلات باليوم الواحد” “How much is the transportation per day”. The interrogative word “ما” “How much” should not be attached to any word.

- Constructing questions semantically such as “ أعلمنا عن الطقس في عمان” “Tell us about the weather in Amman”. Although, the Enhanced ArabChat’s classifier was trained on indirect questions, but questioning phrases like “أعلمنا” “tell us” “قل لي” “tell me” does not involve the stop words list because they are not stop words. In addition, stop words can accept affixes which might lead to the generation of new words and mislead the classifier.

- Using the interrogative word “أ” “Alef” to build the question such as “أأنت هو المسؤول هنا” “Are you the responsible here?”. The Enhanced ArabChat’s classifier does not learn on these instances, and the stop words list does not contain such an interrogative word (“أ”) because it does not appear alone in Arabic.

- Using stop words in the questions not covered in the “Utterance Classification” stop list such as “فبماذا” “so using which” and “ولماذا” “and why”. The interrogative word “ماذا” “What” accepted the prefix “فب” to generate the stop word “فبماذا” “so using which”. When interrogative words or other stop words accept affixes, enumerating and detecting them becomes impossible.

The misclassified non question-based utterances were due to the following reasons:

- Conversing using Colloquial Arabic, which makes it very hard to detect stop words.

- Using interrogative words in their utterances constructed to seem like questions but in reality, were used for purposes other than questioning, such as “ ما أجمل جامعتكم” (“How beautiful your university”) and “ ما أحد حضر إلى الدورة إلا أنا” “No body came to the course except me”.

The reported number of the entered question-based utterances might be considered large. This is may be due to various factors that need further investigation. These main factors are presented in Table 6 and labelled depending on the potentially responsible party including user, engine, the selected domain, and scripts.

TABLE VI. FACTORS OF THE HIGH PERCENTAGE OF QUESTION-BASED UTTERANCES THAT NEEDS VERIFICATION

#	**Factor description**	**Caused by**
1	It might have occurred due to the nature of the applied domain (Information Point Advisor) which is expected to deal with question-based utterances rather than non question-based utterances.	The selected domain
2	It might have occurred as Enhanced ArabChat scripting engine does not deal properly with non question-based utterances.	The engine
3	It might have been caused by the lack of knowledge and experience of users in terms of the nature of CAs and they are dealing with it as QA system.	The user
4	It might have occurred because the Enhanced ArabChat responses do not encourage people to continue conversations based on non-question utterances. Consequently, the users keep asking questions to ArabChat.	The scripts

In Table 6, there are 4 main factors that might have caused the high percentage of question-based utterances. The nature of the applied domain might have encouraged people to find their requested information rather than participate in normal chatting as it works as information point advisor for issues of students. The results in Table 3 show that 74.791% of utterances accessed the core domain (information point advisor) which supports the first factor and leads to accept it as one of the reasons that caused this large number of question-based utterances.

The second factor (caused by the engine) might need further investigation through another experiment in order to investigate it by conversing with the Enhanced ArabChat using non question-based utterances and monitoring the outcomes. Hence, experiment 2 (discussed later) will deal with this issue. The third factor related to the user will be investigated in the Enhanced ArabChat subjective evaluation (discussed later).

Analysing the Enhanced ArabChat logs has revealed that scripted responses might have failed to encourage users to engage in general conversation with the agent through non-question-based utterances. Unfortunately, as depicted in Figure 5, the applied domain is quite large, which led to increases the difficulties of scripting the large amount of rules. Their responses should have a crafted response to try guide the user indirectly and encouraging him/her to keep conversations going. Accordingly, all of these reasons lead to accept the fourth factor which emphasizes that the responses scripts were one of the reasons for the large amount of question-based utterances.

e) The most used Arabic interrogative words in utterances

Evaluation aim

This evaluation aims to determine the most used Arabic interrogative words in the users' utterances. Determining these words will help in discovering the most used types of questions, such as questions regarding quantities, places, time, or yes/no questions. Then, enhancing the Enhanced ArabChat classifier with the ability of classifying such types of questions.

Evaluation methodology

A manual identification for all question-based utterances from the "Brief Log" has been conducted. Then, a manual counting of the Arabic interrogative words has been done individually.

Evaluation results

The used Arabic interrogative words in users' utterances are presented in Table 7. This table presents the most used Arabic interrogative words and their number of usage.

TABLE VII. MOST USED ARABIC INTERROGATIVE WORDS

#	Interrogative word	Interrogative word Count
1	كم (How much)	289
2	ماذا, ما (What)	243
3	كيف (How)	201
4	هل (Is, are)	127
5	لماذا (Why)	93

Discussion

According to Table 7, users are most interested in questions about issues related to fees, quantities, and manners. This was obvious from the first three interrogative words ("How much", and "what"). In Arabic, the interrogative word "ما" ("what") could be used to ask about fee such as " ما سعر هذا الكرسي؟" ("How much is this chair?") or to ask about quantity such as "ما عدد الطلاب في الجامعة؟" "How many students in the university?".

Other question types users were interested in include "Yes/No" questions, which was conducted based upon from the fourth most used interrogative word. While the last most used interrogative word is related to questions about reasons.

As mentioned earlier, determining the most used interrogative words might help in discovering the most used types of questions and then improving the Enhanced ArabChat classifier's capabilities and thus increasing the quality of Enhanced ArabChat response. In order to do this, further research is needed to classify the question-based utterances into other categories such as questions about places, times, people, and yes/no questions. Then, Enhanced ArabChat tested the compatibility between a question type and a response in terms of whether or not the response met the question's type. If no compatibility is found, other research work would need to be done to develop new techniques to deal with this issue.

f) The Enhanced ArabChat interaction speed evaluation

Evaluation aim

This evaluation aims to check Enhanced ArabChat's interaction speed. Interaction speed refers to the time that Enhanced ArabChat takes to reply to a user. This speed might be used to evaluate the usability of Enhanced ArabChat.

Evaluation Methodology

The Enhanced ArabChat stores the elapsed time that it is taken to process each utterance in the "Brief log" [4]. Then, a manual classification for all utterances into valid and invalid utterances was conducted. In addition, another classification has been conducted to categorise the valid utterances based on the number of targeted rules, one rule or many rules (Hybrid Rule). Finally, a manual calculation for the average of elapsed time for all utterances that related to the previous categories was calculated individually.

Evaluation results

The following results were achieved:

1. The general average elapsed time to process an utterance that access one rule is 1.52 seconds.
2. The average elapsed time that is needed to process an utterance that accesses number of rules (Hybrid Rule) is 3.24 seconds.
3. The average elapsed time is needed to process an invalid utterance is 0.6 seconds.
4. The general average elapsed time to process all utterances is 1.869 seconds.

Discussion

Results showed that the time needed by Enhanced ArabChat to process an utterance is based on the status of the processed utterance (valid or invalid). The invalid utterance needs an average of 0.6 seconds to process. The elapsed time that need to process a valid utterance based on the number of rules that need to be fired in order to handle that valid utterance. For instance, the utterance that requires firing one rule needs less time than utterance requires firing many rules.

The reported average of elapsed time for all processed utterances is 1.869 seconds. This amount of time might be considered a good result, especially as the Enhanced ArabChat handled utterances through the Internet.

Summary of the Enhanced ArabChat components evaluation based on the results of the objective (Glass box) approach

Through evaluating the Enhanced ArabChat using the “Glass box” approach, the following outcomes have been found and they are summarised as follows:

- The evaluation of Enhanced ArabChat scripting engine results show that the average RMUT for the 203 users is 67.836%. This result should be better but unfortunately it has been affected negatively by the number of unserious users as discussed earlier. In addition, by analysing the “Unmatched Utterances” log, it has been revealed that all of the unmatched utterances were due number of reasons as mentioned earlier but not due to an engine failure. Given this, it might be considered that the Enhanced ArabChat scripting engine achieved a reasonable performance in terms of its ability to handle conversations successfully.
- The Enhanced ArabChat scripting engine dealt successfully with utterances targeting many topics which requires firing many rules at the same time. The scripting engine replied successfully to 82.354% of utterances targeting two topics where it replied successfully to 75% of utterances targeting three topics at the same time. Consequently, the “Hybrid Rule” feature in the scripting engine that dealt with these kinds of utterances, performed successfully.
- The Enhanced ArabChat classifier achieved a reasonable performance, which demonstrates that the “Utterance Classification” module had a good methodology of classifying utterances. However, it has been noticed that there are large numbers of question-based utterances. As discussed earlier, the nature of the applied domain (information point advisor) and some of the Enhanced ArabChat response styles were two factors that caused this large number of question-based utterances. However, two other factors, the user and the engine need to be verified will be discussed later in this section.
- As discussed earlier, it has been revealed that the fewer amount of unmatched utterances was due targeting uncovered topics. The rest of unmatched utterances were due to the use of colloquial keywords, misspelled words and missing patterns. The uncovered topics that caused a number of unmatched utterances were outside the main scope of the Enhanced ArabChat domain. As a result, the Enhanced ArabChat domain coverage is enough to enable the Enhanced ArabChat to work as an information point advisor and handle users’ conversations successfully.
- According to tables Table 2 and Table 3, the Enhanced ArabChat users (ASU students) were more interested in conversing about issues related to them (core domain) rather than general topics (general domain). In addition, Table 7 confirms that the Enhanced ArabChat users were more interested in asking about quantities, fees and manners.
- Finally, it has been noticed that the elapsed time that Enhanced ArabChat needs to process an utterance depends on the number of rules that needed to be fired to handle the utterances. The general average of the elapsed time to process all utterances is 1.869 seconds. This is can be considered a good result, especially considering that the Enhanced ArabChat works in an online environment (the Internet).

The next section describes the second part of the evaluation which is the subjective evaluation.

2) The Enhanced ArabChat subjective evaluation

The subjective evaluation aim

As discussed earlier, the subjective evaluation will be conducted by asking the Enhanced ArabChat experiment 1 users to give their opinion about various aspects of using the Enhanced ArabChat. Therefore, an online questionnaire was developed and placed on the same web user interface used to converse with the Enhanced ArabChat. The subjective evaluation (the online questionnaire) aims to enable users to evaluate the Enhanced ArabChat user interface, usability, naturalness, the applied domain coverage, speed, availability of Similar Arabic agent, and user general satisfaction.

The subjective evaluation methodology

The online questionnaire has 14 questions designed to meet the above mentioned evaluation aims. For each aim, a number of questions have been assigned to determine the user opinions concerning them. For each question in the questionnaire, a user has 3 options from which to select his/her degree of approval or disapproval for the asking issue. These options are “موافق” (“Agree”), “محايد” (“Neutral”), “ غير موافق” (“Disagree”). The following are the questionnaire questions (14 questions):

1. “واجهة النظام كانت مناسبة جدا” “The user interface was suitable”.
2. “كان النظام قادر على إجابتك على جميع إستفساراتك” “The agent was able to answer all your utterances”.
3. “أجوبة النظام كانت واضحة ومفهومة.” “The agent responses were clear and understandable”.
4. “لم تواجهك أية مشاكل فنية عند إستخدامك النظام” “You experienced no technical problems whilst using the agent”.
5. “الوقت المستغرق من النظام للرد على استفساراتك كان مناسبا” “The elapsed time taken by the agent was reasonable”.

6. “ تفاعل النظام معك كان واقعي وحقيقي شبيه بتفاعل الانسان من حيث الأجوبة وردود الفعل” “The interaction with the agent was realistic and believable”.
7. “ صعوبة التخاطب مع الجامعة عبر الهاتف والبريد الالكتروني وصعوبة الوصول لمعلوماتك المطلوبة عبر موقع الجامعة الالكتروني حعلك تلجأ لإستخدام هذا النظام” “The difficulty of contacting the university by phone or email, and accessing your needed information on the university website were the reasons to use ArabChat”.
8. “لقد ساهم النظام في توفير جهدك و وقتك .” “The agent saves you time and effort”.
9. “ لايوجد خدمة مثيلة باللغة العربية لأي جامعة ,كلية, أو لشركة و أيضا لايوجد نظام اسئلة وأجوبة باللغة العربية” “There is no Arabic university, college or company offering the same services, even there is no question answering system in Arabic”.
10. “كان النظام يشجعك بالاستمرارعلى الحديث” “The agent encourages you to carry on with the conversation”.
11. “تقييمك الإجمالي للنظام بأنه ممتازا” “Your overall rating for this service is excellent”.
12. “سوف تنصح أصقائك باستخدام هذا النظام” “You will recommend your friends to use the ArabChat system”.
13. “ أنت تفضل استخدام هذا النظام بدلا عن التحدث مع الشخص المسؤول في الجامعة” “You prefer to use ArabChat rather than speak with a human advisor”.
14. “سوف تعيد إستخدام النظام في المستقبل” “You will re-use this service in the future”.

The Enhanced ArabChat online questionnaire system will not accept the submission of any questionnaire without completing all the questions.

The subjective evaluation results

159 of 203 of the Enhanced ArabChat experiment 1 users submitted the online questionnaire. Table 8 presents the Enhanced ArabChat online questionnaire results.

TABLE VIII. THE ENHANCED ARABCHAT ONLINE QUESTIONNAIRE RESULTS

#	“Agree” distribution (Percent)	“Neutral” distribution (Percent)	“Disagree” distribution (Percent)
1	142 (89.3%)	11 (6.9%)	6 (3.8%)
2	140 (88.1%)	19 (11.9%)	0 (0%)
3	122 (76.8%)	29 (18.2%)	8 (5%)
4	153 (96.2%)	6 (3.8%)	0 (0%)
5	96 (60.4%)	44 (27.7%)	19 (11.9%)
6	51 (32.1%)	56 (35.2%)	52 (32.7%)
7	126 (79.2%)	33 (20.8%)	0 (0%)
8	115 (72.3%)	37 (23.3%)	7 (4.4%)
9	152 (95.6%)	7 (4.4%)	0 (0%)
10	48 (30.2%)	46 (28.9%)	65 (40.9%)
11	107 (67.3%)	34 (21.4%)	18 (11.3%)
12	95 (59.7%)	45 (28.3%)	19 (11.9%)
13	103 (64.8%)	29 (18.2%)	27 (17%)
14	109 (68.6%)	34 (21.4%)	16 (10.1%)

Discussion

According to Table 8, the questionnaire questions will now be discussed based upon the evaluation aims as discussed before:

- The Enhanced ArabChat user interface evaluation: the user interface was evaluated using item number 1. 89.3% of users agreed that the Enhanced ArabChat user interface was suitable.

- The Enhanced ArabChat usability evaluation: the Enhanced ArabChat usability was evaluated through 3 items in the questionnaire which are 4, 7, and 8. 96.2% of users agreed that they experienced no technical problems while using the Enhanced ArabChat. 79.2% of users agreed that difficulty contacting the university by phone or email, as well as difficulty accessing their needed information on the university website were the reasons that caused them to use the Enhanced ArabChat. Finally, 72.3% of users agreed that the agent saved them time and effort.

- The Enhanced ArabChat naturalness evaluation: the Enhanced ArabChat’s naturalness has been evaluated through 3 items: 3, 6, and 10. 76.8% of users agreed that the Enhanced ArabChat’s responses were clear and understandable. Only 32.1% of users mentioned that the Enhanced ArabChat’s interaction was realistic and believable. 40.9% of users disagreed with the notion that Enhanced ArabChat encouraged them to carry on with their conversation. This inability to encourage further conversations might be due to the response scripting, which fails to encourage users to continue conversations after firing certain rules. This might provide evidence that the large number of question-based utterances in experiment 1 was due Enhanced ArabChat responses not encouraging users to keep conversations going.

- The applied domain coverage evaluation: the applied domain coverage has been evaluated through item number 2. 88.1% of users agreed that Enhanced ArabChat was able to provide all of their requested information, indicating that the applied domain coverage topics were good enough to cover ASU students’ issues.

- The Enhanced ArabChat interaction speed evaluation: the interaction speed of Enhanced ArabChat has been evaluated through item number 5. 60% of users agreed that the elapsed time taken by Enhanced ArabChat to handle their utterances was reasonable.

- The availability of Similar Arabic agent evaluation: The availability of similar Arabic CAs was evaluated through item number 9. 95.6% of users agreed that there is no Arabic university, college or company offering the same services. This high percentage carries two meanings behind it. First, Enhanced ArabChat might be considered the first Conversational Agent responsible for handling user utterances in the Arabic language. Second, it might reflect the users’ inability to differentiate between CAs and QA(Question Answering) systems. According to Table 5, 74.292% of user utterances were questions. As a result, they might consider it as a QA system due to the lack of experience using similar systems. This fact supports the third factor emphasizing the large number of question-based utterances are due to the users' confusions about whether the Enhanced ArabChat is a QA or a CA.

- The user general satisfaction evaluation: the general satisfaction of the Enhanced ArabChat users was evaluated through item numbers 11, 12, 13, and 14. 67.3% of users agreed that their overall rating for Enhanced ArabChat was excellent, while 59.7% agreed to recommend Enhanced ArabChat to their friends. 64.8% of users prefer to use Enhanced ArabChat rather than speak to a human advisor. Finally, 68.6% of users confirmed they would use the Enhanced ArabChat for future needs.

B. *Experiment 2*

As discussed in experiment 1, 74.292% of the utterances were question-based. This high percentage of question-based utterances might be caused by different factors as discussed in experiment 1. All of the mentioned factors were investigated in experiment 1 apart from the factor that the Enhanced ArabChat scripting engine does not deal well with non question-based utterances. Therefore, this experiment has been conducted on Enhanced ArabChat in order to investigate this factor. Given this, the evaluation of Enhanced ArabChat based on experiment 2's results will be limited to the metrics that related to the engine only.

Experiment 2 methodology

In this experiment, 17 users were asked to have a conversation with Enhanced ArabChat. The 17 users were randomly selected to converse with Enhanced ArabChat. All users were students in ASU from different courses. This experiment focused only on the utterance type (question or non-question) and its effect on continuing conversations. The users were requested to chat with the Enhanced ArabChat by entering non-question-based utterances as much as possible. In other words, they were required to avoid asking questions. The number of utterances that a user should enter was not determined. Therefore, different users entered a different number of utterances.

Experiment 2 results

In this experiment, the Enhanced ArabChat handled 104 utterances from 17 users. The number of classified utterances as questions and non question are presented in Table 9. The results of experiment 2 will be discussed with the evaluation of Enhanced ArabChat based upon these results in the next section.

TABLE IX. QUESTION VS. NON-QUESTION UTTERANCES BY ENHANCED ARABCHAT

Utterance classified Type	**Number of classified utterances (percent)**
Question-based	18 (17.3076 %)
Non Question-based	V. (82.692 %)

The Enhanced ArabChat evaluation based on experiment 2 results

The evaluation of the Enhanced ArabChat for this experiment will deal only with metrics that meet the discussed factor (the engine factor). Therefore, only the objective evaluation will be conducted including the "Glass box approach" evaluation. The "Glass box approach" will only be used to evaluate the Enhanced ArabChat "Utterance Classification" module and the scripting engine.

1) *The objective (Glass box) approach evaluation*

a) *Utterance classification evaluation*

Evaluation aim

This evaluation aims to evaluate the performance of the "Utterance Classification" module.

Evaluation methodology

The Enhanced ArabChat handled 104 utterances from 17 users in experiment 2. These utterances were classified into question-based and non-question-based utterances, as presented in Table 9. A manual classification process for the 104 utterances was conducted in order to evaluate the real correct module performance.

Evaluation results

The real number of question-based and non question-based utterances is presented in Table 10.

TABLE X. THE REAL (QUESTION VS. NON-QUESTION) UTTERANCES (MANUAL CHECKING)

Utterance classified Type	**Number of classified utterances (percent)**
Question-based	14 (13.4615 %)
Non Question-based	90 (86.5384 %)

Discussion

Table 9 presents the classified results of the Enhanced ArabChat. According to the table, 82.6923% of the total utterances are non-question-based. However, Table 10 presented the correct number of non-question-based utterances as 86.5384% of total utterances (manual checking). As a result, the Enhanced ArabChat classifier can be considered acceptably accurate for the two types of utterances (question and non question).

b) *The Enhanced ArabChat scripting engine*

Evaluation aim

The evaluation aim is to determine the RMUT of the Enhanced ArabChat.

Evaluation methodology

To conduct this evaluation, the RMUT equation has been used.

Experiment Results

The results show that the average of RMUT for the 17 users is 72.12%.

Discussion

The results reported in the previous section show that 72.12% of the Enhanced ArabChat users' utterances were matched. This technique cannot reveal if the matching led to a successful conversation or a failed conversation. However, the RMUT, as discussed earlier, gives a general overview of the Enhanced ArabChat scripting engine's performance.

VI. Conclusion

This paper described the Enhanced ArabChat which it is a complement of the first version of ArabChat [4]. Therefore, all the developed features in the first version are also included in the Enhanced ArabChat. In addition, some new features that have been revealed from evaluating the first version of ArabChat are added to improve the agent performance. These new features were "Utterance Classification" and "Hybrid Rule" as described in this paper. Integrating these new features ("Utterance Classification" and "Hybrid Rule") has changed the engine working methodology of the Enhanced ArabChat as discussed in this paper. These changes might reflect positively the performance of the Enhanced ArabChat.

A comprehensive evaluation methodology consisting of objective and subjective approaches has been used to evaluate the Enhanced ArabChat. The objective approach has been conducted through automatic evaluation techniques and manual analysing. The "Glass box" approach evaluated the Enhanced ArabChat components individually. The Enhanced ArabChat obtained a 67.836% of RMUT. This result can give a general overview of Enhanced ArabChat performance, but it does not give a full indicator of its performance. Hence, a new comprehensive evaluation technique for CAs should be modelled and developed. The subjective evaluation showed that 67.3% of users who submitted the questionnaire agreed that their overall rating for Enhanced ArabChat was excellent, and 64.8% of them prefer to use it rather than speak with a human advisor.

It has been observed in experiment 1 that users entered more question-based utterances than non-question-based ones. This might be due to four factors, including the nature of the selected domain, the engine, the user and the scripts. In experiment 1, three of these factors have been discussed and verified as accurate reasons for this problem: the nature of the scripted domain, the user and the scripts (the Enhanced ArabChat responses). In experiment 1, it was noticed that non-serious users negatively affected the calculated user satisfaction by chatting with the Enhanced ArabChat in an indecent manner (not covered by the "Bad words" context). Also, these non-serious users tried entering many questions just to trick the Enhanced ArabChat. Moreover, as discussed in the subjective evaluation, 95.6% of Enhanced ArabChat users agreed that this was the first time they used such a service (ArabChat information point advisor). Thus, the numerous question-based utterances might be to the fact that users cannot differentiate between CAs and QA systems.

Therefore, experiment 2 was conducted for the Enhanced ArabChat to check the fourth factor (the engine factor) that might have caused the large number of question-based utterances. Experiment 2 confirmed that Enhanced ArabChat successfully dealt with non-question-based utterances, as the reported user satisfaction rate was 70.488%. Consequently, it was concluded that Enhanced ArabChat scripting engine can deal with non-question-based utterances. This evidence led to the rejection of the fourth potential factor.

Generally, chatting with a CA does not mean that a user will keep entering either questions or non-questions only. The natural conversations between a user and a CA should consist of both (questions and non-questions). Nevertheless, the amount of question and non-question utterances might be based on the following factors:

1) The topical nature of a CA's applied domain; for instance, an entertainment domain might differ from an information point advisor.

2) The users, if they are familiar with the nature of a CA. It can be concluded from experiment 1 of Enhanced ArabChat that many users consider it a question answering system. As a result, a lot of questions were entered. Also, 92.3% of experiment 1's users confirmed that they had never used any similar service before, which points to a lack of experience in handling these services.

3) The way a CA forms its response might also encourage a user to ask questions or continue chatting with non-question utterances.

According to the two conducted experiments (experiment 1 and experiment 2) and the evaluation of the Enhanced ArabChat based on these experiments' results, it can be concluded that the Enhanced ArabChat successfully handled conversations for ASU students.

Acknowledgment

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

References

[1] Turing, A., Computing machinery and intelligence. Mind, 1950: p. pp 433-60.

[2] Turing, A., Computing machinery and intelligence. MIT Press, 1995: p. 11-35.

[3] O'Shea, K., Z. Bandar, and K. Crockett, A Novel Approach for Constructing Conversational Agents using Sentence Similarity Measures. 2008.

[4] Hijjawi, M., et al. ArabChat: An Arabic Conversational Agent. in proceeding of the 6th International Conference on Computer Science and Information Technology (CSIT). 2014. Amman, Jordan: IEEE Explore.

[5] Crystal, D., Dictionary of linguistics and phonetics., Blackwell., Editor. 2008.

[6] Habash, N., Introduction to Arabic Natural Language Processing, ed. U.o.T. Graeme Hirst. 2010: Morgan & Claypool.

[7] Sammut, C. and D. Michie, InfochatTM Scripter's Manual, Convagent Ltd. . 2001: Manchester.

[8] Weizenbaum, J., ELIZA-A computer program for the study of natural language communication between man and machine. Communications of the ACM., 1966. Vol 10.: p. PP 36-45.

[9] Wallace, R. ALICE: Artificial Intelligence Foundation Inc. . 2008 [cited; Available from: http://www.alicebot.org.

[10] Timothy, B. and G. Toni, Health dialog systems for patients and consumers. J. of Biomedical Informatics, 2006. 39(5): p. 556-571.

[11] Maragoudakisa, M., et al., Natural Language in Dialogue Systems, a case study on a medical application. , in Proceedings of Panhellenic

Conference with International Participation in Human–Computer Interaction. 2001: Greece. . p. 197–201.

[12] Shaalan, K., Rule-based Approach in Arabic Natural Language Processing. 2010.

[13] Sammut, C., Managing Context in a Conversational Agent. Electronic Transactions on Artificial Intelligence, 2001.

[14] Ong Sing, G. and F. Chung Che, The design of interactive conversation agents. WSEAS Trans. Info. Sci. and App., 2008. 5(6): p. 901-912.

[15] Hijjawi, M., Z. Bandar, and K. Crockett. User's utterance classification using machine learning for Arabic Conversational Agents. in proceeding of the 5th International Conference on Computer Science and Information Technology (CSIT). 2013: IEEE Explore.

[16] Hijjawi, M., Z. Bandar, and K. Crockett, A Novel Hybrid Rule Mechanism for the Arabic Conversational Agent ArabChat. Global Journal on Technology, 2015(Issue 8): p. 185-194.

[17] ASU. Applied Science University. 2011 [cited; Available from: www.asu.edu.jo.

Understanding a Co-Evolution Model of Business and IT for Dynamic Business Process Requirements

Muhammad Asif Khan
Department of Information Systems
College of Computer Science and Engineering, Taibah University
Madinah al Munawwara, Saudi Arabia

***Abstract*—Organizations adapt existing business processes in order to become competitive but a change in a process affects other processes as well. In order to support the required change suitable technologies must be provided so that business could run smoothly and efficiently. Since in a dynamic business environment requirements are changed frequently it is difficult to update underlying technologies to support changes in business processes. This creates a gap between business and information technology (IT) that directly affects whole business. In this study requirements for a dynamic business and a co-evolution model are presented that may bring both the entities closer to bridging the gap in a dynamic business organization. The co-evolution model has been used in a financial institution and feasibility and viability of the model has been observed.**

***Keywords*—alignment; business-IT gap; business process; co-evolution**

I. Introduction

Today a large number of companies are increasingly automating business processes with latest information technology (IT) in order to meet customers' requirements. Most of the organizations invest hefty budget for acquiring IT but fail to achieve the desired results in terms of return on investment. Researchers and practitioners found that absence of alignment between business and IT causes the failure to meet the desired goals of organizations. Earlier researchers focused on aligning business strategies and IT strategies [1]. Now alignment is studied at strategic, operational and IT project levels [2][3][4]. The main reason of the failure is that business requirements and IT evolve separately consequently the rate of evolution is different in both the entities. If the business requirements and underlying technologies co-evolve then they will have a significant impact on business. Technologies are continuously emerging that help to develop and improve business processes whereas the requirements coerce IT to be evolved in order to meet the business needs. An analysis of co-evolution requirements has been discussed in a study conducted by Khan & Zedan [5] in which dependency, independency and interdependency between both the entities is discussed. In a dynamic business environment it is necessary that underlying technologies are evolved to support the business but due to continuous changes adaptability in technology is difficult and that affects the efficiency of the business. Therefore, an efficient dynamic business requires supporting IT to be evolved so when business requirements are changed the supporting software has to be evolved i.e. a co-evolution should occur. Contrary to this if changing business requirements are not supported by the IT a gap is created due to absence of co-evolution in turn decrease in efficiency of the business happens. Morrison et al. [6] used the co-evolution term for describing evolution in both business and IT at different rate. Business requirements or supporting technologies are evolved due to internal or external changes. If the business processes are independent of each other the evolution is easy and each process can adjust it, but for co-evolution business processes are interconnected and therefore, a change in a process affects other processes as well. To carry out business activities smoothly and effectively all the processes need to be co-evolved. Researchers and practitioners have demonstrated various models to bring both business and IT into alignment and the most well-known model is Strategic Alignment Model (SAM) in which alignment is observed at three levels within organizations [7]. In another study [8] presented three-layers model in a firm and discussed the business-IT alignment as a co-evolutionary process and asserted co-evolution occurs due to external changes and within organizational components. Earlier researchers and practitioners focused on aligning business strategies and IT strategies with the understanding that one could align the other [1]. In a study Chebrolu [9] assessed the relationship between strategic alignment and IT effectiveness in order to ascertain the dominance between the two constructs. Recently Suwatana, Winai, & Do [10] presented both strategic and operational levels within organizations in order to align both business and IT. In a study Richard & Lucy [11] described the importance of alignment between business strategy and IT strategy and stressed the need of technological readiness in terms of human resource before financial investment. Many studies have shown that IT investment has been effective in organizations' performance in different business areas [12]. A four-layer framework aimed at reducing gap between business and IT by considering requirements of business and information management rather information systems [13]. For business-IT alignment a conceptual model-driven approach has been presented by [25] that aim at restriction of freedom in process modeling. A model based on social, cultural and structure aspects was presented in order to achieve alignment between business processes and IT [30]. In order to gain maximum benefits from technology an organization must be responsive and it should accustomed to changing business challenges requirements and opportunities [14]. Khan [15] presented a co-evolutionary framework using K-mediator that aims at reducing gap between business and IT. In the present time the business-IT alignment seems slow rather stationary in view of increasing and emerging technologies around the globe. In a recent study an approach has been proposed that helps evaluation of alignment level between business process and supporting software [29].

In this paper co-evolution requirements for a dynamic business have been presented. Also a co-evolution model for business and IT is presented and later the model is validated in a dynamic financial business organization where a change causes a co-evolution. The efficacy of co-evolution requirements is discussed and finally recommendations for effective co-evolution in organizations are presented.

II. Co-Evolution Model

Companies are adopting IT function in business processes in order to facilitate customers. Now customers can access a firm's database and give and track their orders online from manufacturing to delivery. Business processes not only aligned with the IT functions but co-evolve as the requirements are changed. Ehlrich et al. [21] introduced the term 'co-evolution' and researchers and practitioners have been using this term in research to explain that evolution of one entity is partially dependant on another entity [22][23][24]. In biology co-evolution occurs in an ecosystem in which each living creature has other creatures of the same environment and other creatures are the parts of its environment [24]. In the present study ecosystem comprises of financial institution and related industries that may influence the institution under study. When co-evolution occurs it impacts both individual elements and the environment. Co-evolution concept has been used in variety of areas from biology to commerce and business to technology [8].

In a dynamic business environment business processes keep changing in a fast paced environment and therefore, it is important to know the requirements for a dynamic business environment.In a dynamic business environment new policies, rules, regulations and customer demands change with a passage of time, therefore on one hand new technologies are integrated to support the changes while on the other hand new models are evolved to facilitate in evolution to both the entities. In a dynamic distributed business environment one may expect unusual events and underlying technologies must be able to handle such situations. In dynamic environment business processes should have loose coupling with each other so that propagation of any change is minimum from one process to another [5]. In the event of any partnership the business processes should be changed in order to accommodate new rules, policies that occur due to new collaboration [16].

In view of a dynamic business environment a co-evolution model is presented where business processes (b1, b2, b3) are supported by underlying IT services (t1, t2, t3). Since the business environment is dynamic a change occurs in business processes and a new process b4 is added to the existing processes. In order to meet the evolution requirement in business processes IT services need to be evolved as well and therefore, a new IT services t4 is created. Figure 1 depicts the scenario where different business processes are integrated and process b2 is evolved to b2' and supporting IT service t2 also co-evolved to t2'. Also a new business process b4 and its supporting IT service t4 is added.

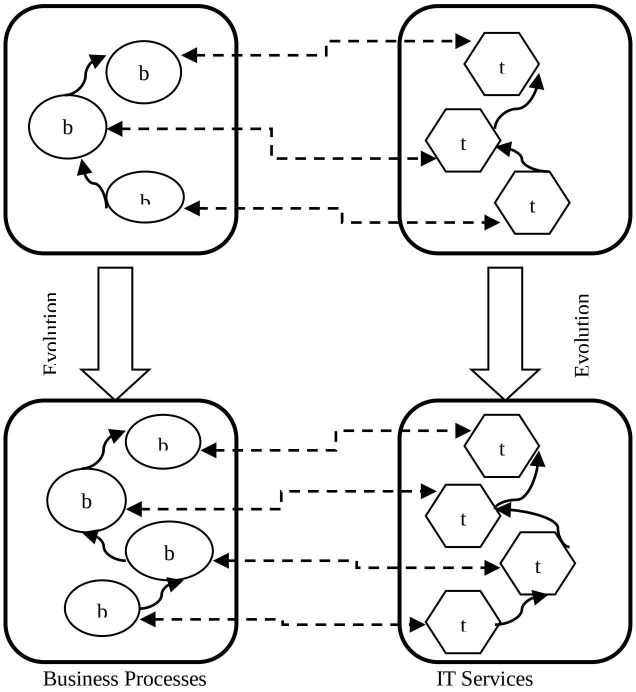


Fig. 1. Business-IT Co-evolution Model

In Figure I it is obvious that adaptability in business or IT function will cause a change in another entity that in turn results a co-evolution. Co-evolution increases efficiency and performance in business and meets customer requirements. It is to be noted that change in a business process or IT can be due to customers' demands or surrounding environment in the marketplace. It is also important for a co-evolution that managers in both business and IT are knowledgeable in each other domains in order to keep co-evolution. When a change occurs in either business requirements or IT function manager of either entity must communicate the supporting function required to co-evolve through the different layers of management and ensure the communication is reached to each level effectively.

A. *Efficacy of the Model*

Dynamic business organizations strive to optimize available resources in order to provide efficient services and to meet customers' needs. Since the alignment between business and IT in organizations is persistent and has potential implications researchers give it priority in organizations [17]. Any change in business process or technology may affect the entire business, therefore, it is necessary to consider any change with great care. In a dynamic business environment a co-evolution between business and IT enables organizations to achieve optimization. In order to see the efficacy of the co-evolution model a service industry was selected because service requires skilled personnel and an effective and efficient technology for running smooth business [10]. A financial institution (called FI for privacy agreement) was selected due to its dynamic nature of business. Financial institutions are more frequently change requirements and adapt processes and underlying technologies. FI provides effective and efficient services to its customers both online and offline. Recently in board meeting of the institution it was discussed to establish a new department that could process customers loan requests in a short period of time. The board decided to approve the establishment of department with required resources. It was noticed the new department required the customers' data on a regular basis so that loan applications could be completed in a minimum time. This process required linkage of other departments with the new department via telecommunication. Customer applications for loan are submitted in a local branch whereas the data is transmitted to other departments immediately so that they could process data and send to the newly established department. In order to handle customers' data in a timely manner different departments also required IT to be updated. Hence the creation of new department caused to update and purchase new telecommunication devices and software. This clearly shows the co-evolution in both business and IT which is in agreement with the co-evolution model. In result of the co-evolution all departments are updated with processes and required technologies. This ensures that the organization is working with the updated information and latest tools available. In the co-evolution model different business processes and technologies are loosely coupled therefore co-evolution is easy and the organizations benefit extensively in result of any change in the business.

III. Validation of the Model

In order to validate the model it was decided to collect data by a survey instrument and a case study in a financial institution. A case study approach helps to understand variables in real life situation. A case study provides rich, extensive and complete details [26]. In this study an intrinsic case study is used that helps to understand specific phenomena in an environment. The questions in the survey instrument were divided in three sections. The first part consisted of questions related to measuring performance of the institution; the second part comprised of questions measuring the business and IT strategies and the third part of the questionnaire focused on the organization architecture. All the questions were prepared using on Likert's scale of 1 to 5 where 1 stands for 'Strongly Disagree' (SDA) and 5 stands for 'Strongly Agree' (SA). The remaining values in between are as 2= 'Disagree' (DA), 3='Neutral' (NU) and 4='Agree' (AG). The survey instrument was distributed to different employees working in different departments at different levels. Table I shows the parameters used in the questionnaire in order to determine the performance in the organization.

TABLE I. Performance Parameters

Parameter	Description
QP	Improvement in quality of the product or service
CS	Increase in level of customer satisfaction
OM	Improvement in organization image
RI	Impact in return on investment
AG	Increase in annual growth

Since a change in requirement has caused creation of a new department in the firm it was necessary to observe the co-evolution in other departments so that the co-evolution model could be validated. There were 52 survey instruments distributed among employees at operational and management levels in the firm. Table II and Table III show the expressions used in the questionnaire in order to get data in both business and IT areas.

TABLE II. Expressions and Parameters for Business Strategy

Parameter	Description
IT	Impact of IT in business process
PR	Business process reengineering
IN	Involvement of IT people in business
GR	Growth in business products
RK	Willingness to take risks

TABLE III. Expressions and Parameters for IT Strategy

Parameter	Description
TB	Business knowledge in IT plan
TA	Acquirement of IT
TU	Updating IT for business requirement
TC	Cost involve in IT acquisition
TE	High level of expertise in IT

Out of 52 distributed surveys 36 questionnaires were received from the financial institution (F1). The survey questionnaires were scrutinized and only 29 questionnaires were found complete in all aspects. Cooper & Schindler [18] stated that a measure is reliable to the degree that it supplies consistent results. Reliability of a questionnaire is important in order to obtain results and hence it was necessary to ensure all items in the questionnaire were reliable. According to [27] an item is called a reliable item when same results are produced by the same items. Internal consistency in reliability is significant which shows consistency in a measuring scale [19]. and to determine internal consistency in a measure reliability of the questions in the survey instrument a reliability test was conducted to determine Cronbach's alpha value [19][20]. After a reliability test it was found 26 questionnaires had Cronbach's alpha value greater than 0.7 that gave confidence to select them for the study. The data collected from the employees were coded into a spreadsheet according to the conventions stated by [28] and assigned numerical values to answers given by respondents in the questionnaires. The values become attributes of the variables. Table IV shows the scores obtained against each parameter used in the questionnaire.

A co-evolution may not take place until all the developing business processes are completely sustained by underpinning evolving technologies. In the financial institution under study it is illustrated from Table IV that absence of latest acquisition of IT (TA) and expertise (TE) may create a gap between evolving business processes and IT as most of the respondents were neutral (value close to 3) for giving answers to such questions. It is obvious from the data that the firm performed well as it received high return on investment (RI) in IT as most of the respondents responded with average value of little higher than 4 (i.e. Agree). This is possible because all the employees of IT department were involved in changing the business requirements as evident from the data parameter IN that has average value higher than 4. These data were also supported by TB and TU where employees agreed that in planning for IT knowledge of business is essential that helps updating IT for any change in business requirements.

TABLE IV. AVERAGE SCORES ON LIKERT SCALE

Parameter	Expression	Average Score
QP	Improvement in quality of the product	3.61
CS	Increase in level of customer satisfaction	3.52
OM	Improvement in organization image	3.23
RI	Impact in return on investment	3.90
AG	Increase in annual growth	3.71
IT	Impact of IT in business process	4.23
PR	Business process reengineering	3.47
IN	Involvement of IT people in business	4.23
GR	Growth in business products	3.80
RK	Willingness to take risks	3.42
TB	Business knowledge in IT plan	4.09
TA	Acquirement of IT	3.04
TU	Updating IT for business requirement	4.14
TC	Cost involve in IT acquisition	3.33
TE	High level of expertise in IT	2.95

This clearly shows that when a change was required in the institution knowledgeable people in business from IT were required to update technology in order to support the change. The co-evolution model proves its viability and good results show its feasibility in financial institution.

IV. CONCLUSION

Co-evolution process is a key of success in a business where both business and IT are evolved to expedite businesses in order to meet requirements. As the Organizations always look for different strategies to optimize their businesses and seldom consider to co-evolve business processes with underlying technologies. In this study it is discussed that the requirements of a dynamic business environment and presented a co-evolution model of business and IT. It is observed in a financial institution that in result of a change in business different business processes and supporting technologies are co-evolved. Co-evolution increases the efficiency and effectiveness of a business that facilitates to meet customer demands.

The model works well when both business and IT co-evolve without any delay, however, as organizations tend to change either business process or IT at slow pace the desired results may not be achieved. Therefore, in future, it is anticipated to work on rate of change in business or IT so that organizations could decide to control the speed of co-evolution. The predetermination of co-evolution speed will help organization to manage resources efficiently and effectively and decision makers may find it convenient to manage businesses. In future it is expected to have an evaluation of the co-evolution model in other industries with different methodology.

REFERENCES

[1] D. Beimborn, J. Franke, H. Wagner, & T. Weitzel, “The Influence of alignment on the post-implementation success of a core banking information system: an enbedded case study”, in Proc 40th Annual Hawaii International Conference on System Science, Hawaii, USA, 2007.

[2] D. Beimborn, F.Schlosser, & T. Weitzel, “Proposing a theoretical model for IT governance and IT business alignment”, in Proc of the 42nd Hawaii International Conference on System Sciences (HICSS 2009), Hawaii, USA, 2009.

[3] L. Silva, E.Figueroa, & J. Gonzalez-Reinhart, “Interpreting IS alignment: a multiple case study in professional organizations, information and organization”, vol. 17, pp. 232–265, 2007.

[4] T.A. Jenkin, & E. Chan, “IS project alignment — a process perspective”, Journal of Information Technology, vol. 25, pp. 35–55, 2010.

[5] M.A. Khan, & H. Zedan, “Requirements Analysis for Co-Evolution of Business and Information Technology”, Advanced Materials Research, vol. 457-458, pp. 968-973, 2011.

[6] R. Morrison, D. Balasubramaniam, G. Kirby, K. Mickan, B. Wardboys, R. Greenwood, I. Robertson, & B. Snowdown,B.,“A framework for supporting dynamic systems co-evolution”, Autom Softw Eng, vol.14, pp. 261-292, 2007.

[7] D.B. Henderson, & N. Venkatraman, “Strategic alignment: leveraging information technology for transforming organizations”, IBM Systems Journal, vol. 32, pp. 4-16, 1993.

[8] H. Benbya, & B. McKelvey, “Using coevolutionary and complexity theories to improve IS aligmnment: a multi-level approach”, Journal of Information Technology, vol. 21, pp. 284-298, 2006.

[9] S. Chebrolu, “How Does Alignment of Business and IT Strategies Impact Aspects of IT Effectiveness”, International Journal of Applied Management and Technology, vol. 12, pp. 1-15, 2013.

[10] C. Suwatana, W. Winai, & K. Do, “Business-IT Alignment: A Practical Approach”, Journal of High Technology Management Research, vol. 25, pp. 132-147, 2014.

[11] L. Richard, & L. Lucy, “Technology strategy and sustainability of business empirical experiences from Chinese cases”, Journal of Technology Management in China, vol. 8, pp. 62-82, 2013.

[12] J. Halamka, “Making smart investments in health information technology: core principles”, Health Affairs, vol. 28, pp. 385–389, 2009.

[13] Strnadl., F.,“Aligning business and IT: the process-driven architecture model”, Information Systems Management, vol. 23, pp. 67-77, 2006.

[14] F. A. Cummins, “Building the agile enterprise with SOA, BPM and MBM”, Burlington, VT: Morgan Kaufmann,2009.

[15] M.A. Khan, M.A., “An integrated framework to bridging the gap between business and information technology – a co-evolutionary approach”, Canadian Journal of Pure and Applied Sciences, vol. 7, pp. 2611-2618, 2013.

[16] J. Meng, S. Sue, H. Lam, & A. Helal, “Achieving dynamic inter-organizational workflow management by integrating business processes, events and rules”, in Proc. Hawaii International Conference on System Sciences, 2002, USA.

[17] L. Kappelman, E. McLean, J. Luftman, & V. Johnson, “Key issues of it organizations and their leadership: the 2013 sim it trends study”, MIS Quarterly Executive, vol. 12, pp. 227-240,2013.

[18] C. R. Cooper, and P.S. Schindler, Business research methods, 10th ed., 2008, Boston: McGraw-Hill.

[19] L.J. Cronbach, “Coefficient alpha and the internal structure of tests”, Psychometrika , vol. 22, pp. 297-334, 1951.

[20] J. Cortina, “What is coefficient alpha? An examination of theory and methods”, Journal of Applied Psychology, vol. 78, pp. 98-104, 1993.

[21] D.R. Ehrlich, & R.H. Raven, R.H., “Butterflies and plants: a study in co-evolution”, Evolution, Vol. 18, pp. 568-608, 1964.

[22] M.P. Koza, & A. Lewin, A., “The co-evolution of strategic alliances”, Organization Science, vol. 9, pp. 255-264, 1998.

[23] B. Mckelvey, “Visioning Leadership vs Distributed Intelligence: Strategy, Microcoevolution, Complexity”, in proceedings of EIASM Workshop, 1999. Brussels.

[24] S. Kauffman, The Origins of Order: Self-organization and selection in Evolution, 1993, Oxford University Press.

[25] J. Martin, W. Jens, W., & E. Werner, “A model-driven framework for business it alignment”, Int. J. Internet and Enterprise Management, vol. 6, pp. 233 – 247, 2010.

[26] F. Bent, Case Study, In Norman K. Denzin and Yvonna S. Lincoln, eds, The Sage Handbook of Qualitative Research, 4th ed, Chapter 17, pp. 301-316, 2011.

[27] E. Carmines, R. Zeller, R., Reliability and Validity Assessment, 1979., SAGE Publications.

[28] W. Mangione,, Mail Surveys: Improving the Quality, Thousand Oaks, 1995. Sage Publications.

[29] L. Aversano, C. Grasso, and M. Tortorelle, “Managing the alignment between business processes and software systems”, Information and Software Technology, vol 72, pp. 171-188, 2016

[30] E. Seman, and J. Salim, “A Model for Business-IT Alignment in Malaysian Public Universities”, Procedia Technology, vol. 11, pp. 1135 – 1141, 2013

Performance Evaluation of Loss Functions for Margin Based Robust Speech Recognition

Syed Abbas Ali
Dept. of Computer & Information Systems Engineering,
N.E.D University of Engineering & Technology.
Karachi, Pakistan

Raheela Asif
Computer Science & Information Technology
N.E.D University of Engineering & Technology.
Karachi, Pakistan

Maria Andleeb
Computer Science & Information Technology
N.E.D University of Engineering & Technology.
Karachi, Pakistan

Danish-ur-Rehman
Dept. of Computer & Information Systems Engineering,
N.E.D University of Engineering & Technology.
Karachi, Pakistan

Abstract*—Margin-based model estimation methods are applied for speech recognition to enhance the generalization capability of acoustic model by increasing the margin. An important aspects of margin based acoustic model for parameter estimation is that, the acoustic models are derived from soft margin concept and hinge loss function used in SVM as loss function to attained enhanced speech recognition performance. In this study, performance evaluation of loss functions (Logistic, Savage, Sigmoid) have been computed in the presence of white noise, pink noise, and brown noise with and without SVM classifiers to analyze the impact of noise on loss functions in comparison with hinge loss function used in SVM for parameter estimation in margin based acoustic model. Experimental results show that hinge loss function in the presence of pink noise and white noise have significant effects on isolated digits (0-9) in both pre-conditioned and recorded data samples in comparison with brown noise. Whereas hinge loss functions show serious anomalies with savage loss and sigmoid loss in term of performance and sigmoid loss function provides exceptionally good results in term of percentage error for all prescribed conditions.

Keywords*—*Loss Functions; Statistical Learning; Automatic Speech Recognition (ASR); SVM Classifiers; Soft Margin Estimation (SME)

I. Introduction

The prime goal of pattern recognition is to find the parameters of recognizers or classifiers that can decrease the error rate by using the existing training data samples. To build an effective pattern recognizer or classifier, two different categories of learning algorithms in machine learning are generative model learning and discriminative model learning. MLE is considered as generative model or non-discriminative learning approach, which is focus on data distribution modeling instead of directly classifying class boundaries. In contrast, discriminative learning approach discriminately learns the parameters of joint probability model to minimize the recognition/classification error [1]. The main idea behind Discriminative training (DT) is to introduce a discriminative criterion to the training method of Hidden Markov Models (HMMs). Several discriminative training methods have been proposed for ASR, such as maximum mutual information estimation (MMIE) [2,3,4], minimum classification error (MCE) [5,6,7]; and minimum word/phone error (MWE/MPE) [8,9]. For Hidden Markov (HMM) based speech recognition, conventional discriminative training criterions directly minimize the empirical risk on the training data sample and do not focus on the model generalization. In other words, the aim of discriminative training criterions is to minimize classification error on training sample as model estimation but do not show any significance performance to improve the generalization capability of the acoustic model for new unseen test data samples [10]. The generalization capability is an ability to translate gains in the training data set to test data set. In the past studies, the discriminative training achieved this generalization ability by optimizing the smoothed empirical error rate on training data samples [11]. Recently, many researches have been reported to incorporate margins (distance between the decision boundary and well classified data samples) into discriminative training method [12,13,14,15,16,17] to further enhance the generalization capability. The generalization problem of learning classifiers have been studied in the field of machine learning [18,19], whereas, machine learning using concept of statistical learning theory since last three decades to provide the framework for studying inference problem that is of making prediction, gaining knowledge, constructing models and making decisions for a set of data samples [20]. From the statistical learning [18] point of view, a test risk bound is defined by the summation of an empirical risk (i.e., training set risk) and a generalization function. Generalization function is often used to measure the possible mismatch between training and testing environments. ASR researchers at York University proposed the concept of large margin estimation (LME) for speech recognition based on the principle of large margin. Large margin estimation (LME) [10,12] and its variant large relative margin estimation (LRME) [21] of HMMs have been proposed with the concept of enhancing separation margin. The main crux of the LME and LRME is that only correctly classified data samples take part in update models whereas, it is important to note that misclassified data samples are also substantial for classifier learning. To address this issue in LME and LRME, the extension of LRME [86] was proposed by considering all the training data samples, particularly moving misclassified data samples in the direction of correct decision boundary. Another margin based approach, Soft margin estimation (SME) was proposed by J.Li et al [22] from Georgia Tech University based on the idea of soft margin in support vector machines [23] to enhance the generalization capability of the learning classifiers. Soft margin estimation (SME) performs well as compared to Maximum likelihood estimation (MLE) and conventional discriminative criterion, and it is steadily better than Large margin estimation (LME) due to the well-defined separation(misclassification) measure and good optimized objective function for generalization [24]. In contrast, with LME, SME make use of both misclassified and correct classified data samples to update models and the performance of the SME can be improve when the distribution of testing and training data samples become quite comparable[25]. Two considerable issues have been identified in [26] related to hinge loss function in Soft margin (SME) 1) hinge loss function performs well when the noise in training sample is insignificant and 2) any misclassified training sample directly affects the time required for optimization and determines the label of the test sample. To improve this limitation of SME, X.Xiao et al [27] proposed feature domain method based on mean and variance normalization (MVN) [28] which showed that SME perform well with feature domain method and reduces the mismatch between training and testing data samples and suggested the combination of SME framework with other noise compensation methods e.g. model adaptation methods for future research. .Issues related to hinge loss function, Geometric margin MCE criteria in soft margin estimation framework based on sigmoid loss function were presented to find the strength of robustness by increasing the geometric margin of the acoustic model [29]. Loss functions such as ramp loss and 0-1 loss also showed comparable noise tolerance capability like sigmoid loss function [30]. In this paper, demonstrative experiments have been performed to observe the behavior of three loss functions (Logistic, Savage, Sigmoid) in the presence of white noise, pink noise, and brown noise with and without SVM (Soft margin) classifiers in comparison with hinge loss function for preconditioned and recorded digit (0-9) taken from environment.

Rest of the paper is organized as follow. The consequent section discusses the loss function for soft margin (SME) including sigmoid, savage and logistic loss functions. In section III, Data collection and recording specifications are defined. The experimental results and discussions are presented with pre-conditioned and recorded digits in section IV. Finally, conclusions are drawn in section V.

II. Loss Function for Soft Margin Estimation (SME)

The objective of recognition and classification systems is to minimize the classification risk on testing data samples by developing a classifier f . The concept behind the risk minimization is to measure the performance of estimator by its risk, in order to select best estimator function we should have a measure of inconsistency between an estimated classification $f(x,z)$ and true classification Y(x) of x as shown in (1) and (2) respectively,

$$f(x,z) = y' \tag{1}$$

$$\mathrm{Y}(x) = y \tag{2}$$

The performance of classifier f can be measure using loss function$\mathrm{L(y, f(x,z))}$, which can be defined as;

$$\mathrm{L(y, f(x,z))} = \begin{bmatrix} 0 & \mathrm{if y = f(x,z)} \\ 1 & \mathrm{if y \neq f(x,z)} \end{bmatrix} \tag{3}$$

Consider a risk or estimator function providing the true or expected value of loss as follows:

$$\mathrm{R_{true}(z)} = \int \mathrm{L\big(y, f(x,z)\big) dP(x,y)} \tag{4}$$

Where $\mathrm{R_{true}(z)} = \int \mathrm{L(y, y') dP(x,y)}$ and $\mathrm{P(x,y)} = \mathrm{P(x)\ P(y/x)}$

There is a need to find function $\mathrm{f(x,z)}$ that minimize the risk function $\mathrm{R_{true}(z)}$(for all functions $\mathrm{f(x,z)}, \mathrm{z} \,\mathcal{E}\, \Lambda$), but we don't know $\mathrm{P(x,y)}$. Soft Margin Estimation (SME) [13,22] is a margin-based model estimator applied for speech recognition with an objective to enhance the generalization capability and decision feedback learning by increasing the margin and to enhancing the separation measures of the model in the classifier design respectively. Concept of test risk bound has been defined in statistical learning theory bounded by the summation of two terms: A generalization function and an empirical risk (i.e. risk on the training set) [18]. The generalization of a model is a monotonically increasing function of its VC dimension which is used to measures the complexity of model bounded by decreasing function of margin[18].Soft margin estimation (SME) combines two target optimization function in a single object function based on soft margin estimation,

$$\Lambda_{\mathrm{SME}} = \frac{\lambda}{\rho} + \mathrm{R_{emp}}(\Lambda) = \frac{\lambda}{\rho} + \frac{1}{\mathrm{N}} \sum_{\mathrm{t=1}}^{\mathrm{N}} \mathrm{l}(\mathrm{O_t}, \Lambda) \tag{5}$$

Λ represent the set of HMM based model parameters, $\mathrm{R_{emp}}(\Lambda)$ represent empirical risk, $\mathrm{l}(\mathrm{O_t}, \Lambda)$ defines the loss function for utterance $\mathrm{O_t}$ and N is the total number of training utterances. Whereas, λ and ρ are the coefficient used to balance the empirical risk minimization and margin maximization and soft margin respectively. Margin based acoustic model derived from soft margin concept and hinge loss function used in SVM is defined as loss function to attained enhanced speech recognition performance. Hinge loss function does not perform well in the presence of significant amount of noise. This experimental setup evaluate the performance of hinge loss function in the presence of noise in comparison with three other loss functions with and without SVM classifier for preconditioned isolated digit and digit taken from real environment. The hinge loss function used in SVM can be defined as [22]:

$$\mathrm{l}(\mathrm{d}(\mathrm{O_t}, \Lambda)) = \mathrm{d}(\mathrm{O_t}, \Lambda). \frac{1}{1 + \mathrm{e}^{-\Upsilon \mathrm{d}(\mathrm{O_t}, \Lambda)}} \tag{6}$$

Υ is a positive value number relating to smoothness of loss function and $\mathrm{d}(\mathrm{O_t}, \Lambda)$ represent the for soft margin (SME). Similarly savage loss [31], standard sigmoid loss and logistic loss function can be written as (7), (8) and (9) respectively.

$$\mathrm{l}\big(\mathrm{d}(\mathrm{O_t}, \Lambda)\big) = 1/\big(1 + \mathrm{e}^{-\Upsilon \mathrm{d}(\mathrm{O_t}, \Lambda)}\big)^2 \qquad (7)$$

$$\mathrm{l}\big(\mathrm{d}(\mathrm{O_t}, \Lambda)\big) = \frac{1}{1 + \mathrm{e}^{-\Upsilon \mathrm{d}(\mathrm{O_t}, \Lambda)}} \qquad (8)$$

$$\mathrm{l}\big(\mathrm{d}(\mathrm{O_t}, \Lambda)\big) = \ln\,(1 + \mathrm{e}^{-\Upsilon \mathrm{d}(\mathrm{O_t}, \Lambda)}) \qquad (9)$$

III. Data Collection and Recording Specifications

The methodology of experiments includes the collection of data that comprises of two individual sets; the set of TI-digits (0-9) standard isolated digit corpus and digits recorded from real environment. For the recording specification of recorded isolated digit corpus, ITU recommendations based standardized procedure was adopted for speech corpora development. Standard recording environment has been used having SNR (signal to noise ratio) greater than and equal to 45dB. We made use of Microsoft Windows 7 built-in sound recorder to record the 10 utterances of each isolated digit (0-9). The recording format is Mono, 32 bit PCM with sampling rate of 8000Hz using microphone with impedance of 32 Ω, Max Input power=40mW, Drive Unit=30mm, Plug Type=3.5MM, Frequency Response=20Hz ~ 20 KHz. Microphone with specified configuration were used to take input digits 0 to digit 9 and recorded in noise free recording studio environment. Afterward white noise, brown noise and pink noise were mixed with both sets of data with the help of audacity software. The purpose of noise addition in isolated TI-Digit and recorded digit samples is to study the behavior of each digit under prescribed conditions with and without white, brown and pink noises. The number of experiments was performed with the evaluation of cepstrum coefficient values for each digit in clean and noisy conditions and four graphs of each digit for both recorded and isolated TI-Digit databases were analyzed separately.

IV. Experimental Results and Discussion

In this section, we present demonstrative experiments to show the performance of four loss function such as sigmoid, hinge, savage and logistic in the presence of noise with and without SVM (Soft margin) classifiers to observe the behavior of four different loss functions with pre-conditioned and recorded data samples. Data samples were used in experiments consist of isolated digits taken from TI-Digit corpus [33] and recorded digits taken from real environment with and without SVM (Soft Margin) classifier in the presence of three different types of noises (White, Brown and Pink) taken from NOISEX-92 noise-in-speech database [32,34]. The experimental frame work was divided in two phases: two sets of data have been used with each phase of experiments comprises on isolated TI-Digits (0-9) samples of recorded digits taken from real environment. First stage shows the results of pre-recorded isolated Ti-digits (0-9) without classifier & with addition of three noises. Second stage contains the same results with performance of SVM classifier. In the next phase of experiment, we studied behavior of different loss functions: hinge loss, sigmoid loss, savage loss and logistic individually for data samples of each set in the presence of white, brown and pink noise with and with SVM (Soft Margin) classifiers. The results obtained from experimental analysis represented in loss function comparative analysis charts clearly evident that behavior of loss functions significantly changes for clean & noisy conditions. Furthermore, we present the graphical analysis of different loss function with and without noise to observe behavior of hinge loss, sigmoid loss, savage loss and logistic loss function for both sets of isolated TI-Digit and recorded digit. SVM (Soft Margin) Classifier took values of loss functions values as an input for both data sets separately. Based on the results obtained in the previous steps, classifier classifies the clean sample from noisy samples. For graphical representation of the experimental results in the proceeding sections, digit “one” was selected with and without SVM (soft margin) classifiers in the presence of white, brown and pink noises for both isolated TI-Digit and recorded data samples. Results generated from the experimental framework were based on numerous pieces of code that were implemented & observed in MATLAB tool version 10.0 and speech processing toolbox.

A. Graphical Representation and Loss Functions Comparative Analysis of Isolated Ti-Digit Without and with SVM Classifiers

To evaluate the performance of data sets in clean and noisy conditions, the cepstrum of the each digit in the data set with and without noise were obtained to determine the peak values of cepstrum coefficient. We made use of these cepstrum coefficients to distinct clean digit from the noisy data sample. Isolated TI-Digit “1” was selected for the graphical representation of entire experimental results in this and later sections to illustrate the behavior of the different loss function under certain conditions. The interpretation of graphical results demonstrated by blue line and red line. The blue line/curve represent the plot of loss function value without noise and the red line/curve represent the plot of loss function value with noise. The gap between two lines or curves obtained from the different loss functions indicate the resultant value of error between loss functions with and without noise. When the gap increases between two line/curve, it provides higher value of error which reflects the poor performance in the presence of noise.

The red and blue lines/ curves in the above figures clearly show that behavior of the different loss functions with three different noises. The gap between lines/curve of sigmoid and savage loss is lesser than the hinge and logistic loss. Loss function comparative analysis for all isolated digits (0-9) in the next section indicates that the savage and sigmoid loss functions perform well in comparison with hinge and logistic loss function except for some anomalies. Similarly, the performance of loss functions have been evaluated with isolated TI-Digit in the presence of white, brown and pink noise using SVM (soft margin) classifiers. We made use of SVM classifiers to separate clean TI-Digit from noisy TI-Digit as shown in Fig. 5, Fig. 6 for digit 1 with white and brown noise respectively.

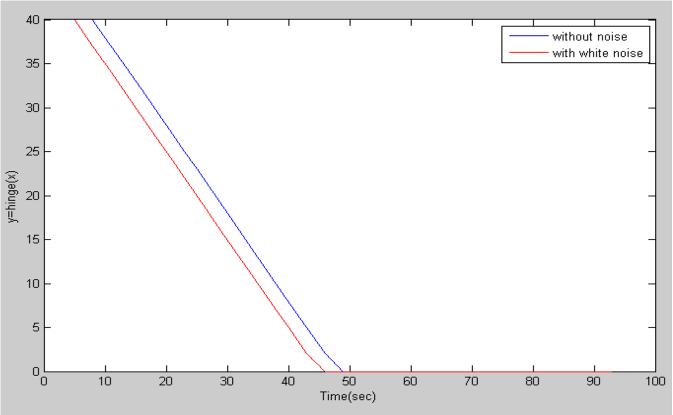


Fig. 1. TI-Digit 1in the presence of white noise without SVM classifier using hinge loss

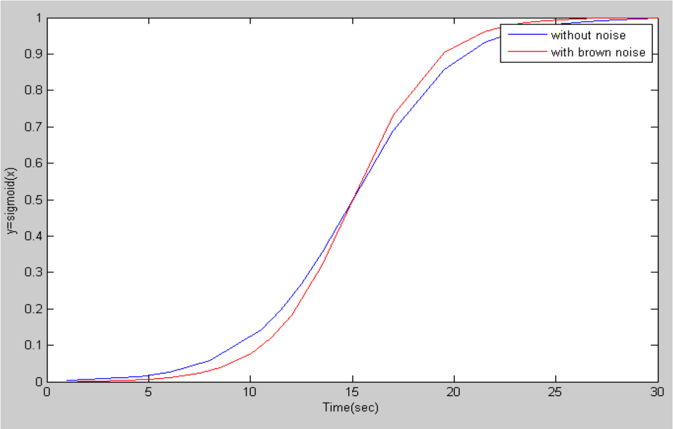


Fig. 2. TI-Digit 1in the presence of brown noise without SVM classifier using sigmoid loss

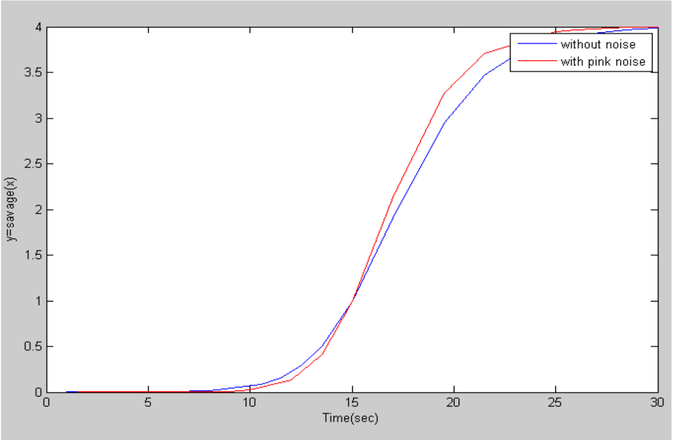


Fig. 3. TI-Digit 1 in the presence of pink noise without SVM classifier using savage loss

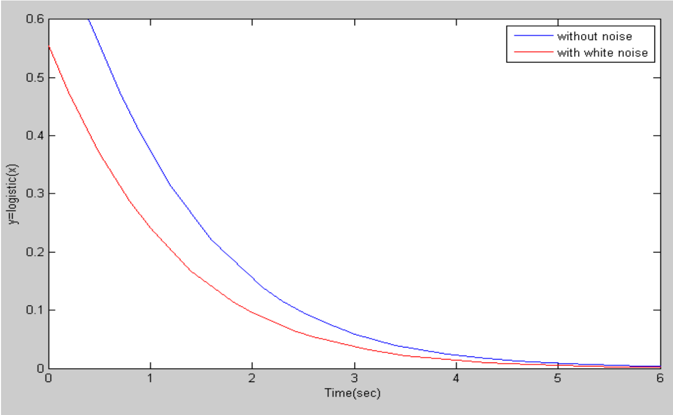


Fig. 4. TI-Digit 1 in the presence of white noise without SVM classifier using logistic loss

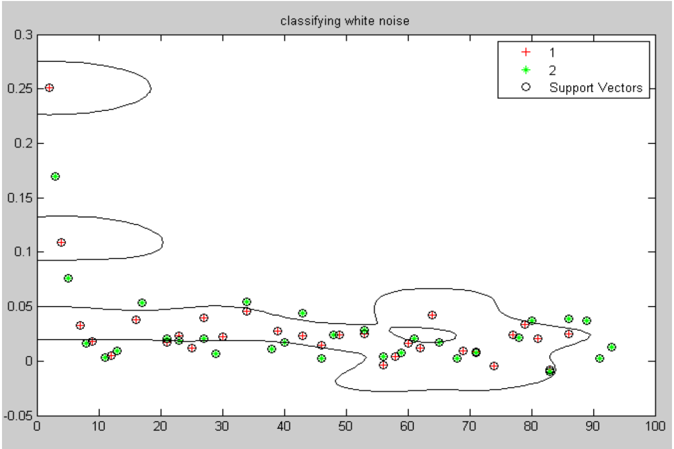


Fig. 5. Plot of TI-digit 1 using SVM classifier with white noise

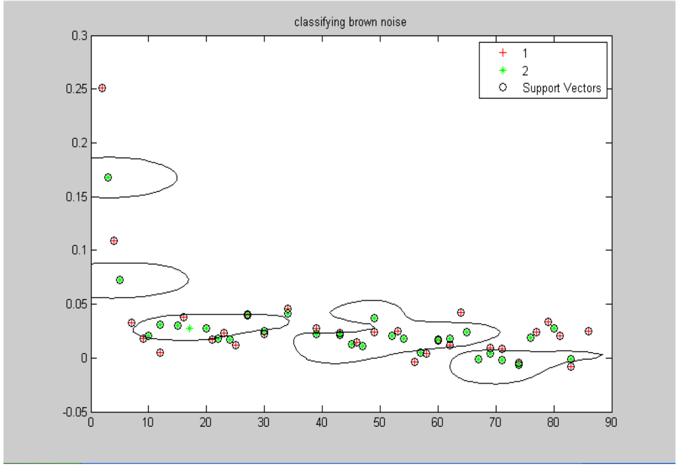


Fig. 6. Plot of TI-digit 1 using SVM classifier with brown noise

The interpretation of the above plots demonstrated by green and red dots which were used to represent the separation of clean and noisy signal respectively using SVM classifiers. The implementation of hinge loss, sigmoid loss, savage loss and logistic loss functions have been done through SVM classifier.

Fig. 7 and Fig. 8 represent hinge and sigmoid plot of Isolated TI-Digit 1 with brown and white noise respectively using SVM classifiers.

Loss functions comparative analysis have been performed among sigmoid loss, hinge loss, savage loss and logistic loss in the presence of white noise, brown noise and pink noise without SVM (Soft Margin) classifiers for Isolated TI-Digit.

The interpretation of experimental results demonstrated by blue bar, red bar, green bar and purple bar for hinge function, sigmoid function, savage function and logistic function respectively.

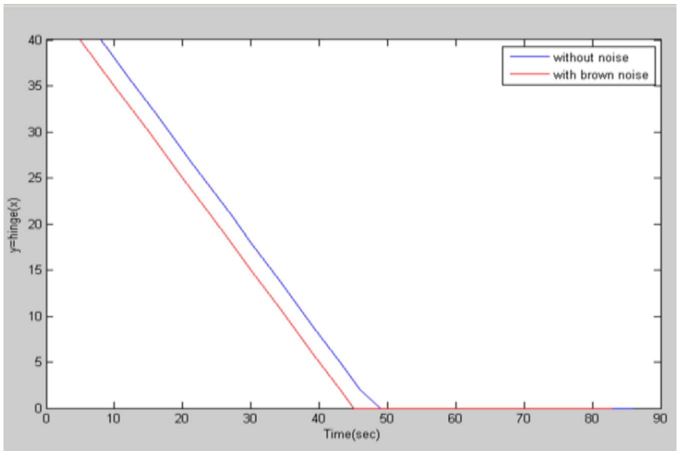


Fig. 7. TI-Digit 1 in the presence of brown noise with SVM classifier using hinge loss

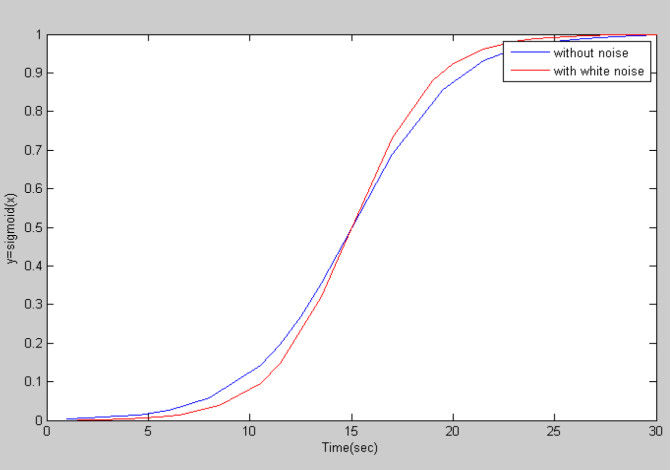
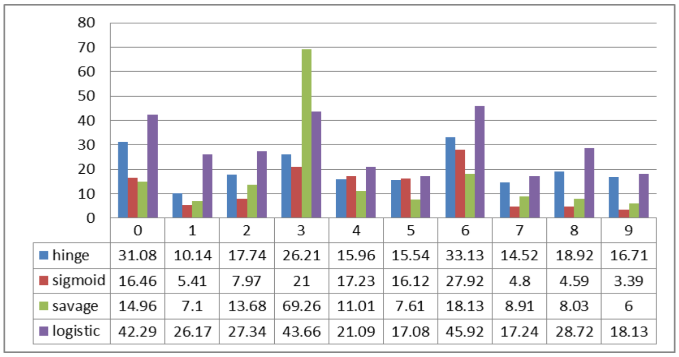


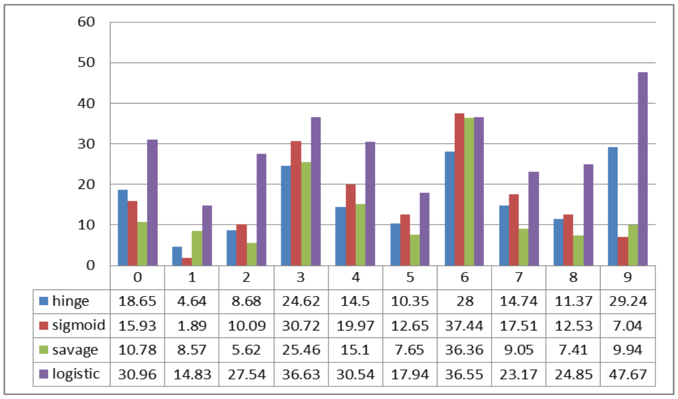
Fig. 8. TI-Digit 1 in the presence of white noise with SVM classifier using sigmoid loss

TABLE I. Loss Functions Comparative Analysis of White Noise for Isolated TI-Digit Without Using Classifier



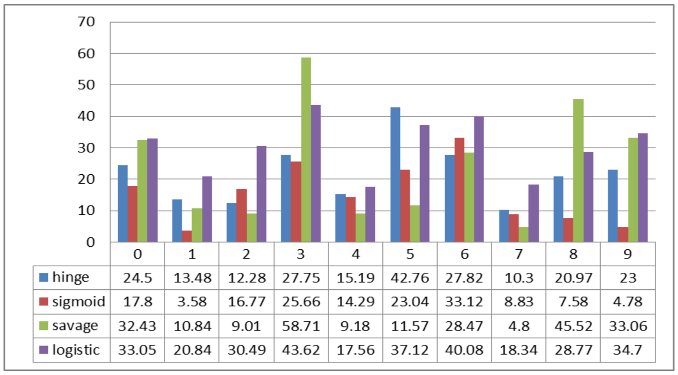
	0	1	2	3	4	5	6	7	8	9
hinge	31.08	10.14	17.74	26.21	15.96	15.54	33.13	14.52	18.92	16.71
sigmoid	16.46	5.41	7.97	21	17.23	16.12	27.92	4.8	4.59	3.39
savage	14.96	7.1	13.68	69.26	11.01	7.61	18.13	8.91	8.03	6
logistic	42.29	26.17	27.34	43.66	21.09	17.08	45.92	17.24	28.72	18.13

TABLE II. Loss Function Comparative Analysis of Brown Noise for Isolated TI-Digit Without Using Classifier



	0	1	2	3	4	5	6	7	8	9
hinge	18.65	4.64	8.68	24.62	14.5	10.35	28	14.74	11.37	29.24
sigmoid	15.93	1.89	10.09	30.72	19.97	12.65	37.44	17.51	12.53	7.04
savage	10.78	8.57	5.62	25.46	15.1	7.65	36.36	9.05	7.41	9.94
logistic	30.96	14.83	27.54	36.63	30.54	17.94	36.55	23.17	24.85	47.67

TABLE III. Loss Function Comparative Analysis of Pink Noise for Isolated TI-Digit Without Using Classifier

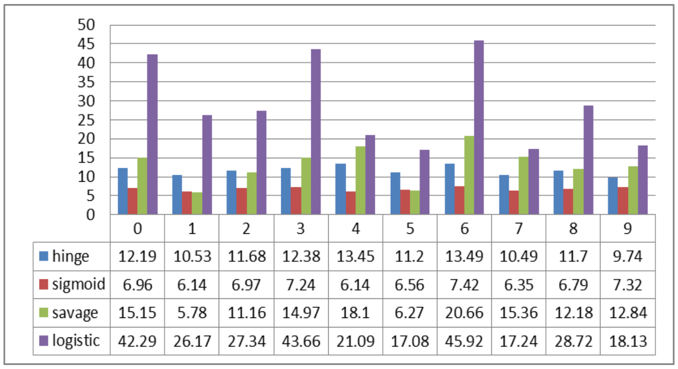


	0	1	2	3	4	5	6	7	8	9
hinge	24.5	13.48	12.28	27.75	15.19	42.76	27.82	10.3	20.97	23
sigmoid	17.8	3.58	16.77	25.66	14.29	23.04	33.12	8.83	7.58	4.78
savage	32.43	10.84	9.01	58.71	9.18	11.57	28.47	4.8	45.52	33.06
logistic	33.05	20.84	30.49	43.62	17.56	37.12	40.08	18.34	28.77	34.7

In Table 1, savage and sigmoid loss function represents substantial anomalies when compared with hinge loss function for digits 3, digit 4 and digit 5. Whereas, perform of hinge loss function quite well than Logistic function. Table 2 indicate that hinge loss function perform well than Logistic loss function in the presence of brown noise, while hinge loss function shows considerable anomalies when compared with savage and sigmoid loss function for digit 1, digit 3, digit 4 and digit 6.

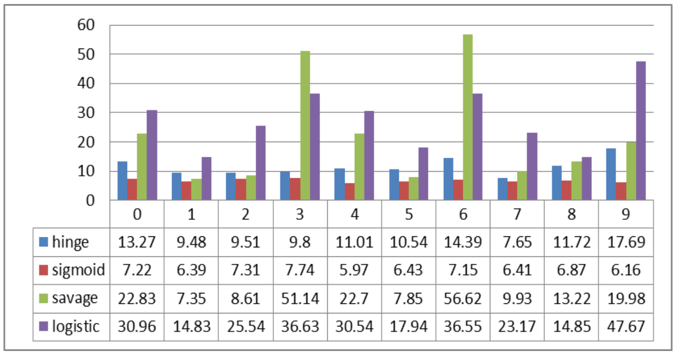
In Table 3, hinge function in the presence of pink noise not performs well than sigmoid function except for digit 2 and digit 6. Whereas, some anomalies have been observed with savage loss function in comparison with hinge loss for digit 0, digit 3, digit 8 and digit 9. Similarity, loss functions comparative analysis have been performed among sigmoid loss, hinge loss, savage loss and logistic loss in the presence of white noise, brown noise and pink noise with SVM (Soft Margin) classifiers for Isolated TI-Digit.

TABLE IV. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF WHITE NOISE FOR ISOLATED TI-DIGIT USING CLASSIFIER



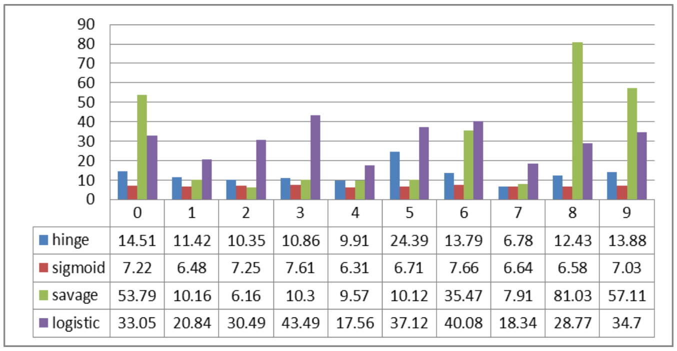
	0	1	2	3	4	5	6	7	8	9
🟦 hinge	12.19	10.53	11.68	12.38	13.45	11.2	13.49	10.49	11.7	9.74
🟥 sigmoid	6.96	6.14	6.97	7.24	6.14	6.56	7.42	6.35	6.79	7.32
🟩 savage	15.15	5.78	11.16	14.97	18.1	6.27	20.66	15.36	12.18	12.84
🟪 logistic	42.29	26.17	27.34	43.66	21.09	17.08	45.92	17.24	28.72	18.13

TABLE V. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF BROWN NOISE FOR ISOLATED TI-DIGIT USING CLASSIFIER



	0	1	2	3	4	5	6	7	8	9
🟦 hinge	13.27	9.48	9.51	9.8	11.01	10.54	14.39	7.65	11.72	17.69
🟥 sigmoid	7.22	6.39	7.31	7.74	5.97	6.43	7.15	6.41	6.87	6.16
🟩 savage	22.83	7.35	8.61	51.14	22.7	7.85	56.62	9.93	13.22	19.98
🟪 logistic	30.96	14.83	25.54	36.63	30.54	17.94	36.55	23.17	14.85	47.67

TABLE VI. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF PINK NOISE FOR ISOLATED TI-DIGIT USING CLASSIFIER



	0	1	2	3	4	5	6	7	8	9
🟦 hinge	14.51	11.42	10.35	10.86	9.91	24.39	13.79	6.78	12.43	13.88
🟥 sigmoid	7.22	6.48	7.25	7.61	6.31	6.71	7.66	6.64	6.58	7.03
🟩 savage	53.79	10.16	6.16	10.3	9.57	10.12	35.47	7.91	81.03	57.11
🟪 logistic	33.05	20.84	30.49	43.49	17.56	37.12	40.08	18.34	28.77	34.7

In Table 4, hinge is better than Logistic function for all digits but severe anomalies can be seen when comparing performance with other loss functions. Table 5 shows that hinge is better than Logistic function for all digits but severe anomalies can be seen when comparing performance with other loss functions. In Table 6, logistic loss function not performs well than hinge loss, but considerable anomalies have been observed when hinge loss function compared with savage loss function.

B. Graphical Representation of Loss Functions Comparative Analysis of Recorded Digit without and with SVM Classifiers

The performance of the loss functions have been evaluated in this section, using recorded digit samples taken from environment to study the behavior of the hinge loss, sigmoid loss, savage loss, and logistic in the presence of noise. Fig. 9 and Fig. 10 represent hinge and logistic plot of Isolated TI-Digit 1 in the presence of pink and pink noise respectively without SVM classifiers.

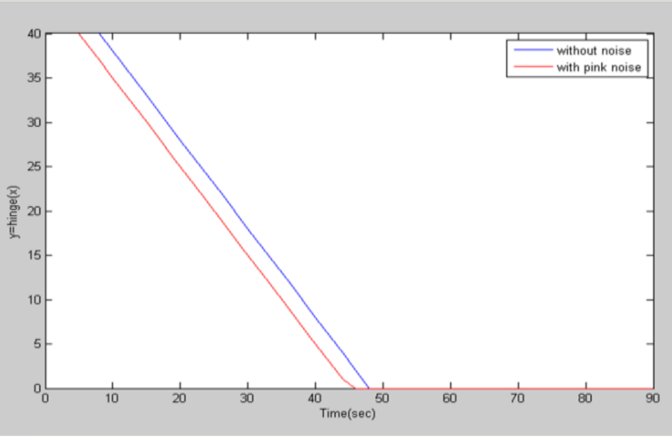


Fig. 9. Recorded digit 1 in the presence of pink noise without SVM classifier using hinge loss

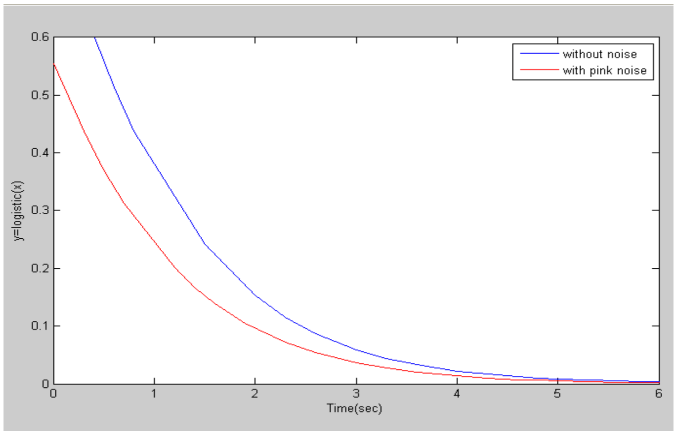


Fig. 10. Recorded digit 1 in the presence of pink noise without SVM classifier using logistic loss

Similarly, we evaluated the performance of loss functions with recorded digit taken from environment in the presence of white, brown and pink noise using SVM classifiers. We made use of SVM classifiers to separate clean recorded digit sample from noisy samples. Fig. 11 and Fig. 12 represent hinge and sigmoid plot of recorded digit 1 with brown and pink noise respectively with SVM classifiers.

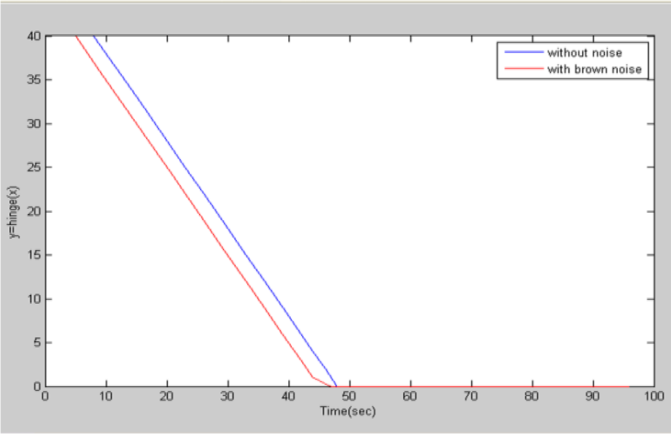


Fig. 11. Recorded digit 1 in the presence of brown noise with SVM classifier using hinge loss

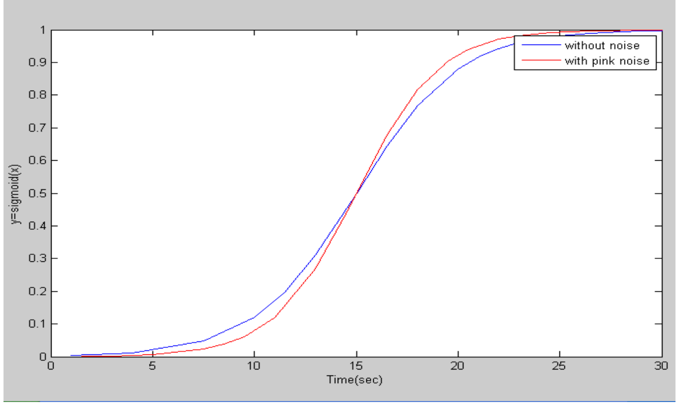
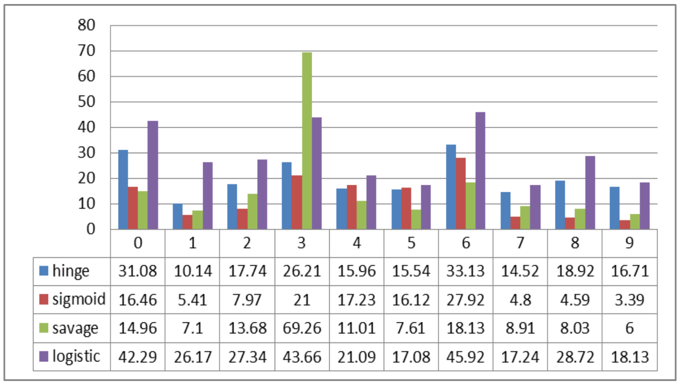


Fig. 12. Recorded digit 1 in the presence of pink noise with SVM classifier using sigmoid loss

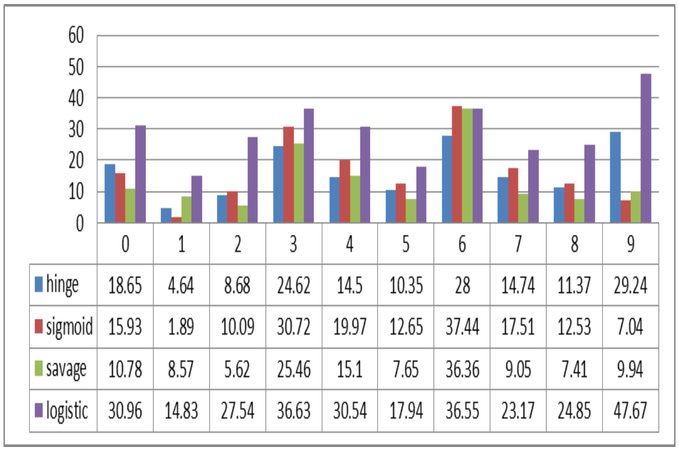
Loss function comparative analysis among hinge loss, sigmoid loss, savage loss and logistic loss have been performed in the presence of white noise, pink noise and brown noise without SVM (Soft Margin) classifiers for recorded digits taken from real environment.

TABLE VII. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF WHITE NOISE FOR RECORDED DIGIT WITHOUT CLASSIFIER



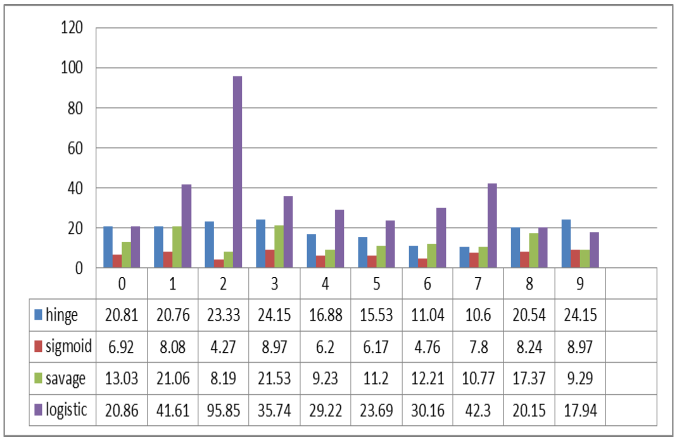
	0	1	2	3	4	5	6	7	8	9
🟦 hinge	31.08	10.14	17.74	26.21	15.96	15.54	33.13	14.52	18.92	16.71
🟥 sigmoid	16.46	5.41	7.97	21	17.23	16.12	27.92	4.8	4.59	3.39
🟩 savage	14.96	7.1	13.68	69.26	11.01	7.61	18.13	8.91	8.03	6
🟪 logistic	42.29	26.17	27.34	43.66	21.09	17.08	45.92	17.24	28.72	18.13

TABLE VIII. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF BROWN NOISE FOR RECORDED DIGIT WITHOUT CLASSIFIER



	0	1	2	3	4	5	6	7	8	9
🟦 hinge	18.65	4.64	8.68	24.62	14.5	10.35	28	14.74	11.37	29.24
🟥 sigmoid	15.93	1.89	10.09	30.72	19.97	12.65	37.44	17.51	12.53	7.04
🟩 savage	10.78	8.57	5.62	25.46	15.1	7.65	36.36	9.05	7.41	9.94
🟪 logistic	30.96	14.83	27.54	36.63	30.54	17.94	36.55	23.17	24.85	47.67

TABLE IX. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF PINK NOISE FOR RECORDED DIGIT WITHOUT CLASSIFIER



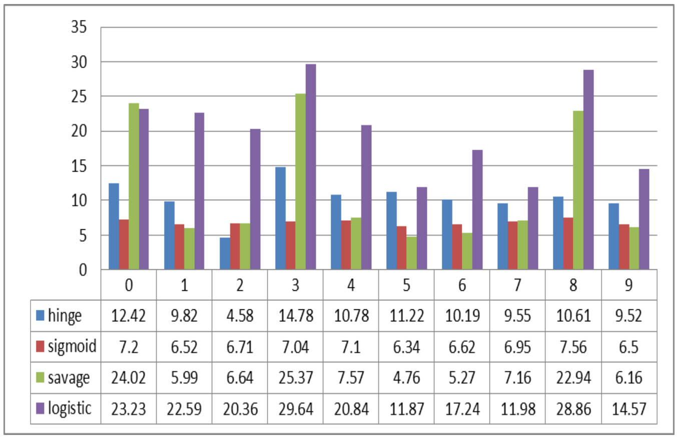
	0	1	2	3	4	5	6	7	8	9
🟦 hinge	20.81	20.76	23.33	24.15	16.88	15.53	11.04	10.6	20.54	24.15
🟥 sigmoid	6.92	8.08	4.27	8.97	6.2	6.17	4.76	7.8	8.24	8.97
🟩 savage	13.03	21.06	8.19	21.53	9.23	11.2	12.21	10.77	17.37	9.29
🟪 logistic	20.86	41.61	95.85	35.74	29.22	23.69	30.16	42.3	20.15	17.94

Table 7, Table 8 and Table 9 provide the comparative analysis of hinge, sigmoid, savage and logistic loss function for recorded digit without SVM (Soft Margin) Classifiers. In Table 7, savage and sigmoid loss function performs well than hinge loss function in the presence of white noise except digits 3, digit 4 and digit 5.

In Table 8, serious anomalies have been observed with hinge loss function when compared with savage and sigmoid in some digits with brown noise. Table 9 displays that savage and sigmoid perform well in comparison with hinge loss except digit l and digit 7 whereas logistic function not performs well than hinge loss function.

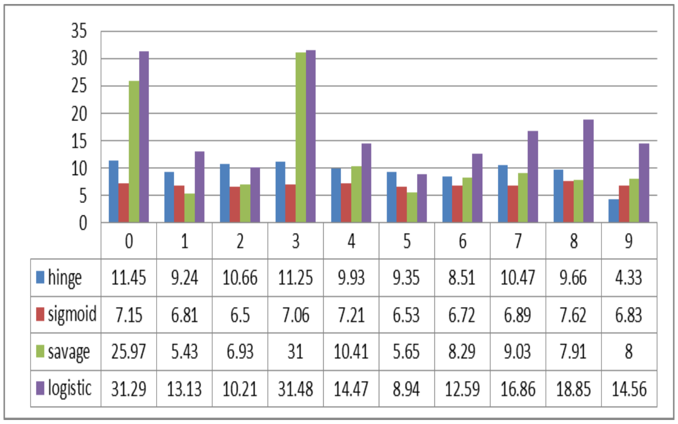
Similarly, loss function comparative analysis among hinge loss, sigmoid loss, savage loss and logistic loss in the presence of white noise, brown noise and pink noise with SVM (Soft Margin) classifiers for recorded digits taken from real environment.

TABLE X. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF WHITE NOISE FOR RECORDED DIGIT WITH CLASSIFIER



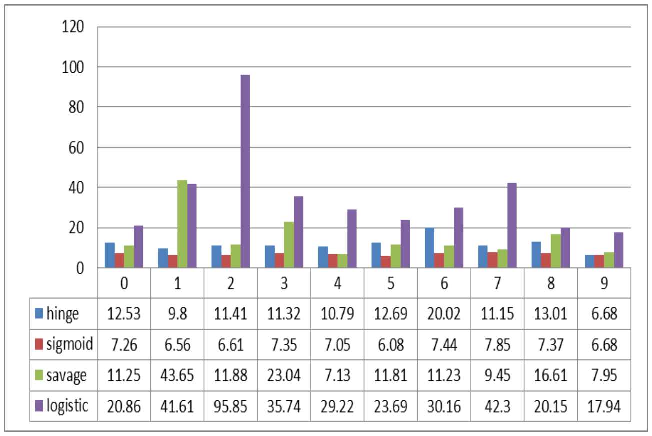
	0	1	2	3	4	5	6	7	8	9
🟦 hinge	12.42	9.82	4.58	14.78	10.78	11.22	10.19	9.55	10.61	9.52
🟥 sigmoid	7.2	6.52	6.71	7.04	7.1	6.34	6.62	6.95	7.56	6.5
🟩 savage	24.02	5.99	6.64	25.37	7.57	4.76	5.27	7.16	22.94	6.16
🟪 logistic	23.23	22.59	20.36	29.64	20.84	11.87	17.24	11.98	28.86	14.57

TABLE XI. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF BROWN NOISE FOR RECORDED DIGIT WITH CLASSIFIER



	0	1	2	3	4	5	6	7	8	9
🟦 hinge	11.45	9.24	10.66	11.25	9.93	9.35	8.51	10.47	9.66	4.33
🟥 sigmoid	7.15	6.81	6.5	7.06	7.21	6.53	6.72	6.89	7.62	6.83
🟩 savage	25.97	5.43	6.93	31	10.41	5.65	8.29	9.03	7.91	8
🟪 logistic	31.29	13.13	10.21	31.48	14.47	8.94	12.59	16.86	18.85	14.56

TABLE XII. LOSS FUNCTIONS COMPARATIVE ANALYSIS OF PINK NOISE FOR RECORDED DIGIT WITH CLASSIFIER



	0	1	2	3	4	5	6	7	8	9
🟦 hinge	12.53	9.8	11.41	11.32	10.79	12.69	20.02	11.15	13.01	6.68
🟥 sigmoid	7.26	6.56	6.61	7.35	7.05	6.08	7.44	7.85	7.37	6.68
🟩 savage	11.25	43.65	11.88	23.04	7.13	11.81	11.23	9.45	16.61	7.95
🟪 logistic	20.86	41.61	95.85	35.74	29.22	23.69	30.16	42.3	20.15	17.94

Table 10, Table 11and Table 12 provide the comparative analysis of hinge, sigmoid, savage and logistic loss function for recorded digit with SVM (Soft Margin) Classifiers. In Table 10, hinge loss not perform well as compared to savage and sigmoid loss function but some anomalies have been observed with digit 0, digit 2, digit 3 and digit 8 in the presence of white noise. In Table 11, hinge loss performs well than logistic loss except for digit 2 whereas, some anomalies have been observed with digit 0, digit 3, digit 4 and digit 9 in the presence of brown noise when compared with savage and sigmoid function. Table 12 indicates that logistic loss function not perform well in comparison with hinge loss but savage function and sigmoid function perform well than hinge loss function except for digit 1, digit 3 and digit 8. The following observation has been acquired from demonstrative experiments:

- Pink noise and white noise illustrate significant effects on isolated digits (0-9) in both pre-conditioned and recorded conditions in comparison with brown noise. Hinge loss function doesn't perform well than sigmoid loss and savage loss functions but it performs better than logistic loss function in the presence of white and pink noise, however, some anomalies are observed in the presence of brown noise.

- In all four prescribed conditions in demonstrative experiments for both recorded digits taken from environment and pre-conditioned TI-Digits with and without classifiers, logistic loss function not perform well in comparison with hinge loss function whereas hinge loss function show serious anomalies with savage loss sigmoid and functions in term of performance.

- In comparison with hinge loss and Logistic loss functions, sigmoid loss function provides exceptionally good results in term of percentage error for all prescribed conditions in experiments. Whereas, few inconsistencies can be seen in the performance of savage loss function in comparison with hinge loss.

V. CONCLUSION

Motivated by the issue related to hinge loss function used in SVM for parameter estimation in margin based acoustic model, this paper presented the comparative analysis of three loss functions (Logistic, Savage, Sigmoid) in comparison with hinge loss to observe the behavior of loss functions in the presence of white noise, pink noise, and brown noise with and without SVM (Soft margin) classifiers for preconditioned and recorded data samples. Demonstrative experiments have been made on NOISEX-92 (speech and noise-in speech) databases, TIDIGIT corpus and recorded data samples (0-9) taken from real environment. The demonstrative experiments indicated that hinge loss function doesn't perform well than savage loss and sigmoid loss functions but it performs better than logistic loss function in the presence of pink and white noise as compared to brown noise for all prescribed conditioned. Whereas, sigmoid loss function shows remarkably better results in comparison with hinge and other loss function in term of percentage error.

REFERENCES

[1] X. He and L. Deng, Discriminative Learning for Speech Recognition: Theory and Practice. Morgan & Claypool, 2008.

[2] Y. Normandin, “Maximum Mutual Information Estimation of Hidden Markov Models,” In Automatic Speech and Speaker Recognition, edited by C. H. Lee, F. K. Soong and K. K. Paliwal, Eds. Kluwer Academics Publishers, Norwell, M.A, 1996, pp.1-159.

[3] L R. Bahl, P. F. Brown, P. V. Desouza and R. L. Mercer, “Maximum Mutual Information Estimation of Hidden Markov Models parameters for speech recognition,” In Proc. ICASSP, vol.1, pp. 49-52, 1986.

[4] V. Valtchev, J. Odell, P. Woodland and S. Young, “Maximum Mutual Information Estimation training for large vocabulary recognition systems,” Speech Communication, vol.22, no.4, pp. 303-314, 1997.

[5] B. -H. Juang, W. Chou, and C.-H. Lee, “Minimum Classification Error rate methods for speech recognition,” IEEE Trans. on Speech and Audio Proc. vol.5, no.3, pp.257-265, 1997.

[6] B.-H. Juang, and S. Katagiri, “Discriminative learning for Minimum Error Classification,” IEEE Trans. on Signal Processing, vol.40, pp.3043-3054, 1992.

[7] R. Schlueter, W. Macherey, B. Muller, and H. Ney “Comparison of discriminative training criteria and optimization methods for speech recognition,” Speech Communication, vol.34, pp.287-310, 2001.

[8] D. Povey and P. Woodland, “Minimum Phone error and I-smoothing for improved discriminative training,” In Proc. ICCASP, vol.1, pp. 105-108, 2002.

[9] D. Povey, Discriminative training for large vocabulary speech recognition, Ph.D. dissertation, Cambridge University, Dept. Eng., Cambridge, UK, 2004.

[10] H. Jiang, X. Li and C. Liu, “Large Margin Hidden Markov models for speech recognition,” IEEE Trans. on Audio, Speech and Language Processing, vol.14, no.5, pp.1584-1595, 2006.

[11] D. Yu, L. Deng, X. He and A. Acero, “Large Margin minimum classification training for Large-Scale Speech Recognition Tasks,” In Proc. ICASSP, 2007.

[12] H. Jiang and X. Li, “Solving large margin HMMs estimation via semi-definite programming,” In Proc. ICASSP, vol.4, no.5, pp.IV-629-IV-632, 2007.

[13] J. Li, M. Yuan and C.-H. Lee, “Soft margin estimation of Hidden Markov Model parameters,” In Proc. Interspeech, pp.2422-2425, 2006.

[14] X. Li, and H. Jiang, “A constrained joint optimization methods for large margin HMM estimation,” In Proc. ASRU Workshop, pp.151-156, 2005.

[15] C. Liu, H. Jiang and L. Rigazio, “Recent improvement of minimum relative margin estimation of HMMs for speech recognition,” In Proc. ICASSP, vol.1, pp.269-272, 2006.

[16] F .Sha, and L. Saul, “Large-Margin Gaussian mixture modeling for phonetic classification and recognition,” In Proc. ICASSP, vol.1, pp.265-268, 2006.

[17] D. Yu, L. Deng, X. He and A. Acero, “Use of incrementally regulated discriminative margins in MCE training for speech recognition,” In Proc. Interspeech, pp. 2418-2421, 2006.

[18] V. Vapnik, The nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[19] B. Scholkopf and A. J. Smola, Learning with Kernels: support vector machine, regularization, optimization, and beyond. Cambridge: The MIT Press, 2002.

[20] O. Bousquet, S. Bouchern and G. Lugosi, “Introduction to statistical learning theory. Advanced lectures on machine learning lecture notes in artificial intelligence 3176, pp.167-207. (Eds) Springer, Heidelberg, Germany, 2004.

[21] C. Liu, H. Jiang and X. Li “Discriminative training of CDHMMS for maximum relative separation margin,” In Proc. ICASSP, pp.1101-1104, 2005.

[22] J. Li and C. -H. Lee, “Soft margin feature extraction for automatic speech recognition”, Proc. Interspeech, 2007.

[23] C. Burges, “A tutorial on support Vector machine for pattern recognition,” Data Mining and Knowledge Discovery, vol.2, no. 2, pp.121-167, 1998.

[24] J. Li, C.-H. Lee and R. H. Wang, “A Study of Soft margin estimation for LVCSR,”In Proc. IEEE automatic speech recognition and understanding workshop, pp.268-271, 2007.

[25] R. K. Aggarwal and M. Dave, “Acoustic modeling problem for ASR system: advances and refinements (Part II),” International Journal of Speech Technology, Springer, pp.309- 320, 2011.

[26] S.A. Ali, N. G. Haider and M. K. Pathan., “Margin Based Learning: A Framework for Acoustic Model Parameter Estimation”, I.J. Intelligent Systems and Applications, vol. 2, no.12, pp. 26-31, November 2012.

[27] X. Xiao, J. Li, E. S. Chng, H. Li, C. H. Lee, “ A study on the generalization capability of Acoustic modes for Robust Speech Recognition,” IEEE Trans. on Audio, Speech and Language Processing, Vol. 18, No.6, August 2010.

[28] O. Viikki, K. Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” Speech Communication, vol.25, pp.133-147, 1998.

[29] S.A. Ali and N.G. Haider, “Margin Based Learning Framework with Geometric Margin Minimum Classification Error for Robust Speech Recognition,” International Journal of Sciences: Basic and Applied Research (IJSBAR), vol.11, No 1, pp. 39-48, 2013.

[30] A. Gosh, N. Manwani, and P. S. Sastry. Making risk minimization tolerance to Label Noise. CoRR, abs/1403.3610, March 2014.

[31] H. Masnadi-Shirazi and N. Vasconcelos, “On the design of loss functions for classification: theory, robustness to outliers, and Savage Boost”, Advances NIPS, 2007.

[32] A.P. Varga, H. J. M Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition,” In Technical Report, DRA Speech Research Unit, 1992.

[33] Clean digits are from “TIDIGIT Corpus”, available at www.ee.columbia.edu/~dpwe/sounds/tidigits/

[34] Noise Corpus are from “NOISEX-92” http://Spib.rice.edu/spib/select_noise.

Improving Quality of Vietnamese Text Summarization Based on Sentence Compression

Ha Nguyen Thi Thu
Information Technology Faculty
Electric Power University
Hanoi, Vietnam

Cuong Nguyen Ngoc
Department of Computer and Mathematics
The people's Security University
Hanoi, Vietnam

Tu Nguyen Ngoc
Information Technology Faculty
Electric Power University
Hanoi, Vietnam

Hiep Xuan Huynh
Information Technology Faculty
CanTho university
Hanoi, Vietnam

***Abstract*—Sentence compression is a valuable task in the framework of text summarization. In previous works, the sentence is reduced by removing redundant words or phrases from original sentence and tries to remain information. In this paper, we propose a new method that used Grid Model and dynamic programming to calculate n-grams for generating the best sentence compression. These reduced sentences are combined to text summarization. The experimental results showed that our method really effective and the text is grammatically, coherence and concise.**

Keywords—Sentence compression; topic modeling; text summarization; Grid model; n-grams; dynamic programming

I. Introduction

Text summarization is technique allows computers automatically generated text summaries from one or more different sources. To base oneself on features of the main content and to recapitulate content from original documents that text summarization is one of the fields is interested in researchers from the 60's of the 20th century and it is still a hot topic of the forums and seminars on the current world [1].

The traditional text summarization method usually bases on extracted sentences approach [1], [9]. Summary is made up of the sentences were selected from the original. Therefore, in the meaning and content of the text summaries are usually sporadic, as a result, text summarization lack of coherent and concise. Figure 1 below illustrates the approach to extract sentences. Text summarization has one sentence with extraction rate 30%.

Vừa qua, Microsoft đã chính thức ra mắt người dùng bản cập nhật Service Pack đầu tiên dành cho Windows 7. Nhưng liệu chúng ta có nên tốn thời gian để cài đặt bản cập nhật này cho Windows không. Bạn chưa bao giờ cập nhật các bản Service Pack dành cho Windows của mình. Trước tiên, chúng ta cần lưu ý rằng Service Pack thực chất chỉ là một gói tổng hợp các bản cập nhật nhỏ lẻ đã được tung ra từ trước đó thông qua Windows Update với việc sửa chữa các lỗi và bổ sung một vài thay đổi. Microsoft tung ra Service Pack nhằm mục đích giúp những người dùng phải cài lại hệ điều hành vì một lý do nào đó có thể cập nhật Windows một cách nhanh chóng chỉ thông qua một gói tải về duy nhất.Tăng cường chất lượng âm thanh của thiết bị dùng cổng HDMI. In tài liệu XPS với nhiều khổ giấy khác nhau. Thay đổi hành động của tính năng. Hỗ trợ Advanced Vector Extensions. Hỗ trợ Advanced Format 512e cho các thiết bị lưu trữ.

Figure 1.a. Original text

Thay đổi hành động của tính năng. Hỗ trợ Advanced Vector Extensions. Hỗ trợ Advanced Format 512e cho các thiết bị lưu trữ.

Figure 1.b. Result with extraction rate 30%

Fig. 1. Extraction summary

Some other text summarization methods are the problem of natural language processing that made summary has a good linguistic score and seamlessly coherence the content of the original. One of its is a sentence compression technique [2], [3], [7]. With the compression approach, researchers focused using supervised learning techniques or using legal vocabulary or deep level language analysis techniques based on syntax tree [10]. These methods have the following characteristics:

- High cost when building the corpus for training.
- Need a long time for construction meticulously by language experts, especially construction corpus related legal vocabulary.
- Higher computational complexity.

Therefore, in this paper, we use a sentence compression method to create a text summary basing on grid model with the target:

- Use unsupervised learning to reduce costs.

- Use unsupervised learning techniques to not waste time to build corpus crafts.
- Minimize computational complexity by using dynamic programming algorithm.

The rest of the paper is organized as follows: In section 2, we will introduce some related works. In section 3 is the presentation of our method for Vietnamese feature reduction, the methodology of Vietnamese sentence compression is presented in section 4. Experiments and results will show in section 5. And finally, section 6 is a conclusion and future works.

II. Related Works

The sentence compression task is defined as the curtailment of redundant components, in sentence to produce a shorter sentence. Figure 2 below is an example

Tháng 11 ~~năm ngoái~~, Quốc hội Việt Nam thông qua dự án xây hai nhà máy điện hạt nhân ~~đầu tiên của Việt Nam~~ tại ~~tỉnh~~ Ninh Thuận.

Figure 2.a. Original sentence

Tháng 11, Quốc hội Việt Nam thông qua dự án xây hai nhà máy điện hạt nhân tại Ninh Thuận.

Figure 2.b. Reduced sentence

Fig. 2. The task of sentence compression

Text summarization is based on a sentence compression approach, allows connecting multiple sentences were reduced to make a shorter document that have the meaning and grammar are accepted, to guarantee a coherent level of content and meaning.

Some studies of sentence compression showed the importance of this approach in the problem text summarization. The first people who proposed sentence compression model is Jing and McKeown in 2000, they presented one of the earliest approaches on sentence compression using machine learning and classifier based techniques. This research work was focused on removing inessential phrases in extractive summaries based on an analysis of human written abstracts. In their experiments, they used human-written abstracts, and the corpus was collected from the free daily news and headlines provided by the Benton Foundation [4].

Noisy channel is the typical of the sentence compression method, in studies of Marcu et al., they suggested two methods for sentence compression: one is the noisy channel model where the probabilities for sentence compression (P{compress|S)} 1) are estimated from a training set (Sentence, Sentencecompress) pairs, manually crafted, while considering lexical and syntactical features. The other approach learns syntactic tree rewriting rules, defined through four operators: SHIFT, REDUCE DROP and ASSIGN [9].

In the work of Le Nguyen and Ho in 2004, two sentence compression algorithms were also proposed. The first one is based on template translation learning, a method inherited from the machine translation field, which learns lexical transformation rules, by observing a set of 1500 (Sentence, Sentencereduced) pairs, selected from a website and manually tuned to obtain the training data. Due to complexity difficulties found in the application of this big lexical rule set, they proposed an improvement where a stochastic Hidden Markov Model is trained to help in the decision of which sequence of possible lexical reduction rules should be applied to a specific case [11].

Some other works used unsupervised approach. Turner and Charniak, in their work, corpus for training are automatically extracted from the Penn Treebank corpus, to fit a noisy channel model [3], similar to the one used by Knight and Marcu [8]. And Clarke and Lapata devise a different and quite curious approach, where the sentence compression task is defined as an optimal goal, from an Integer Programming problem. Several constraints are defined, according to language models, linguistic, and syntactical features. Although this is an unsupervised approach, without using any parallel corpus, it is completely knowledge driven, like a set of craft rules and heuristics incorporated into a system to solve a certain problem [2].

All these works applied to English. For Vietnamese, there are some methods for sentence compression. Minh Le Nguyen et.al and Ha Nguyen Thi Thu et. al. Minh Le Nguyen proposed two methods for sentence compression, one of its applied HMM to Vietnamese sentence compression and other used syntax control for reducing sentences [10], [11]. Ha Nguyen Thi Thu using unsupervised learning and supervised learning for creating Vietnamese text summarization based on sentence compression [5], [6].

III. Vietnamese Text Feature Reduction

A. Feature reduction problem

Considering a number of applications such as in a data processing system (the voice signal, image or pattern recognition generally) if we consider setting of features as a set of vectors of real value. Assuming that the system is only effective if the dimension vector of each individual is not too large.

The problem of dimensionality reduction occurs when data have greater dimension processing capabilities of the system [16]. Example: A face recognition/classification system based on multi-level gray image that has size mxn, corresponding to mxn dimensional vector of real value. In the experiment, an image may have m = n = 256 or 65536 dimensions. If using a multilayer perceptron network to perform the classification system. It will become difficult to build an MLP.

Therefore, the feature reduction is an important problem when we work with the data that has many features such as image, voice, text, ... The feature reduction illustrated like following figure

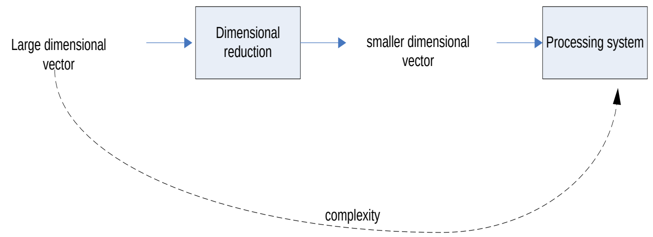


Fig. 3. Dimensional vector reduction model

B. Methodology of Vietnamese text feature reduction

Define 1: (topic word): Topic word is the nouns that have been extracted from sentences.

Example 1: *table, human, computer,…* is the topic words

For the Vietnamese text, some text processing problem often uses word segmentation tool for separate words in text. In previous works, we proposed a method for feature reduction that is published in [5], [6]. Documents can reduce complexity computing of large feature set by using a word, segmentation tool for separating word into two word sets: nouns set (called topic word) and other words set. In any text, nouns contain information of the text. So, when we extract nouns from text, a remarkable reduction of large feature set.

Example 2: Have an original Vietnamese text include 34 words.

"*Các nhà nghiên cứu thuộc trường Đại học Michigan vừa tạo ra một nguyên mẫu đầu tiên cho hệ thống tính toán quy mô nhỏ, có thể chứa dữ liệu một tuần khi tích hợp chúng vào trong những bộ phận rất nhỏ như mắt người.*"

Translate in to English:

"Researchers at the University of Michigan have created a first prototype system for small-scale computing, which can contain data for a week while integrating them into very small parts as the human eyes."

Like this document, we must calculate weight for 34 words. And representation matrix with 1 row and 34 columns like below

$$T = \{t_{1,1}, t_{1,2}, ..., t_{1,34}\} \quad (1)$$

In Example 2, we separate document d into two sets, the first set include noun and the second set is remain of words.

Noun set T' = {*nhà, nghiên_cứu, trường, đại_học, Michigan, nguyên_mẫu, hệ_thống, quy_mô, dữ_liệu, tuần, chúng, bộ_phận, mắt, người*}.

Other set O'= {*Các, thuộc, vừa, tạo, ra, một, đầu_tiên, cho, tính_toán, nhỏ, có_thể, chứa, một, khi, tích_hợp, vào, trong, những, rất, như* }.

Use text separation technique in two sets, the size of the matrix T will be reduced, for example, with the original text in example 1, instead of using the T matrix contains one row and 34 columns, we only need the matrix T' consists of one row and 14 columns:

$$T = \{t_{1,1}, t_{1,2}, ..., t_{1,14}\} \quad (2)$$

IV. Using Grid Model for Vietnamese Text Summarization

Methods Vietnamese sentence compression based on unsupervised learning techniques with grid model combined with dynamic programming to choose the best sentence shortened. The calculation is based on the set of noun in a sentence that limit the loss of information in a sentence.

To calculate the probability of a sequence P $(w_1, w_2, .., w_n)$. Use chain rule of probability:

$$P(X_1...X_n) = P(X_1)P(X_x | X_1)P(X_3 | X_1^2)...P(X_n | X_1^{n-1}) = \prod_{k=1}^{n} P(X_n | X_1^{k-1}) \quad (3)$$

Apply chain rule for the words, to receive:

$$P(w_n^n) = P(w_1)P(w_2 | w_1)P(w_3 | w_1^2)...P(w_n | w_1^{n-1}) = \prod_{k=1}^{n} P(w_n | w_1^{k-1}) \quad (4)$$

With N - grams, the conditional probability approximation of the next word in the sequence is

$$P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-N+1}^{n-1}) \quad (5)$$

In this paper, we use bi-gram to reduce complexity in calculation, therefore, when using the bi-gram model to predict the conditional probability of the next word can use approximately formula as follows:

$$P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-1}) \quad (6)$$

Define 2. (Word substring): *Word substring is initialized by a topic word and stop with a topic word, no any topic word between its.*

Example 3: In his memory, **Flowers** bloom along the **river.**

Define 3. (**The most Likelihood word substring**) *is the word substring in which, every word has maximum bi-gram.*

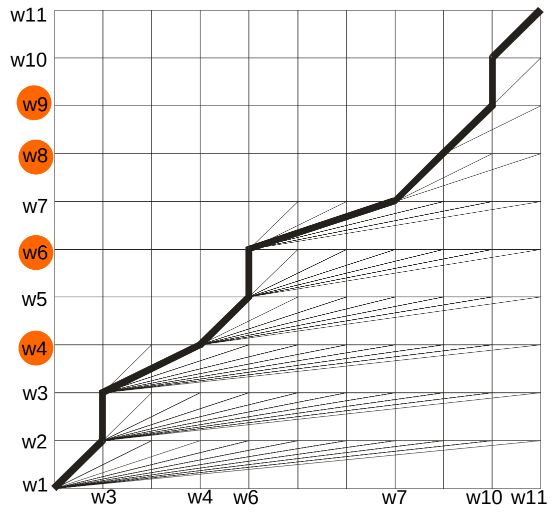


Fig. 4. Grid model for a 11 words

Suppose there are 11 word in the sentence *W* , *W* is represented by :

$$W = \{w_1, w_2, ..., w_{11}\} \qquad (7)$$

Using word identification and word segmentation tools into two sets of a separate word, inside the collective noun includes words w_4, w_6, w_8, w_9. The grid model with this sentence is created as Figure 4 below

In the figure 4 illustrated a original sentence, that has 11 words. In which, w_4, w_6, w_8, w_9 are all the topic words (nouns). We have 3 word substrings. Reduced sentence can be generated by these steps:

- Step 1: Let start with the first word substring $w_1..w_4$: Initialize with w_1, w_1 has 3 ways: $w_1 \rightarrow w_2$, $w_1 \rightarrow w_3$, $w_1 \rightarrow w_4$.
- Step 2: Calculate probability of these ways from w_1. $S_{12}=w_1 \rightarrow w_2$, $S_{13}=w_1 \rightarrow w_3$, $S_{14}=w_1 \rightarrow w_4$. After that, chose the most likelihood probability.
- Step 3: Track the point that has the most likelihood probability. If not the end of word substring, Loop step2: continue to new way from track point to another word in word substring.
- Step 4: Continue with the next word substring
- Step 5: Reduced sentence is the projection of words on the horizontal line.

SENTENCE COMPRESSION BASED ON GRID MODEL ALGORITHM

```
Input
        W: original sentence;
Output
        S: reduced sentence;
Initialization
$i \leftarrow 0; f \leftarrow 0; j \leftarrow 0; S \leftarrow \emptyset; N \leftarrow \emptyset; O \leftarrow \emptyset;$
 1. Separate sentence W to two word sets.
        For $i=1$ to length(W) do
            If $w(i)$ is noun then
                $N \leftarrow w(i);$
            Else
                $O \leftarrow w(i);$
2. Calculate initial probability
        For each $w(i)$ in sentence W
            If $Pr(w(i)|start) > 0$ then
                $f=i;$   // start of reduced sentence.
                $S = S \cup w(f)$
3. Representation w(i) on the Grid
4. Generate sentence compression
    Loop
        While not end sentence W do
            For $i=f$ to length(W) do
              If $w(i)$ is noun then break;
                For $j=f$ to $i$ do
                  $K=argmax(Pr(w(j)|w(f))$
                  $f=j;$
                  $S = S \cup w(j);$
    End Loop
```

Fig. 5. Sentence compression based on Grid model

Like figure 4, reduced sentence contain these words: w_1, w_3, w_4, w_6, w_7, w_8, w_9, w_{10}, w_{11}. Figure 5. is the algorithm of Vietnamese sentence Compression using grid model.

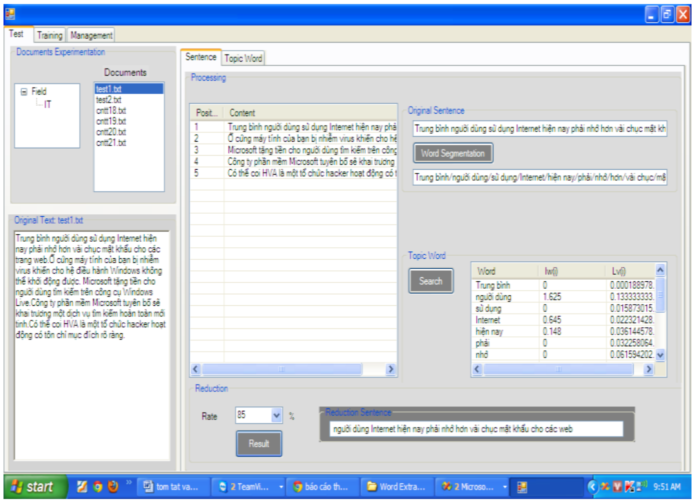
V. EXPERIMENTAL

A. Corpus

Our experiment used the corpus of 100 Vietnamese text. We collected from Vnexpress online news (http://VnExpress.Net). We then used the VLSP word segmentation tool (http://vlsp.vietlp.org:8080/demo/?page=seg_pos_chunk) to segment Vietnamese text into words. After correcting them manually, we obtained more than 200,000 words, which was used in previous works [5].

B. Building text summarization system

This system has been built based on our proposed and use for automatic testing and experimental. Here is the interface of system.



test1.txt
test2.txt
cntt18.txt
cntt19.txt
cntt20.txt
cntt21.txt

Original Text: test1.txt

Trung bình người dùng sủ dụng Internet hiện nay phải nhớ hơn vài chục mật khẩu cho các trang web.Ổ cũng máy tính của bạn bị nhiễm virus khiến cho hệ điều hành Windows không thể khởi động được. Microsoft tặng tiền cho người dùng tìm kiếm trên công cụ Windows Live.Công ty phần mềm Microsoft tuyên bố sẽ khai trương một dịch vụ tìm kiếm hoàn toàn mới tinh.Có thể coi HVA là một tổ chức hacker hoạt động có tôn chỉ mục đích rõ ràng.

Posit...	Content
1	Trung bình người dùng sủ dụng Internet hiện nay phả
2	Ổ cũng máy tính của bạn bị nhiễm virus khiến cho hệ
3	Microsoft tặng tiền cho người dùng tìm kiếm trên công
4	Công ty phần mềm Microsoft tuyên bố sẽ khai trương
5	Có thể coi HVA là một tổ chức hacker hoạt động có t

Original Sentence: Trung bình người dùng sủ dụng Internet hiện nay phải nhớ hơn vài chục mật kh

Trung bình/người dùng/sủ dụng/Internet/hiện nay/phải/nhớ/hơn/vài chục/mậ

Word	Iw(i)	Lv(i)
Trung bình	0	0.000188978.
người dùng	1.625	0.133333333.
sủ dụng	0	0.015873015.
Internet	0.645	0.022321428.
hiện nay	0.148	0.036144578.
phải	0	0.032258064.
nhớ	0	0.061594202.

Rate 85 %

Reduction Sentence: người dùng Internet hiện nay phải nhớ hơn vài chục mật khẩu cho các web

Fig. 6. Text summarization system

It's difficult to compare our method with previous ones, because there were no widely accepted benchmarks for Vietnamese text reduction sentence. Therefore, we compare our proposed method with manual sentence compression generated by humans, called Human, sentence compression method using syntax control, called Syn.con, proposed by M.L. Nguyen [10] and another our method based on determining likelihood substring (Called DLSS) [5].

C. Results

In this experiment, we use the evaluation way as Knight and Marcu [9]. Table I shows the sentence compression results that are carried out by our method, Human and Syn.con for Vietnamese text.

TABLE I. EXPERIMENTAL RESULTS

Method	Compression	Grammatically	Information
Baseline	x	X	X
DLSS	63.26	6.83 ± 1.3	6.78 ± 1.2
Our method	78.1	8.2	8.8
Human	61.2209	8.333333	8.34524
Syn.con	67	6.5 ± 1.7	6 ± 1.1

Table 1 shows compression ratios in the second column, which indicates that the lower the compression ratio the shorter the reduced sentence. The Grammaticality in the third column, which indicates the appropriateness of reduced sentence in term of grammatical.

VI. CONCLUSION

Many Vietnamese text summarization researches were published that showed the importance of Vietnamese information processing problem today. In this paper, the text summarization method based on the sentence compression approach to the target reduce time when applying unsupervised learning method and to not waste cost to build corpus crafts and to reduce computational complexity by using dynamic programming algorithm. The experimental results illustrate our approach is satisfactory requirement of the text summary and can be applied to a number of different languages .

ACKNOWLEDGEMENTS

We would like to thank the experts of University of Engineering and technology, Vietnam of University, and Japan Advanced Institute of Science and Technology , Dr Nguyen Le Minh, Dr Nguyen Van Vinh**,** Dr Nguyen Phuong Thai, Vu Xuan Luong for their great help in building the experimental summarizing application and Vietnamese dictionary corpus.

REFERENCES

[1] Ani Nenkova and Kathleen McKeown, Automatic Summarization, Foundations and Trends in Information Retrieval, Vol. 5, No. 2–3 , p 103–233, 2011.

[2] Clarke, J., & Lapata, M.. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 377–384, 2006.

[3] Clarke, J., & Lapata, M., Global inference for sentence compression: An integer linear programming approach. Journal of Artificial Intelligence Research, 31, 399–429, 2008.

[4] Hongyan Jing and Kathleen R. McKeown. Cut and paste based text summarization, In Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2000), pages 178–185, 2000.

[5] Ha Nguyen Thi Thu, Quynh Nguyen Huu “Method of Sentence reduction in Vietnamese Text Based on Determining Likelihood Substring”. International Conference on intelligen Network and Computing, November, 2010.

[6] Ha Nguyen Thi Thu and An Nguyen Nhat, A Method for Generating Vietnamese Text Sentence Reduction Based on Bayesian Network, International Journal of Innovative Computing, Information and Control, pp 407-416, Vol 11, No2. , 2015.

[7] Hal Daume´ III and Daniel Marcu, A Noisy-Channel Model for Document Compression, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 449-456.

[8] Turner, J., & Charniak, E. Supervised and unsupervised learning for sentence compression. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 290–297, 2005.

[9] Knight, K., & Marcu, D. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. Artificial Intelligence, 139 (1), 91–107, 2002.

[10] M.L. Nguyen and S. Horiguchi, “A Sentence Reduction Using Syntax Control”, Proc. Of 6th Information Retrieval with Asian Language, pp. 139-146, 2003.

[11] M.L. Nguyen and S. Horiguchi, “Example-Based Sentence Reduction Using the Hidden Markov Model” ACM Transactions on Asian Language Information Processing, Vol. 3, No. 2, pp146-158, 2004.

[12] Maria Soledad Pera and Yiu-Khai Ng, A Naïve Bayes Classifier for web document summaries created by using word similarity and significant factors, International Journal on Artificial Intelligence Tools, Vol. 19, No. 4, pp. 465–486, 2010.

[13] Trevor Cohn, Mirella Lapata, Sentence Compression as Tree Transduction, Journal of Artificial Intelligence Research 34, pp. 637-674,2009.

[14] Youngjoong Ko, Jinwoo Park, Jungyun Seo, Improving text categorization using the importance of sentences, Information Processing and Management 40, pp, 65–79, 2004.

[15] Ziegler, C. and M. Skubacz, 2007. Content extraction from news pages using particle swarm optimization on linguistic and structural features. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Nov. 2-5, IEEE Computer Society Washington, DC., USA., pp: 242-249.

[16] Galavotti et al., 2000] Galavotti, L. –Sebastiani, F. –Simi, M.: Feature selection and negative evidence in automated text categorization. Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, ECDL-00, Lisbon, 2000

Virtual Identity Approaches Evaluation for Anonymous Communication in Cloud Environments

Ibrahim A.Gomaa[*§]
Department of Computers & Systems
[*] National Telecommunication Institute
[§] Helwan University, Cairo, Egypt

Emad Abd-Elrahman[* ϕ]
RST Department
[ϕ] Telecom SudParis, Evry, France

Mohamed Abid
IResCoMath Research Unit
National School of Engineers of Gabes, Tunisia

***Abstract*—Since the era of Cloud computing beginning, the Identity Management is considered as a permanent challenge especially for the hybrid IT environments that permit for many users' applications to share the same data center depending on servers' virtualization. This paper introduces a complete study about Identity forms in different domains and applications. Also, a performance evaluation of new approaches for Virtual Identity was done. Virtual Identity, a new terminology used in virtual environments, was introduced to enhance the anonymous communication in such types of complex networks. Based on the work analysis and motivations done through an online survey, two techniques were used to implement the Virtual Identity; Identity Based Encryption (IBE) and Pseudonym Based Encryption (PBE). Both techniques were validated using MIRACL library for security algorithms. In addition, the performance of both approaches was evaluated under different configurations and network conditions through OPNET Modeler. The results showed the impact of the number of cloud users and their locations (either local or remote) on the application response time in cloud environments using the proposed virtual identities. Moreover, Application Characterization Environment whiteboard was used to simulate the overall flow of data across different tiers from start to end of the application task for Virtual Identity creation. The results and outcomes for both methodologies showed that they are suitable paradigm for achieving high degree of security and efficiency in such sophisticated network access to many online services and applications.**

Keywords*—Cloud Environments; Virtual Identity; Performance Evaluation and Security

I. INTRODUCTION

Nowadays, the Internet is inundated by data generated from different sources. Users share personal details, opinions, videos, pictures and very often their identities for each service with public or even their friends. Identity-based service is registered either explicitly, by users who share their identities through social networks, or implicitly by access services through applications of portable devices.

Therefore, the virtualized services access management over the Internet is becoming a critical technology for maintaining privacy and performance especially after the transition to cloud computing. As keeping service provider assets secure is a suitable approach to all parties, anonymous communications between users and virtual service provider became critical issue for users to preserve their personal details.

Users continue to depend on the virtual environments for services' delivery and more individuals are using multiple types of devices to access those services, and applications, Hence, it is necessary hide who has access to the service. Often, users suffer from password bore, having to create and remember at least one password for each service/application. Adding to the challenges of cloud networking security is the increasingly wide range of structured and unstructured data that is exchanged across the network and the heterogeneity of devices used to access it from any place. The service provider must handle access from smart phones, tablets, PCs, and other form factors, often with different operating systems. Each device may equipped with access enterprise applications, mobile apps, social media, streaming video and traditional data each time in one access. Transactions mentioned before, creates a highly sophisticated environment in which the service provider must control how and who has access to what and when.

Table I summarizes and evaluates the four anonymous communication techniques that are available nowadays on the Internet.

This paper highlights the concept of anonymous communication by another way depending on the user identity to create virtual one for services anonymous login. A secure communication is established by creating this virtual identity using a triangle negotiation between the user, the Private Key Generator (PKG) and the service Provider (SP). This process is followed the Identity-Based Cryptography (IBC) methodology either using IBE or PBE techniques. Moreover, both techniques effects in Virtual Identity (V_{ID}) generation are simulated to confirm the solution feasibility for cloud service access. The rest of the paper is organized as follows: in Section 2, related work background and motivations will be presented and reviewed. Section 3 introduces identity in the cloud and virtual environments. Section 4 introduces the proposed V_{ID} mechanisms and its implementation using MIRACL library. Section 5 presents the performance evaluation using OPNET modeler and its results. Finally, Section 6 concludes the paper and future directions for anonymous communication using virtual identities in cloud environments is discussed.

TABLE I. ANONYMOUS COMMUNICATION TECHNIQUES

Parameter	**anonymous communication familiar techniques**			
	Private Browsing	***Proxy***	***VPN***	***Access Routers***
Definition	"Incognito" in Google Chrome "Private Browsing" in Firefox "InPrivate Browsing" in Internet Explorer	prevent the destination from logging IP address and other relevant information of Internet user	Virtual Private Network encrypts all of the packets sent out from users to VPN server.	A random path consisting of multiple nodes are selected, and original data are encrypted and re-encrypted using the public key of the selected nodes.
Pros	Cleans browsing cookies	Hide user's relevant data from the final destination.	all packets that are sent out from the user are encrypted	Use asymmetric cryptography and multiple layers of encryption
Cons	Disclosed by real time attacks as the cleaning is done after the browser is closed.	Disclosed by ISP or an attacker along the route to ISP using traffic analysis	Disclosed by obtaining the secret key or if a VPN server gets hacked	Disclosed by sniffing traffic at the exit node
Anonymity assessment testing	0% Pass	25% Pass	50% Pass	75% Pass
Dynamicity and path changing	Do not support	Do not support	Limited	Support
Cost	Free	Low cost	High-cost	Free
Anonymous Level	Low	Better	Best	High

II. RELATED WORK AND BACKGROUND

Users are very interested in having privacy over all their data, so only their service provider is able to have access to their data. Therefore, Virtual Identity is becoming more and more important for anonymous communication on Internet environment since it can protect people's rights to online privacy even from their service providers.

This section is divided into two parts; the first one presents a review identity categories in social networks, real and cloud environments. In the second part, identity challenges and work motivations is introduced, in addition to, extraction and analysis of an online survey with some questions about using identities in social networking and virtual environments [2].

A. Identity Categories

Two identity categories are discussed:

1) Identity in Social Networks: Undoubtedly, the online social networks (OSNs) became the most visited websites on Internet, with almost one third of all daily online users' transactions visiting them. Non-anonymity communications in social networking platforms disclose privacy. Moreover, users cannot express opinions more freely. In addition, non-anonymity can turn the Internet into a horrible platform because of its built-in nature which making the Sybil attack [3] is piece of cake due to users' identities became easily traceable.

According to the 2015 data breach investigation report [4], threat resulted of social engineering attack is increasing dramatically due to the continually progress in the use of social networking platforms by ordinary and non-technical people, as shown in the Figure 1:

Recently, the most visited online social networks at all, is Facebook that has grown to become one of the most popular social networking platform in the world.

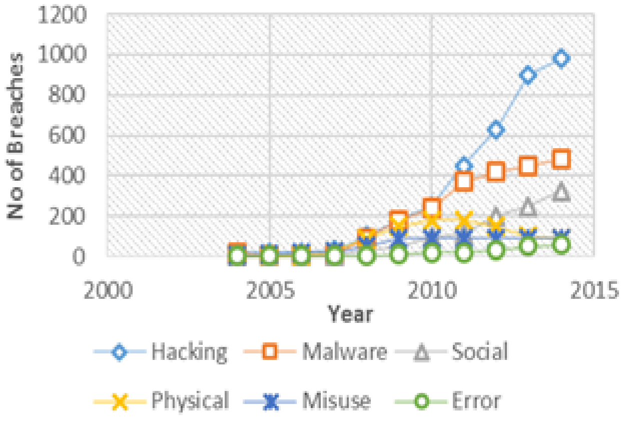


Fig. 1. Number of Breaches per Threat Action Category

It is not only used by many peoples to communicate and share information, but also it turned to be a productive marketing and advertising channel for lots companies, retailers, business entities. In June 2013, there were approximately 1.15 billion monthly active Facebook users [5]. Therefore, Facebook has turned out to become a high-potential target for cyber criminals.

Anonymity on Internet especially on social networks should help users to protect their privacy from getting disclosed. Numerous anonymity techniques are available and used by many technical users on Internet. First of all most of Internet browsers have added anonymous mode to their browsers such as "Incognito" in Google chrome, "private browsing" in Firefox and "InPrivate Browsing" in Internet explorer. Although, the anonymous browsing cleans cookies after the browser is closed, many tools can capture the cookies and use them for real time attack. Another available and widely used anonymity technique is the proxy server, which prevent the server side from logging real IP address of client.

Despite of the client hiding the IP address from target destination, the IP address is still disclosed to the Internet Service Provider (ISP). Therefore by using traffic analysis attacker can get the private information. To solve the proxy server problem, Virtual Private Network (VPN) was introduced to encrypt all of the packets sent from the client. This technique satisfies a higher level of security. There is no chance to decrypt the packets until the VPN server gets hacked. Furthermore, The Onion Router (TOR) is used as the ideal solution for anonymous communication in social networks. This solution has multiple encryption layers and employs asymmetric cryptography. Later, professional attackers pointed out certain vulnerabilities of TOR at exit node; As a solution, TOR uses dynamic IP address to prevent continuous monitoring of exit node [1]

Many works have employed trust in social networks to enhance system anonymity, one of the earlier approaches is [6], which uses personal digital certificates issued by a trusted certificate authority. Then, it applies certain technique such as idemix to make the certificates anonymous, un-linkable, and non-transferable. Consequently, many researches were conducted in this field such as [7-13]. In [14] authors Introduced faceTrust which provide light weighted, flexible and relaxed identity attribute credentials in online social networks. The work in [15] proposed an identity-based Strong Designated Verifier Signature (SDVS) scheme that resists to the key-compromise attack.

2) Identity in Real Time Environment: Identity management using traditional model of username and password became insecure at Internet-scale. The website Experian [16] stated that, in average, the user today shares just five passwords for 26 online identities. Therefore, the problem “break once, break everywhere” arises significantly. Mark Burnett in “10000 Top Passwords” stated that approximately 23% of passwords appear on the table of the three most commonly used passwords as in Table II [17].

Increased security vulnerabilities and growing user frustration have prompted a list of alternatives such as tokens, multi-factor authentication, mutual authentication, biometric identification, and federated identity.

TABLE II. MOST-USED ONLINE PASSWORDS

Frequency	Password
4.7%	password
8.5%	password or 123456
9.8%	password, 123456 or 12345678

Tokens are physical devices generating randomized code that can be used to assure the identity of the user or service which has control of them. Tokens provided by way of either hardware or software, an extremely high level of satisfied authentication because of the multiple exchanges they employ to verify the identity of the user.

Multi-factor authentication approaches intensify security by combining multiple factors: something known (such as a password, PIN) with something we have (such as a token, smart card) with something owned (such as a biometric: retina, hand) and / or something we do (such as voice, handwriting).

Mutual authentication is a model where both the source and the destination entity must fully identify themselves before communication is allowed. It may be accomplished in a number of ways: Diffie-Hellman (DH) key exchange can be used, it provides a more secure method of message exchange and protects the secret being used for an authentication process. This method has a weak point which is the Man In the Middle attack MITM. The solution can be the use or pre-shared key or certificate to avoid this attack. Another method that may be used for mutual authentication is using certificates. The Certificate Authority (CA) must be known by parties to both parties to verify the identities at both sites, and the public keys for both must be shared from the trusted CA.

Biometric identification provides a higher level of authentication than other techniques. It may be used as a main factor of multi-factor authentication, or on a standalone basis. Biometric signatures include fingerprint, iris pattern, facial recognition and heartbeat. Fingerprint-based technology became featured in laptops from manufacturers such as HP, Lenovo and Sony. ING Direct Canada, an online bank, has issued customers with computer mice equipped with fingerprint recognition system. Later, Apple’s added fingerprint feature to the iPhone 5S.

Federated identity approaches enable the existing online accounts to be used to sign into new authenticated-access websites (single sign on). Facebook and Google have moved to leverage their vast user-bases to offer such “federated authentication” services. More details about federated identity are presented in Section 3.

Three case studies of trusted service providers such as financial institutions, governments and mobile operators are discussed hereafter. They are well-placed to offer the importance of identity management services for providing the high-security and high-reliability authentication [18]:

a) Financial Institutions: Financial institutions around the world recognized the risk in online transactions where the parties never meet. Therefore, they must rely only upon electronic identity credentials. IdenTrust is the global leader in trusted identity solutions, applied in more than 170 countries, recognized by global financial institutions, government agencies, and commercial organizations around the world [19]. Also, Mint.com, offers personal financial services and credit monitoring, depending on centrally-authenticated access to all of a user’s online bank and credit card accounts to collect transactions and balance information across the user’s financial issue.

b) Government: Germany, Italy, Spain, Pakistan and Morocco Governments have applied the Electronic Identity Card or “eID” format: a physical identity card with embedded microchip. This approach allows both virtual and physical authentication. The world’s first electronic parliamentary elections were held in Estonia in 2007, powered by the Estonia eID card. The European Union’s “STORK” program (“Secure idenTity across boRders linked”) is working towards a “digital single market by 2015”, allowing recognition of national electronic identity (eID) across the European single market. The Netherlands’ “DigiD” service now provides single-password to access over 500 local and national public service organizations, while the US government is working to allow all federal services to be accessed by passwords from approved third parties, such as Google or PayPal, through the "Federal Cloud Credential Exchange" program.

c) Mobile Operators and Manufacturers: The Subscriber Information Module (SIM) allows Mobile Network Operators (MNOs) to authorize access to services, providing a crypto-graphically-protected unique identifier for each user. Turkish operator "Turkcell" charges 5 Turkish Liras (£1.56, or $2.74) per month for its "Mobil imza" application, which it launched in 2007 to facilitate secure, legally-binding consumer and enterprise transactions. Not only MNOs and MVNOs able to take advantage of mobile identity but also the latest smart phones which contain an embedded secure element, providing SIM-like security, beyond the control of the MNO. This creates opportunities for Mobile Operators and Manufacturers like Apple, Google, Samsung and Microsoft to develop wider Identity Management capabilities.

III. Identity in the Cloud & Virtual Environments

In the Cloud Computing Technology Roadmap, the National Institute of Standards and Technology (NIST) highlighted this concern: "… the need for trusted identities, secure and efficient management of these identities while users' privacy is protected is a key element for the successful adoption of any cloud solution." [20].

The best way to address these concerns is to deploy identity management processes and technologies to ensure that only authorized users have access to cloud applications.

Identity Management process depends on two concepts the first one is Single Sign-On (SSO) and the second one is Federated Identity Management (FIM). SSO makes it possible for a user to log in once and gain access to numerous systems or networks available in a federation without being prompted to log in again [21,22].

Federated identity, describes the technologies, standards and use-cases which serve to enable the portability of identity information across different autonomous security domains. Consequently, users of one domain can access to all the services offered by another domain without burdening them. Hence, with suitable FIM, users should be able to access data across different domains. One important approach in identity management is the Identity Meta-systems, which is defined as an "interoperable architecture for digital identity that enables people to have and employ a collection of digital identities based on multiple underlying technologies, implementations, and providers" [23].

Numerous identity and federation manager products that support federation via Security Assertion Markup Language (SAML) versions 1.1 and 2.0 are available. Actually, there are three major protocols for federated identity: SAML, OpenID and OAuth. SAML [20] [24] is deployed in SSO systems, large enterprises, government agencies and service providers as their standard protocol for communicating identities across the Internet. SAML is an eXtensible Markup Language (XML) based standard for exchanging authentication and authorization Simple Object Access Protocol (SOAP) messages between security domains, that is, between an identity provider and a service provider. In [25] authors introduced an in-depth analysis of 14 major SAML frameworks and showed that 11 of them, including Salesforce, Shibboleth, and IBM XS40, have critical XML Signature Wrapping (XSW) vulnerabilities.

OpenID is used to implement federated identity management in many web sites like Facebook, Microsoft, Google, PayPal, Symantec, and Yahoo. OpenID is an open, decentralized user identification standard, permitting users to log onto different services with the same digital identity. In OpenID the user is authenticated using third-party services called identity providers through simple URL. Users can choose their preferred identity providers to log onto websites that accept the OpenID authentication scheme. OpenID has some vulnerabilities like Phishing Attacks and Authentication Flaws.

OAuth is the third major open standard protocol for federated identity. OAuth is being used exclusively for authorization purposes and not for authentication purposes like OpenID and SAML. OAuth 2.0 relies entirely on the underlying Secure Socket Layer/Transport Layer Security (SSL/TLS) to provide confidentiality and integrity and does not support signature, encryption, channel binding, and client verification. Therefore, it is described as an inherently insecure protocol.

Finally, there is a growing number of other federated identity approaches.

Higgins, is a new open source protocol that allows users to control which identity information is released to an enterprise or with diverse identity management systems.

Windows U-Prove, is Microsoft new identity meta-system controlled by users, that provides interoperability between identity providers and relying parties.

MicroID, is a new identity layer to the web and micro-formats that allow anyone to simply claim verifiable ownership over their own pages and content hosted anywhere.

Liberty Alliance [26], is a large commercially oriented protocol providing inter-enterprise identity trust. It is the largest existing identity trust protocol deployed around the world.

SXIP [26], is commercially available product that offers users the ability to control their own identity information and authentication in use with blogs and other applications.

INames [26], a new service offering a centralized user controlled identity data store as well as providing authentication trust between enterprises.

OpenSSO, is a Sun Microsystems open source version of their commercial product OpenSSO Enterprise. Shibboleth is a distributed web resource access control system that permits federations to communicate together for sharing web-based resources. It is an open source project that uses OpenSAML toolkit.

Lastly, the Ping Identity [27], Next Generation Identity platform facilitates trusted interaction among groups of application providers and consumers on the Internet, through APIs, and from any mobile or desktop screen. Regardless of which product is selected, as long as it conforms to the standards of SAML, all products can be used interchangeably with no compatibility issues.

A. *Identity Challenges*

It is convenient to use a different Virtual Identity V_{ID} for each service. In that way, each V_{ID} is only exposed meanwhile it is used to access to its associated service and it only contains the required attributes for accessing to one service (so less attributes are exposed in a single access to a service).

Furthermore, V_{ID} should be a string that does not include any information about user identity, terminal being used, or service to be accessed. On that way, any sniffer attacker in the access network is only able to know the home domain of the user, but no other information.

It is desirable to maintain a matching between identifiers and services into a private repository, in order to generate the "identity specific side" of V_{ID} in a pseudo-random way. The values stored in this local repository must be maintained equals to the ones in SAML authority side.

B. *Work Motivations*

1) Analysis of Work Motivations: It is clear that using Identity has been changed over the years; nowdays, we find that all services offered over Internet impose using identities. In addition, each required service enforces users to remember an identity for each one. Another one proposes one identity for all services which is not practical. Therefore, using a new kind of identity will overcome the main issues related to having many identities or one single Identity for all services. Also, the traceability of main identities is reduced as the use of virtual ones cannot lead to the original identity.

In order to extract and analyze these work motivations, we did an online survey composed by questions about using identities in social networking and virtual environment [2]. Then, we compared our proposed mechanism with some existing methodologies to evaluate its performances.

a) Survey Analysis: The results of the survey are used to extract some directives to be used in developing/proposing a new solution that enables personalized Identity for services, mapping Identity to user or service’s needs. Through this questionnaire [2], there are a series of questions used to assess user requirements & user satisfactions and to suit their needs from using Identity over social & virtual environments. Moreover, users are asked users to answer some relevant questions about service virtualization knowledge and their identities while accessing social networks like YouTube, Facebook, or Dailymotion. The answer's investigation revealed the existence of two types of user: users from inside the National Telecommunication Institute (Egypt) [28] and users from outside. The participants in this of this survey are mainly scientific users and the institute colleagues’ staff. Therefore, their's culture played an important role in the questionnaire answers and overall analysis. The survey has some direct questions about using new Internet services if the access to them needs some personal qualifications. Actually, the major of this survey was to ask about identity and its privacy. Therefore, the following two sections describe the most important points in the results' analysis.

b) Knowledge and Identity Use: All target users have a good knowledge about social networking access (as the survey indicated 100%). Among them, about 70% prefer using one main identity for all online social sites by means of creating one identity for each service automatically by the operators as shown in Figure 2. By this, the users are searching for an easy solution in order to avoid remembering many identities for all services.

c) Privacy and Virtual Identity: The need for privacy and virtual identity for users is investigated through YES/NO questions. Figure 3 illustrates some questions samples and the answers globally indicate 70% of users are interested in privacy and virtual identity aspects.

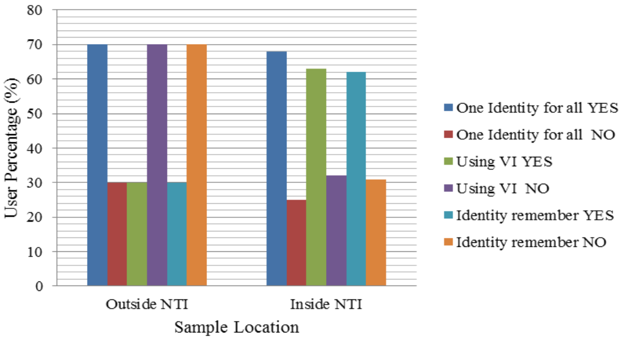


Fig. 2. Identity background questions/answers in the online survey

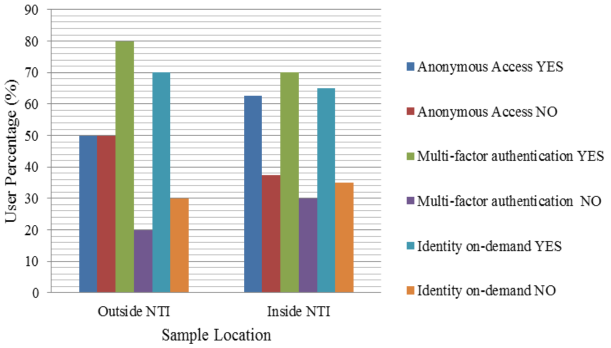


Fig. 3. Identity security and privacy questions/answers in the online survey

2) Survey Conclusion: This survey leads to the major conclusion, which is the importance of using V_{ID} in social environment. The users are motivated to use this principle of access and searched for more privacy by asking for this technique.

IV. PROPOSED VIRTUAL IDENTITY MECHANISMS

In this work, we introduce the IBE and PBE as two approaches for generating virtual identities by collaboration with the Private Key Generator (PKG). The PKG is the security server which is used to generate the IDs deployed in cloud service access based on the type of service required by cloud's user as shown in Figure 4.

Two secure mechanisms for creating V_{ID} are proposed (one based on IBE and the other based on PBE); they are mainly using public-key cryptography for encryption and digital signatures. The key length for most used public key cryptography algorithms has increased over recent years, and this has put a heavier processing load on applications using these algorithms. This burden has ramifications, especially for social and commerce sites that conduct large numbers of secure transactions. Thus, Elliptic Curve Cryptography (ECC) is showing up in standardization efforts, including the IEEE P1363 Standard for Public-Key Cryptography. The principal attraction of ECC, compared to others, is that it appears to offer equal security to RSA for a far smaller key size, thereby reducing processing overhead [29]. Therefore, ECC was chosen in the design of the new proposed solutions. The two solutions need a Private Key Generator (PKG) to calculate the V_{ID}. However, these approaches assume that a centralized Trust Authority (TA) is in charge of the private key generation. Thus, the anonymous communications are not anonymous to the TA. Nevertheless, they use different encryption techniques.

We implemented the two mechanisms IBE and PBE using Multi-precision Integer and Rational Arithmetic C/C++ (MIRACL) library [30] to evaluate the feasibility, performance and scalability of the proposed solutions. Figure 5 and Figure 6 show the two algorithms messages exchanges.

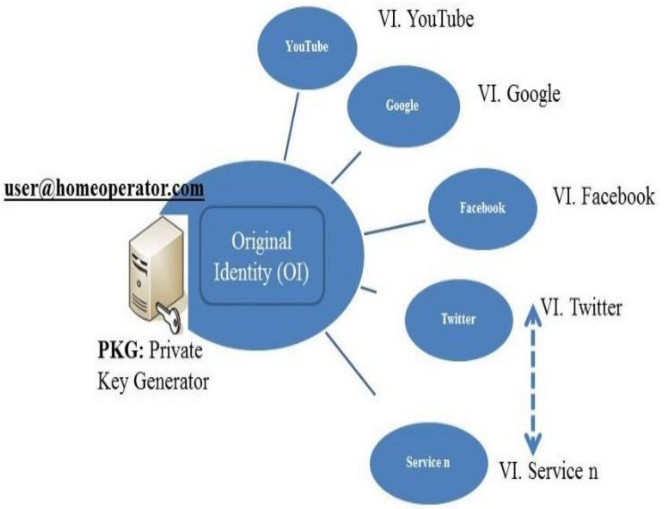


Fig. 4. Virtual Identities generation framework

A. *The First Approach using Identity-basedEncryption (IBE)*

Public-key based solution, such as Identity-Based Cryptographic (IBC) is an asymmetric key cryptographic technique, in which a user's public key can be an identifier of the user and the corresponding private key is created by binding the identifier with a system master secret [31].

The first approach is based on the IBC, which can be traced back to the IBE firstly proposed by Shamir. The construction of the proposed IBE scheme is shown in Figure 5.

Since we use for this solution IBE and ECC, we have to set up the ECC parameters. The equation of the elliptic curve that we used is $y^2 = x^3 + ax + b \bmod p$. The points of this curve define a finite field; their number must be a prime number. In order to satisfy this condition we used the ECPG algorithm. we fixed a prime number (p) and random integer (a). Then we initialize the variable (b) and calculate (n) (the number of points) on the elliptic curve. We used a function in MIRACL that can calculate the number of the points in a finite field. The principle of the algorithm is as in Table III.

The main steps of the proposed solution are:

a) System setup: Each user send U_{ID}: User ID and Ser: Requested Service to Private Key Generator PKG. The Private Key Generator (PKG) or the trust Authority (TA) selects an elliptic Curve E over GF (p) where p is a big prime number. We also denote P as the generator point of E and q (big number), as the order of P. The master Key X = (x1, x2….xn-1, xn). The public Key Y= (y1, y2…yn-1, yn) where yi=xi*P for i=1: n.

b) Key extraction: Given U_{ID}, Ser. PKG generates V_{ID}

The Virtual Identity V_{ID} (V_{ID}=Original identity (mail, service) * Point on elliptic curve).

The User public key UP=H*V_{ID} (H is a secure hash function).

The User private key UD=S*UP (S is the master secret key of PKG).

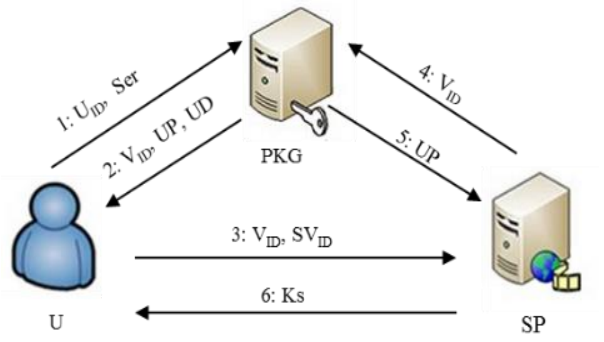


Fig. 5. Proposed IBE Messages Exchanges

TABLE III. ALGORITHM 1 : ECPG ()

Algorithm 1 : ECPG ()
1 : Choose p and a
2 : Initialize b
3 : Calculate n
4 : If n is prime, n will be the proper parameter Else, increase b by 1
5 : Go to 3

c) Signature generation: The announcing user receives V_{ID}, UP and UD from PKG. In order to sign the user virtual identity V_{ID} using a private key UD derived from the PKG to determine V_{ID} and signature SV_{ID}, the announcing user:

- Receives V_{ID}, UP and UD from PKG
- Execute EcdsaSign (V_{ID}, UD) as in Table IV to determine SV_{ID}

TABLE IV. ALGORITHM 2 : ECDSASIGN (V_{ID}, UD)

Algorithm 2 : EcdsaSign (V_{ID}, UD)
1 : Generate n a large prime number
2 : Calculate d= UD mod (n-2)
3 : Computes Q = d* UP
4 : Select a unique and unpredictable integer k in the interval [1, n-1].
5 : Compute k* UP = (x1, y1) and r = x1 mod n. If r = 0, then go to 4.
6 : Compute k-1 mod n.
7 : Compute s = k-1*(h (V_{ID}) + d*r) mod n (h is the Secure Hash Algorithm)
8 : The signature for V_{ID} is the pair of integers (r, s) = Sig (V_{ID}).
9 : Return Sig (V_{ID}) = (r, s)
10 : Publish (Sig (V_{ID}), n, Q)

d) Signature Verification: Once the service provider SP receives the signed virtual identity V_{ID}, it asks PKG for the public key for checking the signed virtual identity SV_{ID}, Algorithm 3 steps are given in Table V.

TABLE V. ALGORITHM 3: ECDSAVER (V_{ID}, UP)

Algorithm 3 : EcdsaVer (V_{ID}, UP)
1: Verify that r and s are integers in the interval [1, n-1].
2: Compute w = s-1 mod n and h (V_{ID}).
3 : Compute h(V_{ID})*w mod n and r*w mod n
4: Compute h(V_{ID})*w mod n*UP + r*w mod n*Q = (x0, y0), v = x0 mod n.
5 : Accept the signature if and only if v=r.

e) Encrypt future communication: If the verification of the signature is successful, the service provider SP generates Shared Secret Key Ks and sends it to user U. Otherwise Ks is discarded. After the generation of the pre-shared key Ks, the future messages are encrypted using pre-shared key Ks as EcdhEncrypt (m), Algorithm 4, Table VI. The resulting ciphertext is denoted by c. The decryption of ciphertext c using the same pre-shared key Ks is given as EcdhDecrypt(c), Algorithm 5, Table VI.

TABLE VI. ALGORITHM 4: ECDHENCRYPT (M); ALGORITHM 5: ECDHDECRYPT (C)

Algorithm 4 : EcdhEncrypt (m)	**Algorithm 5 : EcdhDecrypt (c)**
1 : Gen. random number a ∈ GF (p).	1 : Gen. random number b ∈ GF (p).
2 : Calculate multi_a= a.UP	2 : Calculate multi_b= b.UP
3 : Publish (multi_a)	3 : Publish (multi_b)
4 : Receive multi_b	4 : Receive multi_a
5 : Calculate Ks=a*multi_b	5 : Calculate Ks=b*multi_a
6 : Encrypt m with Ks, {m} Ks	6 : Encrypt m with Ks, {m} Ks
7 : Return c= {m} Ks	7 : Return {m} Ks

B. The Second Approach using Pseudonym Based Encryption

The second approach is based on Pseudonym Based Encryption (PBE), which was proposed for Key management for anonymous communication in mobile ad-hoc networks [32]. In this approach, user uses PBE to calculate its own V_{ID}. The PKG just computes the user's private key, which depends on its secret master key. The PKG will act as an authority that certifies that the user has the private key corresponding to his/her public key. Figure 6 shows the second solution based on PBE messages exchanges.

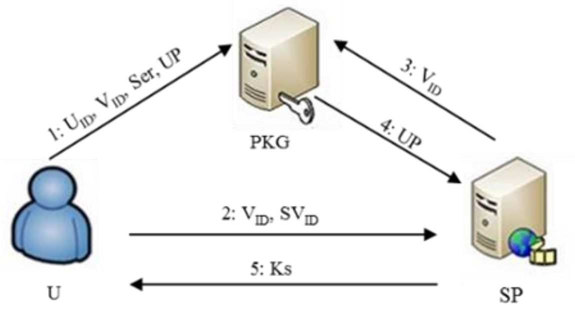


Fig. 6. Proposed PBE Messages Exchanges

The user sends to the PKG his/her identity (e.g., user@homeoperator.com), the requested service, the public key by choosing an elliptic curve with its generator point P and chooses his/her V_{ID} (as pseudonym). The PKG calculates the user's private key UD and doesn't need to send the key pair (public/private) to the user because the UP and UD are already computed by the user. The user wants to be authenticated by SP; therefore, he/she uses an Identity-Based Signature (IBS) [33] to calculate SV_{ID} and sends it with V_{ID} to the SP. The SP sends V_{ID} to the PKG and asks for public key corresponding to the V_{ID}. The SP verifies the SV_{ID} by decrypting it using the UP. If it retrieves the V_{ID}, then the authentication succeeds. At the end, the SP generates and sends a shared secret key to the user to encrypt future communication between them.

We implement the second solution using the same steps as done in the first one except for the second step in IBE (as the trusting is verified by the cloud service provider in this case).

- Each user sends U_{ID}: User ID and Ser: Requested Service V_{ID}: Virtual ID (Pseudonym), UP: User Public Key to Private Key Generator PKG. The PKG is in charge of the private key generation within an anonymous communication system. Therefore, the anonymous communications are not anonymous to the trust authority (TA).
- The PKG/TA just computes the user's private key, which depends on its secret master key. PKG selects an elliptic Curve E over GF (p) where p is a big prime number. The PKG calculates the user's private key UD and doesn't send the key pair (public/private) to the user because the UP and UD are already computed by the user.
- Other steps follow the same way as described before in the first solution.

C. Processing time for IBE and PBE resulting from MIRACL

We used MIRACL Library during the evaluation of our solution's performance to observe the processing time for all functions and executed entities. The results for the two proposed solutions IBE and PBE are illustrated in the following two tables (Table VII and Table VIII).

TABLE VII. PROCESSING TIMES FOR IBE

Message ID	Source	Destination	Depends On	Processing Time (sec)
1	U	PKG	Beginning	N/A
2	PKG	U	ID:1	0.034
3	U	SP	ID:2	0.004
4	SP	PKG	ID:3	0.0015
5	PKG	SP	ID:4	0.0015
6	SP	U	ID:5	0.009
Six messages total				0.05

TABLE VIII. PROCESSING TIMES FOR PBE

Message ID	Source	Destination	Depends On	Processing Time (sec)
1	U	PKG	Beginning	N/A
2	U	SP	ID:1	0.033
3	SP	PKG	ID:2	0.0015
4	PKG	SP	ID:3	0.028
5	SP	U	ID:4	0.009
Five messages total				0.0715

The results we got to calculate the processing times for all messages are around 0.05 Sec and 0.0715 Sec for all executed entities and functions for IBE and PBE respectively, using a computer machine has specs, Intel Core 2 Duo CPU E8400 @ 3.00GHz x 2, memory 4G in Linux Ubuntu 12.10. Table VII and VIII show the processing times for IBE and PBE as captured during the two scenarios validation.

V. PERFORMANCE EVALUATION USING OPNET MODELER

Optimized Network Performance (OPNET) Modeler [34] is a discrete event simulation tool. It provides a comprehensive development environment supporting the modeling and simulation of communication networks. This contains data collection and data analysis utilities. OPNET allows large numbers of closely spaced events in a sizeable network to be represented accurately. This tool provides a modeling approach where networks are built of nodes interconnected by links. Each node's behavior is characterized by the constituent components. The components are modeled as a final state machine. Actually, we used Application Characterization Environment (ACE), which is included in OPNET Modeler to visualizes, analyzes, and troubleshoots networked applications.

A. Network Model Scenarios

ACE has a number of predictive features that enable us to determine how network and application changes will affect application performance. Therefore, we used it to evaluate the proposed solutions. First, we use ACE whiteboard to draw exchanging messages among the three entities User (U), Private Key Generator (PKG) and Service Provider (SP). Therefore, we set the processing time for each message as obtained from MIRACL validations. After that, from the ACE whiteboard, we draw the two scenarios of network topology for each proposed solution. The first scenario is for local clients and the second scenario is for remote clients. For each proposed solution we evaluated the performance in two cases, the first one is when clients (users) need single service and the second one is when users need more than one service access (exactly ten services). Finally, we compared the results obtained and conclude this performance evaluation.

B. Simulation Results

As shown in the following sections, we implemented four different scenarios. In all scenarios, we measured the application response time, which is described as the time taken for all the tasks in the custom application to complete.

1) First solution using IBE with Single Service: As shown in Figure 7, the application response time is not zero for all users. Nevertheless, it has small values such that the application response time resulted when one user used one service is 0.052857 seconds and when 200 users used one service is 0.406446 sec. We noticed the differences when remote users use one service.

2) First solution using IBE Multiple Services: Application response time increased when users used many services. Actually, we simulated 10 services for each user. Figure 8 highlights the two use cases of accessing either locally or remotely for multiple services (10 services) IBE virtual identity-basedgeneration.

3) Second solution using PBE Single service: As shown in Figure 9, the values of application response time that we got close to the values mentioned earlier in the case of IBE single service. As mentioned before, these values are not zero but the application response time resulted when one user used one service is 0.075512 seconds and when 200 users used one service is 0.406088 sec. We noticed the differences when remote users use one service.

4) Second Solution using PBE Multiple Services: As mentioned before, application response time increased when cloud users requested many services. We note that, PBE application response time is better than IBE application response time when using users multiple services scenario as it is clear in Figure 10.

5) Global Results: We note the difference when remote users used many services. In fact the scenario of remote users multiple services is the actual one. Most cloud users used many services remotely. Therefore, we can emulate the number of users and application response time resulted from this scenario to calculate the overall delay for actual cloud networks used IBE to create V_{ID} for anonymous communication.

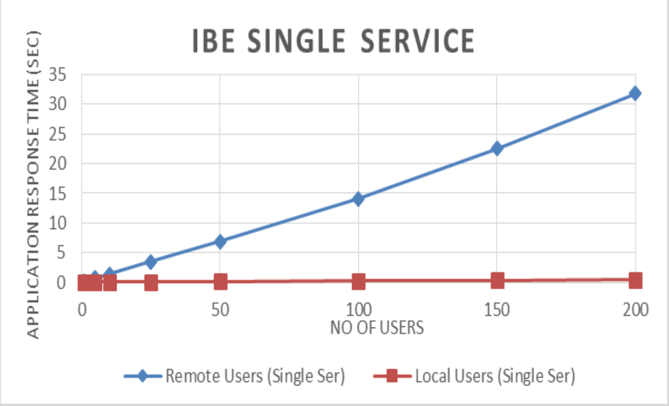


Fig. 7. IBE Single Service

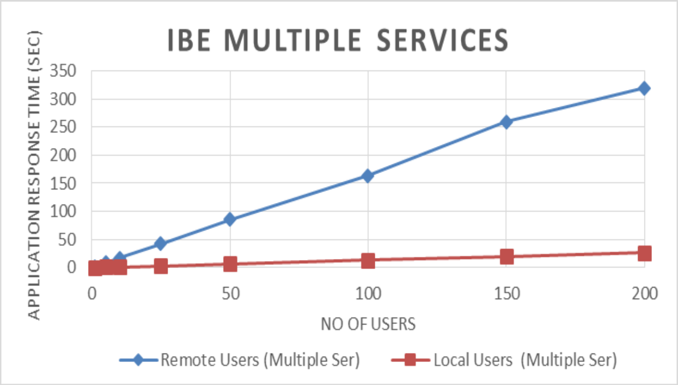


Fig. 8. IBE Multiple Services

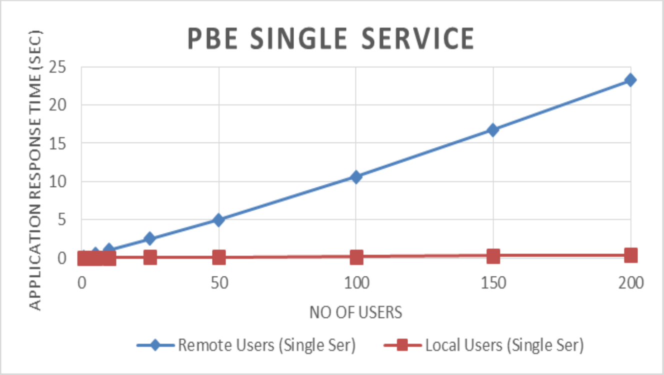


Fig. 9. PBE Single Service

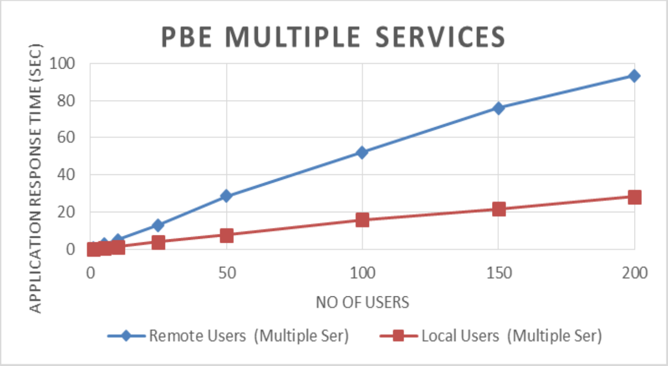


Fig. 10. PBE Multiple Services

VI. Conclusions

In cloud computing environment, users want to protect their privacy and their identities. In the literature, different manners to use identities in the network application are found. In this paper, two novel approaches to generate virtual identities are proposed. The first one is based on the Identity Based Encryption (IBE) and the second one, on Pseudo Based Encryption (PBE). We started by complete study about the identity management in general and the virtual identity in particular. Then, and in order to validate our work motivations, an online survey about issues and opportunities in virtual environment is analyzed to draw our framework architecture for supporting VID solutions. The proposed solution defined a single sing on and anonymous communication to help Cloud and Internet users protecting their privacy and private information from any disclosure. Both approaches implementations are validated using MIRACL library. Furthermore, another performance evaluation is done using OPNET Modeler. The evaluation drew our attention through the proposed solutions feasibility in cloud scale applications or services based on simulating single and multiple services for both local and remote users access. As future directions for this work, we will validate our solutions in a real cloud platform like OpenStack in correlation with keystone security server for best integration with cloud scalability.

References

[1] Hoang N., and Pishva D., “Anonymous Communication and its Importance in social networking”, ICACT2014, February 16-19, 2014.

[2] Survey about using Identities in Social Networks and Virtual Environments, http://www.ntiegypt.sci.eg/survey/index.php/212212/, last visit: December, 2015.

[3] Danezis G. and Mittal P., “Sybil Infer: Detecting Sybil Nodes using Social Networks”, NDSS, 2009.

[4] Threat Actions, the 2014 data breach investigations report, Verizon enterprise, page 9, http://www.verizonenterprise.com/DBIR/2015/, last visit: December, 2015

[5] Facebook reports second quarter 2013 results, Facebook, Retrieved 24 July 2013.

[6] Camenisch J. and Herreweghen E., “Design and Implementation of the idemix Anonymous Credential System”, ACM CCS, 2002.

[7] Pujol J. and Delgado R., “Extracting Reputation in Multi Agent Systems by Means of Social Network Topology”, AAMAS, 2002.

[8] Sovran Y., Libonati A., and Li J., “Pass it on: Social Networks Stymie Censors”, IPTPS, 2008.

[9] Ramachandran A., and Feamster N., “Authenticated Out-of-Band Communication over Social Links”, WOSN, 2008.

[10] Yardi S., Feamster N., and Bruckman A., “Photo-Based Authentication Using Social Networks”, WOSN, 2008.

[11] Tran D., Min B., Li J., and Subramanian L., “Sybil-Resilient Online Content Rating”, NSDI, 2009.

[12] Lesniewski-Laas C. and Whanau M., “A Sybil-proof Distributed Hash Table”, NSDI, 2010.

[13] Post A., Shah V., and Bazaar A., “Strengthening user reputations in online marketplaces”, NSDI, 2011.

[14] Sirivianos M., Kim K., Gan J. and Yang X., “Assessing the veracity of identity assertions via OSNs”, IEEE, 2012.

[15] Lin H., “Toward Secure Strong Designated Verifier Signature Scheme from Identity-Based System”, IAJIT, Vol.11 No.4, July 2014.

[16] Experian, http://press.experian.com/, last visit: December, 2015.

[17] Burnett M., “10,000 Top Passwords”, https://web.archive.org/web/20150315000117/ https://xato.net/#.Vm2Zs0p97IV, last visit: December, 2015.

[18] Dargue M. and Wadsworth W., “Cartesian: Identity in the Internet Age” the management network group, September 2013.

[19] IdenTrust: Bank Assurance for Government, White Paper, IdenTrust Inc., USA.

[20] National Institute of Standards and Technology (NIST), Computer Security Division, Information Technology Laboratory, July 2013.

[21] Galpin R. and Flowerday S., “Online Social networks: Enhancing user trust through effective controls and identity management”, IEEE, 2011.

[22] Hamlen K., Liu P., Kantarcioglu M., Thuraisingham B. and Yu T., “Identity Managememt for Cloud Computing: Developments and Directions”, CSIIRW ’11, October 12 -14, 2011.

[23] Lewis K. and Lewis J., “Web Single Sign-On Authentication using SAML”, IJCSI, Vol. 2, pp.41-48, 2009.

[24] Prasanalakshmi B. and Kannammal A., “Secure Credential Federation for Hybrid Cloud Environment with SAML Enabled Multifactor Authentication using Biometrics”, International Journal of Computer Applications, Volume 53, No.18, September 2012.

[25] Somorovsky J., Mayer A., Jorg S., Schwenk j., Kampmann M., and Jensen M., “On Breaking SAML: Be Whoever You Want to Be”, 21st USENIX Security Symposium, August 8-10, 2012.

[26] Authentication world, http://www.authenticationworld.com/Authentication-Federation, last visit: December, 2015.

[27] Ping Identity, https://www.pingidentity.com/en/products/next-gen-identity.html, last visit: December, 2015

[28] National Telecommunication Institute, http://www.nti.sci.eg/, last visit: December, 2015

[29] Stallings W., Cryptography and Network Security: Principles and Practice, 5/E, Prentice Hall, 2011.

[30] Multi-precision Integer and Rational Arithmetic C/C++ (MIRACL) library, http://info.certivox.com/, last visit: December, 2015.

[31] Chen L., “An Interpretation of Identity-Based Cryptography”, Foundations of Security Analysis and Design IV, Lecture Notes in Computer Science, Volume 4677, 2007.

[32] Huang D., “Pseudonym-based cryptography for anonymous communications in mobile ad hoc networks”, Int. J. Security and Networks, Vol. 2, 2007.

[33] Boneh D. and Franklin M., “Identity-Based Encryption from the Weil Pairing”, CRYPTO 2001, LNCS 2139, Springer-Verlag, 2001.

[34] Riverbed, http://www.riverbed.com/, last visit: December, 2015.

A Multi-Criteria Decision Method in the DBSCAN Algorithm for Better Clustering

Abdellah IDRISSI
Computer Sciences Laboratory (LRI), Computer Sciences Department
Faculty of Sciences, Mohammed V University of Rabat
Rabat, Morocco

Altaf ALAOUI
Computer Sciences Laboratory (LRI), Computer Sciences Department
Faculty of Sciences, Mohammed V University of Rabat
Rabat, Morocco

***Abstract*—This paper presents a solution based on the unsupervised classification for the multiple-criteria analysis problems of data, where the characteristics and the number of clusters are not predefined, and the objects of data sets are described by several criteria, and the latter can be contradictory, of different nature and varied weights. This work focuses on two different tracks of research, the unsupervised classification which is one of data mining techniques as well as the multi-criteria clustering which is part of the field of Multiple-criteria decision-making. Experimental results on different data sets are presented in order to show that clusters, formed using the improvement of the algorithm DBSCAN by incorporating a model of similarity, are intensive and accurate.**

Keywords—Data mining; Clustering; Density-based clustering; Multiple-criteria decision-making

I. INTRODUCTION

Many studies showed that the resort to multiple-criteria analysis of the data in the classification establishes an effective approach for the extraction of the information, and that in optimal way in big databases described by several criteria, which are sometimes of different nature [1], [2]. To do it, several algorithms of different principles have been used in various different types of work. For example, UTADIS [3], [4][5] which presents the first and the only method belonging to the unique criterion synthesis approach. Basing on the utility functions apply only in the case cardinal data. In the first methods of assignment based on outranking relations approach, there is Trichotomic segmentation [6] and N-tomic (A Support System for Multicriteria Segmentation Problems) [7], had a limited number of categories and a fuzzy assignment. On the other side Electre-Tri [8] [9] [10] with its rather strong explanatory character, can handle any number of categories. There have been many developments since then [12]. But always with fuzzy assignment, an ordinal sorting and preorder structure. Thus the filtering method based on fuzzy preference introduced the fuzzy assignment approach and a binary relation of preference. The last techniques based on fuzzy indifference modeling, PROAFTN [13] [14], [15] and TRINOMFC [16] are the methods of nominal sorting which require no particular structure.

However, it is noticed that all these methods have for basic principle supervised learning. This tendency is confirmed by the studies of D' Henriet [16], Zopounidis [2], Belacel [17] and others who list the various algorithms of multiple-criteria classification, and those classified in the family of multiple-criteria assignment based on supervised learning.

In spite of the superiority of the algorithms based on the supervised classification, their contribution remains limited in face to certain problems in which the information or/and the experience in the domain remain insufficient to predefine the clusters. To overcome this problem, some studies have begun researches by exploiting unsupervised learning.

In this sense, F.Anuska [1] introduces the research by evoking the multiple-criteria clustering problem and proposes the attempts of solution based on:

- The reduction of the multiple-criteria analysis problem in clustering to clustering problem with single criterion obtained as a combination of the criteria;
- The application of the techniques of clustering to grouping obtained by using single criteria clustering algorithms for each criteria;
- The application of constrained clustering algorithms where a chosen criterion is considered as the clustering criterion and all others are determining for the constraints;
- The modification of a hierarchical algorithm which would allow to solving the problem directly.

However, the indirect solutions proposed by F. Anuska direct towards NP-complete problems. And even direct solutions based on a hierarchical clustering method would be limited, because all the hierarchical clustering algorithms are efficient when the size of dataset does not exceed 100 objects [18], and they also are adapted for specific problems associated with areas having the separation or the regrouping of the objects, following the example of taxonomy in biology and in the natural evolution of the species [19].

Then Y. De Smet [21] and Rocha [20] used partition-based clustering algorithms as K-means. The first proceeded to the improvement of the K-means algorithm [22] by integrating a structural procedure preference (P, I, J) considering a triplet of binary relations, where p models strong preference, I Indifference relation and J incomparability relation. The second, more recent proposed the classification approach of a set of alternatives to a set of partially ordered categories by using the K-means method. Thereafter, these categories are classified according to their centroid by using an ordinal classification process such as ELECTRE [23]. In spite of the notoriety of K-means with a large number of variables, may be computationally faster than other clustering (if K is small), however the partitioning methods in clustering require fixed number of clusters can make it difficult to predict it. Moreover, this method is based on calculation of the distance, which obliges to establish the metric ones [24].

Taking into consideration all the limits evoked previously, this present paper proposes an approach of an unsupervised clustering algorithm based on the density. This algorithm is contributing to the resolution of the problem of clustering in a multidimensional way by using algorithm DBSCAN [25] and integrating a model of similarity inspired of the concept of the multiple-criteria decision analysis [26], [27] [28] [29] [30] [31]. This approach based on the density makes it possible to work on great databases without however determining beforehand the nature and the number of clusters, in this family of clustering much of work exists, quoting by way an example algorithm DGLC [32], OPTICS [33], DENCLUE [34], WaveCluster [35], CLICKS [36], CURD [37] AND DBSCAN [38]. And the choice of DBSCAN algorithm is justified by the fact that beyond supporting several types of data of which those of space, it is particularly effective when the groups are touched or in the presence of noise. It is also effective for the detection of non-convex clusters [38] [39]. It is also advisable to stress that the fact of working with the no modified version of DBSCAN algorithm, which leaves the result of this exploitable work by all other improved DBSCAN algorithms, following the example of OPTICS [33], DVBSCAN [40], VDBSCAN [41], DBCLASD [42], LDD-DBSCAN [43], NDCMD [44], ST-DBSCAN [45].

II. Approach Proposed: Involving the Multi-Criteria Concept in the DBSCAN Algorithm

Both data mining research and Multiple-criteria decision-making have each specific and limited asset. As a result, the hybrid algorithm (DBSCAN modified) synergies the strengths of each algorithm in solving clustering problems.

A. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN [25], A Density Based Spatial Clustering of Application with Noise, is a density based clustering technique for discovering clusters of arbitrary shape as well as distinguishing noise. DBSCAN accepts a radius value $Eps:(\varepsilon)$ based on a user defined distance measure and a value $MinPts$ for the number of minimal points that should occur within Eps radius.

The following are some concepts and terms that explain the DBSCAN algorithm as presented in [25]:

- Eps-neighborhood: The Eps-neighborhood of a point p " $N_{\varepsilon}(p)$ " is defined by:
- $N_{\varepsilon}(p)=\{\forall q\in D|\ p\neq q\wedge dist(p,q)\leq Eps\ \}$, with D as database of n objects.
- Core object: A core object contains at least a minimum number $MinPts$ of other objects within its Eps-neighborhood.
- Directly density-reachable: A point p is directly density-reachable from a point q if $p\in N_{\varepsilon}(q)$ and q is a core point.
- Density-reachable: A point p is density-reachable from the point q with respect to Eps and $MinPts$ if there is a chain of points p_1,..., p_n, with p_1 = q and p_n = q such that p_{i+1} is directly density reachable from p_i with respect to Eps and $MinPts$, for $1 \leq i \leq n$, $p_i \in D$.
- Density-connected: A point p is density connected to a point q with respect to Eps and $MinPts$ if there is a point $o\in D$ such that both p and q are density-reachable from o with respect to Eps and $MinPts$.
- Nose: A point in D is a noise object if it does not belong to any cluster.
- Cluster: A cluster C with respect to Eps and $MinPts$ is a non-empty subset of D satisfying the following conditions:
- $\forall p,q$: if $p\in C$ and q density-reachable from p with Eps and $MinPts$, then $q\in C$ (Maximality);
- $\forall p,q$: p is density-connected to q with Eps and $MinPts$. (Connectivity).

B. The model of similarity and dissimilarity

The model of comparison used in our algorithm is composed of four stages by calculating the following functions (e.g. first object: alt1 and second object: alt2) [30] [38]:

- The functions of similarity: $Similarity(alt1,alt2)$ (1);
- The functions of the weighted similarity: $WeigthedSi\ milarity\ (alt1,alt2)$ (2);
- The function of strong dissimilarity: $StrongDissimilarity_{i}\ (alt1,alt2)$ (3);
- The functions of the overall similarity: $OverallSimilarity\ (alt1,alt2)$ (4).

1) The function of similarity

In order to calculate the similarity (1) between two alternatives for each criterion "i" of the whole of criteria, we use the following functions:

$Similarity_{i}(alt1,alt2):D\times D\rightarrow\{-1,1\}$; Such D is the group of the objects (alternatives).

$$Similarity_i(alt1, alt2) = \begin{cases} +1 \text{if } |\text{alt1}_i - alt2_i| \leq \sigma_i \\ -1 \text{if } |\text{alt1}_i - alt2_i| \succ \sigma_i \end{cases} \quad (1)$$

Each criterion is determined by a threshold σ_i , denotes marginal similarity of the criterion "i" with $0 \leq \sigma_i \leq MaxCr_i - MinCr_i$, where $MaxCr_i$ and $MinCr_i$ are respectively the maximal value and the minimal value of the criterion "i".

According to the results of the first function, we can conclude that the similarity of two alternatives "alt1" and "alt2" come as follows:

- If $Similarity_i(alt1, alt2) = +1$, then "alt1 "and "alt2" are similar on criterion "i";
- If $Similarity_i(alt1, alt2) = -1$, then "alt1" and "alt2" are not similar on criterion "i".

2) The function of the weighted similarity

In this stage, the importance of every criterion is introduced, the function of the weighted similarity (2) is the sum of product of similarity $Similarity_i(alt1, alt2)$ (1) and the weight "pi" of every criterion "i".

$$WeigthedSimilarity\,(alt1, alt2) : D \times D \rightarrow [-1,1];$$

Such D is the group of the objects (alternatives).

$$WeigthedSimilarity(alt1, alt2) = \sum_{i=1}^{n} p_i * Similarity_i(alt1, alt2) \quad (2)$$

The results of this function can be classified in three cases:

- If $0 < WeigthedSimilarity(alt1, alt2) \leq 1$, it implicates that it is more sure than not that "alt1" is similar to "alt2";
- If $-1 \leq WeigthedSimilarity(alt1, alt2) < 0$, it implicates that it is more sure that "alt1" is not similar to "alt2" than the opposite;
- If $WeigthedSimilarity(alt1, alt2) = 0$, in this case we are in doubt whether object "alt1" is similar to object "alt2" or not.

To reinforce results and to limit doubt, by passing to the third stage, this latter can calculate strong dissimilarity between two alternatives.

3) The function of strong dissimilarities

This stage of the model allows to calculating strong dissimilarity (3) between two alternatives by using the following function:

$$StrongDissimilarity_i\,(alt1, alt2) : D \times D \rightarrow \{0,1\}.$$

Such D is the group of the objects.

.

$$StrongDissimilarity_i\,(alt1, alt2) = \begin{cases} 1\,if\ |\text{alt1}_i - alt2_i| \geq \delta_i^+ \\ 0\ elseif \end{cases} \quad (3)$$

Where δ_i^+ is the threshold of strong dissimilarity, such as: $\delta_i < \delta_i^+ \leq MaxCr_i - MinCr_i$.

If $StrongDissimilarity_i\,(alt1, alt2) = 1$ implicates that "alt1" and "alt2" are strongly dissimilar on criterion "i".

In certain cases two alternatives can be similar in most criteria but there is a strong dissimilarity on the other criteria.

4) The functions of overall similarities

The last stage of the model of comparison allows us to introduce a total similarity (4). With the aid of following functions, we can finalize this model of comparison. $OverallSimilarity\,(alt1, alt2) : D \times D \rightarrow [-1,1];$

$$\begin{aligned} &OverallSimilarity\,(alt1, alt2) = \\ &mm(WeigthedSimilarity(alt1, alt2), \\ &-StrongDissimilarity\ _1(alt1, alt2), \ldots, \\ &-StrongDissimilarity_m\,(alt1, alt2)) \end{aligned} \quad (4)$$

With this function: $mm : [-1,1]^q \rightarrow [-1,1]$.

$$mm(p_1, \ldots, p_q) = \begin{cases} \max(\,p_1, \ldots, p_q)\ if\ \ p_i \geq 0 \\ \min(\,p_1, \ldots, p_q)\ if\ \ p_i \leq 0 \\ 0\ elseif \end{cases} \quad (5)$$

- If $WeigthedSimilarity(alt1, alt2) > 0$ and there is no strong dissimilarity between both alternatives "alt1" and "alt2", it implicates that $OverallSimilarity\,(alt1, alt2) = WeigthedSimilarity(alt1, alt2)$ In that case we can conclude that both alternatives "alt1" and "alt2" are similar;
- If $WeigthedSimilarity(alt1, alt2) > 0$ and there is a strong dissimilarity between both alternatives "alt1" and "alt2" with one or several criteria, it implicates that $OverallSimilarity\,(alt1, alt2) = 0$. In this case, we must prove the number of criteria where there is a strong dissimilarity and a weight of these criteria;
- If $WeigthedSimilarity(alt1, alt2) \leq 0$ and there is a strong dissimilarity between both alternatives "alt1 "and "alt2" on one or several criteria, it implicates that $OverallSimilarity\,(alt1, alt2) = -1$. Therefore both alternatives are dissimilar.

C. Description of the algorithm

D denote a set of n objects, where each object of this list is described on m criteria of nominal, interval, ordinal and/or cardinal type. The evaluation of an object on criteria j can be encoded in real interval bounded by the minimal and maximal value of this criteria "i": $\left[MaxCr_i, MinCr_i\right]$.

The relative importance of which criterion intervenes in assessing the comparison between two objects is not always equivalent and can influence the final result of a multi-criterion analysis. Therefore, the presence of a coefficient related to every criterion; witch reflects the importance in comparison with other criteria; is a primordial aspect in an algorithm to appoint a weight to every criterion with: $p_i \in [0,1]$ and $\sum_{i=1}^{n} p_i = 1$.

The algorithmic approach can be structured into the following steps:

1) Choose an arbitrary object "$alt_i \in D$" of the set of alternatives;

2) Calculate similarity (1) and strong dissimilarity (3) of this object "alt" with every object of the set of alternatives;

3) Calculate the weighted similarity (2) of this object "alt_i";

4) Calculate the overall similar (4) this object "alt_i";

5) Test the value of overall similar (4) and the presence of strong dissimilarity (3) which allows the determination if the alternative is considered to be a neighborhood of the object "alt_i";

6) Recover all objects density-connected to the object "alt_i" on the parameters of overall similar (4) and the parameter "$MinPts$";

- If "alt_i" is a core object, a cluster is formed;
- If "alt_i" is a point of border, therefore any points can be density-connected to "alt_i" and the algorithm visits the following object of the set of alternatives;

7) This sequence continues until the density-connected cluster is completely and definitively found.

D. Multi-Criteria-DBSCAN Algorithm

```
Algorithm MC-DBSCAN(D, $\sigma$ , $\delta^+$,  $p$ , $MinPts$ )
//Inputs:
//D={a1,a2,..., an} set of alternatives(objects)
// $\sigma$ : threshold denote marginal similarity discrimination threshold of the criterion
// $\delta^+$ : is the threshold of strong dissimilarity
// $p$ : set of weights of every criterion
// $MinPts$ : the number of minimal points that should occur within Eps radius

//Output:
//C={c1, c2,..., ck} set of clusters
 Cluster_Label=0
 for i=1 to |D|
 if $a_i$ is not cluster then
 for j=1 to |D|
 $L_1 = Similarity(a_i, a_j)$   (1)
 $L_2 = StrongDissimilarity(a_i, a_j)$   (3)
 $L_3 = WeightedSimilarity(L_1)$   (2)
 $X = X \cup OverallSimilarity(L_2, L_3)$   (4)
 end for
 if |X| < $MinPts$ then
 marke $a_i$ as noise
 else
  Cluster_Label= Cluster_Label+1
  add $a_i$ to cluster
  for i=1 to |X|
   if $a_i'$ is not cluster then
   for j=1 to |D|
 $L_1' = Similarity(a_i', a_j')$   (1)
 $L_2' = StrongDissimilarity(a_i', a_j')$   (3)
 $L_3' = WeightedSimilarity(L_1')$   (2)
 $X' = X' \cup OverallSimilarity(L_3', L_2')$   (4)
   end for
   if |X'| >= $MinPts$ then
   X = X U X'
   if $a_i'$ is not cluster then
              add $a_i'$ to cluster
 end for
 end for
```

Algorithm 1: MC-DBSCAN Algorithm

III. Experimentation and Results

To test and to assess the performances of our algorithm, we implemented the DBSCAN and the MC-DBSCAN algorithms by using Java as a language to implement the algorithms.

Performances of both algorithms DBSCAN and MC-DBSCAN are assessed on a few well-known datasets such as the Stulong [48], Iris [46], BasketBall [48], ColorHistogram

[48] and other ones from UCI Machine Learning Repository [47] and KEEL Knowledge Extraction based on Evolutionary Learning [48].

For these tests to reflect correctly the performance of an algorithm, we compare the number of groups created by both algorithms and the percentage of non classified objects by varying parameters knowing that common parameters, ray locating maximum neighbors "$Eps:(\varepsilon)$" and the minimum number of points that have to be present in Eps- neighborhood of this object "$MinPts$", we have the same values as both algorithms.

In the results table "Tab. 1" due to the global parameter Eps and MinPts, DBSCAN classifies objects in one class because it is not able to consider several criteria simultaneously.

The results presented in "Fig. 1" and "Fig. 2" prove that the classes obtained by the multi criteria clustering algorithm are very similar to groups that have been proposed by experts and that the percentage of non-classified objects is too low.

TABLE I. COMPARISON OF RESULTS BETWEEN TWO ALGORITHMS: DBSCAN AND MC-DBSCAN

Data	**Number of alternatives**	**Number of criteria**	**Parameters (ε, δ_i^+, MinPts)**	**DBSCAN**		**MC-DBSCAN**	
				Number of clusters	***Non classified objects***	***Number of cluster***	***Non classified objects***
Iris	150	4	0.2, 2.2 et 6	2	4.5%	3	0%
			0.4, 0.9 et 6	1	0%	7	15%
			0.6, 2.2 et 6	1	0%	3	4.5%
			0.9, 2.2 et 6	1	0%	2	3%
Stulong	1417	5	0.9, 80, 9	1	0%	7	0.28%
			0.6, 80, 9	1	0.49%	7	0.28%
			0.1, 80, 6	1	8.04%	7	0.28%
Color Histogram	65535	32	0.2, 0.9, 3	77	98.52%	8	0.0120%
			0.01, 0.9, 4	12	99.53%	9	0.01%
			0.25, 0.9, 9	2	0.19%	9	0.016%
			0.6, 0.9, 5	1	6.25%	9	4.19%
BasketBall	96	5	0.6, 0.9, 6	1	0%	7	4.19%
			1.6, 9, 8	1	0%	7	5.20%
			0.4, 0.9, 4	1	6.25%	9	4.19%

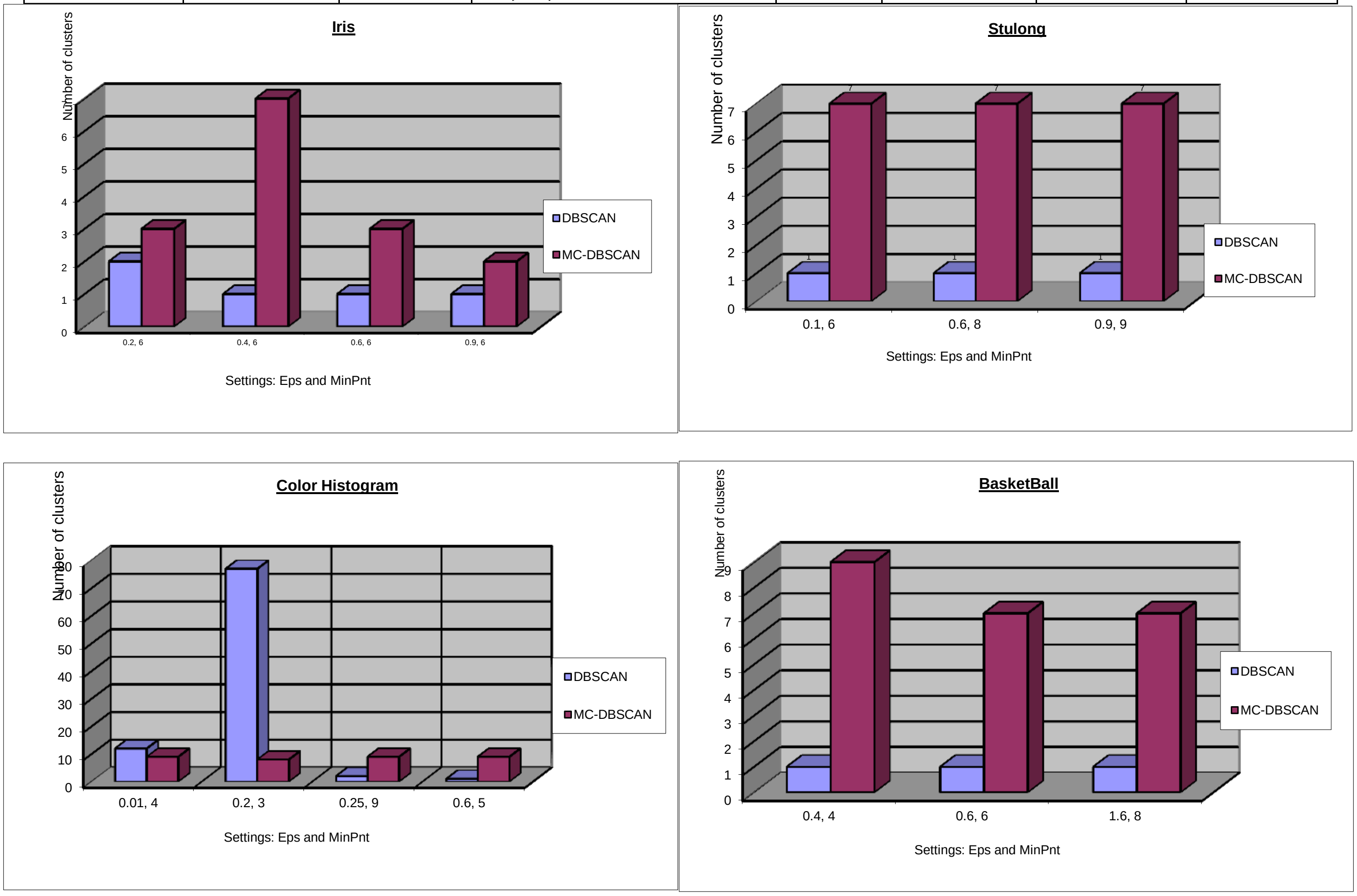
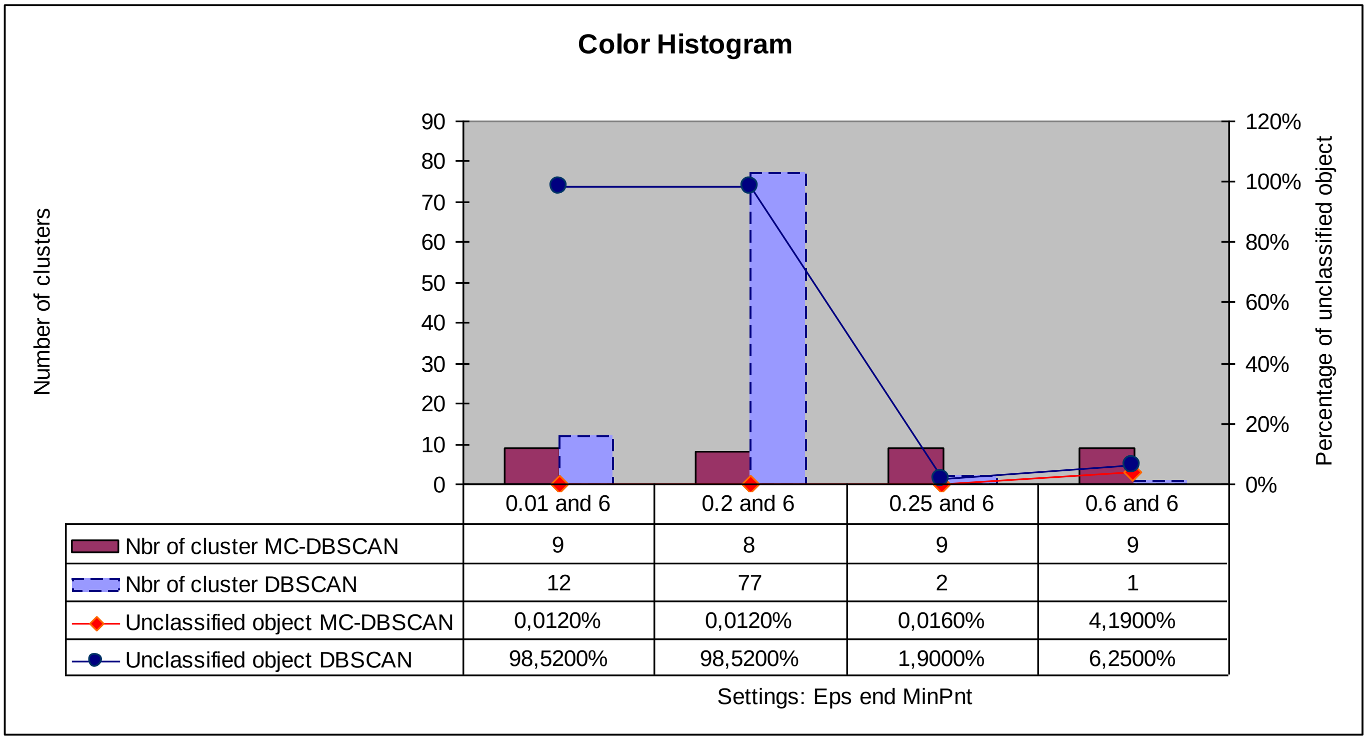


Fig. 1. Results assimilation of several clustering database by varying the parameters



	0.01 and 6	0.2 and 6	0.25 and 6	0.6 and 6
▬ Nbr of cluster MC-DBSCAN	9	8	9	9
▭ Nbr of cluster DBSCAN	12	77	2	1
◆ Unclassified object MC-DBSCAN	0,0120%	0,0120%	0,0160%	4,1900%
● Unclassified object DBSCAN	98,5200%	98,5200%	1,9000%	6,2500%

Fig. 2. Assimilation of the number of class obtained and the percentage of unclassified objects to the database "Color Histogram" between the two algorithms DBSCAN and MC-DBSCAN

The proposed algorithm allows for an experimental comparative study between the results by varying the relative importance regarding the criteria involved in the evaluation of assimilation between two actions.

Regarding the proposed algorithm, the weight change may influence the final outcome of a multi-criteria analysis "Fig. 3", while DBSCAN algorithm does not consider the indifference between the relative importances of each criterion.

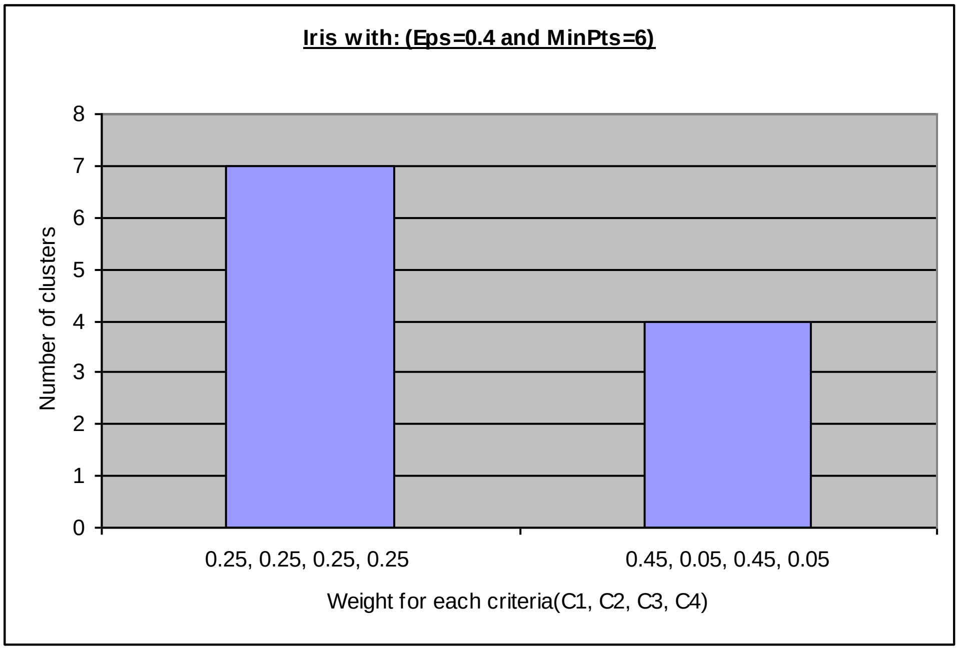


Fig. 3. Class Number obtained by the MC-DBSCAN algorithm applied to the Iris database by setting the input parameters (Eps = 0.4 and MinPts = 6) and by varying the relative importance to each criteria.Abbreviations and Acronyms

The purpose of this final test is to evaluate the performance of the suggested algorithm on the same database by increasing its size. In this test, we apply the MC-DBSCAN algorithm on the database "Color Histogram" of a varying size between 1300 and 65000 objects by changing the input parameters.

Reading the "Fig.4" show that even if the size of the database increases from 1300 up to 65,000 objects, the results remain in the standard, which explains that the added objects by increasing the size will affect the created classes but not the creation of new classes.

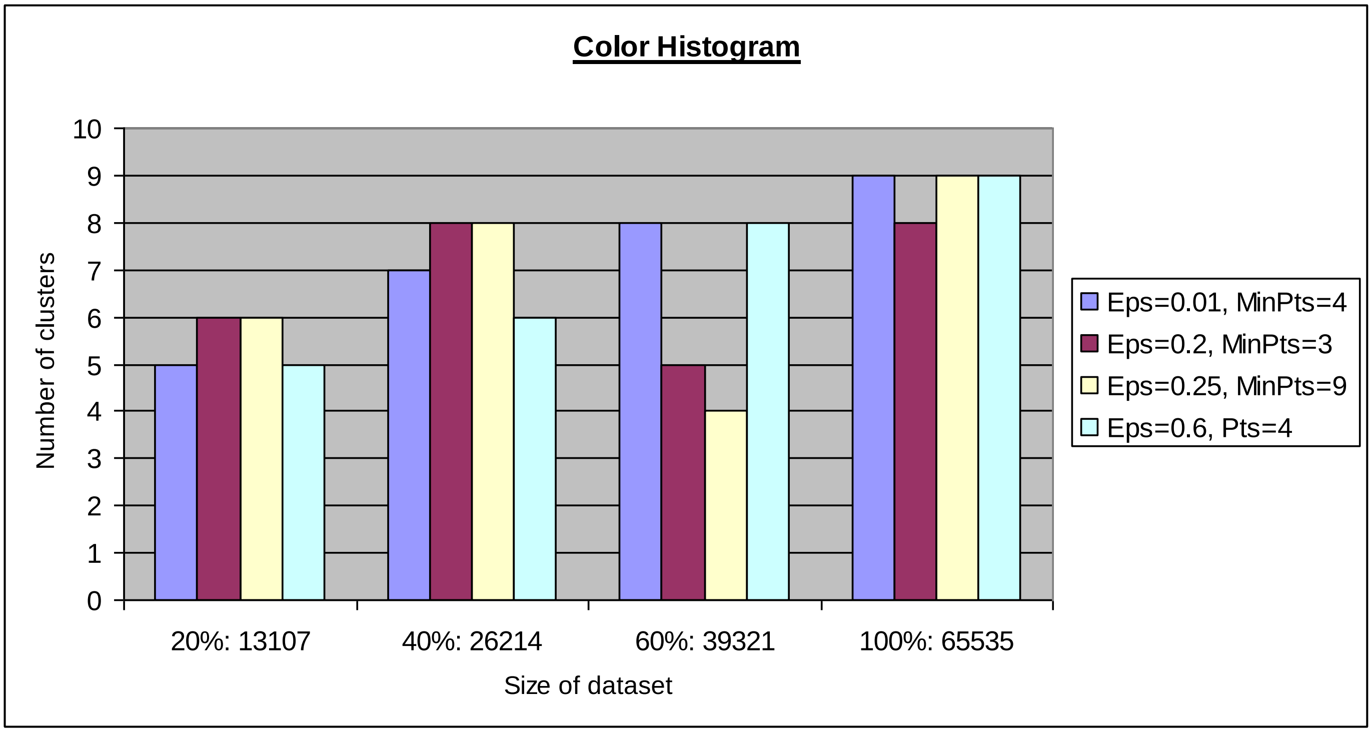


Fig. 4. Test result applied to the database "Color Histogram" by increasing the size of DataSet (Number of class obtained by increasing the size)

IV. Conclusion

This work has eventually reached a new clustering algorithm which contributes to resolving the multiple-criteria clustering problem with various weights to the relative importance to each criterion.

This new approach is based on the clustering by the enhancement of the DBSCAN algorithm which was merged with multiple-criteria decision-making.

However, it is necessary to highlight the need to further improve the performance of the algorithm. Because MC-DBSCAN like most clustering algorithms requires in advance a manual determination of input parameters.

It becomes clear that is by minimizing the human intervention relative to the determination of the input parameters will give us a better result.

References

[1] A. Ferligoj, and V. Batagelj, Direct multicriteria clustering algorithms. Journal of Classification, 9, 43-61, 1992.

[2] C. Zopounidis and M. Doumpos, Multicriteria classification and sorting methods: A literature review, European Journal of Operational Research, vol.138, no.2, pp.229-246, 2002.

[3] M. Doumpos and C. Zopounidis, A multicriteria discrimination method for the prediction of financial distress: the case of Greece, Multinational Finance Journal 3(2): 71–101,1999.

[4] M. Doumpos and C. Zopounidis, A multicriteria classification approach based on pairwise comparisons. European Journal of Operational Research, 158, 378–389, 2004.

[5] E. Dehghan Manshadi, M. Reza Mehregan and H. Safari, Supplier Classification Using UTADIS Method Based on Performance Criteria. International Journal of Academic Research in Business and Social Sciences February 2015, Vol. 5, No. 2 ISSN: 2222-6990

[6] B. Roy, A multicriteria analysis for trichotomic segmentation problems, In P. Nijkamp and J. Spronk (Eds), Multiple criteria analysis: Operational methods (pp. 245-257), Aldershot: Gower Press, 1981.

[7] M. Massaglia, and A. Ostanello, "N-TOMIC: A decision support for multicriteria segmentation probles", in: P. Korhonen (ed.), International Workshop on Multicriteria Desicion Support, Lecture Notes in Economics and Mathematics Systems 356, Springer-Verlag, Berlin, 167-174, 1991.

[8] V. Mousseau and R. Slowinski, Inferring an ELECTRE TRI model from assignment examples. Journal of Global Optimization, 12(2):157-174, 1998.

[9] Mousseau, V., Figueira, J., Naux, J.-Ph., Using assignment examples to infer weights for ELECTRE-TRI method: Some experimental results. European Journal of Operational Research 130 (2), 263–275, 2001.

[10] B. Roy, Presentation et interpretation de la methode ELECTRE TRI pour affecter des zones dans des categories de risque. Document du LAMSADE 124, Universite Paris-Dauphine, Paris, France, 2002.

[11] M. M. Köksalan, V. Mousseau, Ö. Özpeynirci, and S. Özpeynirci, A new outranking-based approach for assigning alternatives to ordered classes, Nav Res Log 56, 74–85, 2009.

[12] N. Belacel and M.R. Boulassel, Multicriteria fuzzy classification procedure PROCFTN: methodology and medical application. Fuzzy Sets and Systems, 141# 2, 203-217, 2004.

[13] N. Belacel, "Multicriteria assignment method PROAFTN: Methodology and medical applications", European Journal of Operational Research, 125, 175-183,2000.

[14] F. Al-Obeidat, N. Belacel, J. A. Carretero and P. Mahanti, "An evolutionary framework using particle swarm optimization for classification method PROAFTN". Applied Soft Computing 11 (8): 4971–4980. doi:10.1016/j.asoc.2011.06.003,2011.

[15] J. Léger and J.-M. Martel, A multicriteria assignment procedure for a nominal sorting problematic. European Journal of Operational Research, 138#2, 349-364, 2002.

[16] L. Henriet, "Evaluation systems and Multicriteria Classification To Help In The Decision, Construction Model And Assignment Procedures", Doctoral Thesis, University Paris Dauphine, 2000. ["Systèmes D'évaluation Et De Classification Multicritères Pour L'aide A La Décision, Construction De Modèles Et Procédures D'affectation". Thèse de doctorat en sciences, Université Paris Dauphine, 2000] (French).

[17] N. Belacel, " Multi-criteria classification methods: methodology and application to aid medical diagnosis", Doctoral Thesis, University Bruxelles, 1999. [Méthodes de classification multicritère: méthodologie et application à l'aide au diagnostic médical. thèse de doctorat: Université libre de Bruxelles (Service de mathématiques de la gestion). 1999.] (French)

[18] P. Hansen and B. Jaumard, Cluster Analysis and Mathematical Programming. Mathematic Programming, 79, 191-215, 1997.

[19] G. Cleuziou; "A method for unsupervised learning rules and seeking information classification", [Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. Other. Université d'Orléans, 2004] (French)

[20] C. Rocha, L.C. Dias and I. Dimas, "Multicriteria Classification with Unknown Categories: A Clustering-Sorting Approach and an Application to Conflict Management", Journal of Multi-Criteria Decision Analysis, Vol. 20, No. 1-2, 13–27, 2013.

[21] Y. De Smet and L. Montano Guzm´an, Towards multicriteria clustering: An extension of the k-means algorithm. European Journal of Operational Research, 158(2):390–398, oct 2004.

[22] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceeding of the fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[23] B. Roy and D. Bouyssou, " Multicriteria Decision: Methods and Case " [Aide Multicritère à la Décision: Methodes et cas. Economica: Paris, 1993.] (French)

[24] K. Kameshwaran1 and K. Malarvizhi, Survey on Clustering Techniques in Data Mining, International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2272-2276, 2014.

[25] M. Ester, H.P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland: Oregon, pp. 226-231,1996.

[26] B. Roy. "multi-criteria decision support methods" [Méthodologie multicritère d'aide à la décision. Economica, Paris, 1985.] (French)

[27] K. Jabeur, J.-M. Martel and S. Ben Khélifa, A distance-based collective preorder integrating the relative importance of the groups members, Group Decision and Negotiation 13, 327-349.26, 2004.

[28] R. B., Multicriteria methodology for decision aiding, Kluwer Academic Publishers, Dordrecht. 1996

[29] J.-L. Maricha, Fuzzy measures and integrals in the mcda sorting problematic, Th. Doct. Univ. Libre de Bruxelles 202; 2003.

[30] R. Bisdorff, P. Meyer and A.-L. Olteanu, A clustering approach using weighted similarity majority margins, Springer-Verlag Berlin Heidelberg, Volume 7120 of the series Lecture Notes in Computer Science pp 15-28, DOI. 10.1007/978-3-642-25853-4-2, 2011.

[31] P. M. Raymond Bisdor and A. L. Olteanu, Weighted similarity majority margins based multiple attributes clustering, ADMA. Advanced Data Mining and Applications Part I (LNAI 7120)15-28, 2011.

[32] J. Ruiz-Shulcloper, E. Alba-Cabrera and G. Sanchez-Diaz, DGLC: a density-based global logical combinatorial clustering algorithm for large mixed incomplete data, Geoscience and Remote Sensing Symposium, Proceedings, IGARSS, IEEE 2000 International, 2000.

[33] M. Ankerst, M. Breunig, H.P. kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," In Proc. 1999 ACM-SIGMOD Int. Conf. Management of data (SIGMOD'96), 1999.

[34] S. Xue-gang, C. Qun-xiu and M. Liang, "study on topic-based web clustering", The Journal of Chinese Information Processing, Vol 17, No. 3, pp.21-26, 2003.

[35] G. Sheikholeslami, S. Chatterjee and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases", Proceeding 24th International Conference on Very Large Data Bases, pp. 428-439, New York City, NY, 1998.

[36] R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", Proceeding ACM SIGMOD '98 International Conference on Management of Data, pp. 94-105, Seattle, WA, 1998.

[37] M. Shuai , W. TengJiao, T. ShiWei, Y. DongQing and G. Jun, A New Fast Clustering Algorithm Based on Reference and Density. Proc. of WAIM, pp. 214-225, 2003

[38] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD96) 226-231, 1996.

[39] G. Sheikholeslami, C. Surojit and Z. Aidong, Wavecluster: A multi-resolution clustering approach for very large spatial databases, Proceeding 24th International Conference on Very Large Data Bases 428-439, 1998.

[40] R. Anant, J. Sunita, S. Anand and K. Manoj, "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases", International Journal of Computer Application Volume 3,No.6, June 2010.

[41] Peng Liu, Dong Zhou, Naijun Wu," VDBSCAN: Varied Density Based Spatial Clustering of Application with Noise", in proceedings of IEEE Conference ICSSSM 2007 pg 528-531, 2007.

[42] A.K.M Rasheduzzaman Chowdhury and Md.Asikur Rahman, "An efficient Mehtod for subjectively choosing parameter k automatically in VDBSCAN",proceedings of ICCAE 2010 IEEE,Vol 1,pg 38-41, 2010.

[43] S. Richa, M. Bhawna and R. Anant, "Local Density Differ Spatial Clustering in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 3, March 2013

[44] Khushali Mistry, Swapnil Andhariya, Prof. Sahista Machchhar, "NDCMD: A Novel Approach Towards Density Based Clustering Using Multidimensional Spatial Data", International Journal of Engineering Research & Technology, ISSN: 2278-0181 www.ijert.org IJERTIJERT Vol. 2 Issue 6, June – 2013

[45] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial-temporal data." Data Knowl Eng 60(1):208–221, 2007.

[46] L.C. Dias and V. Mousseau. IRIS : A DSS for multiple criteria sorting. Journal of Multi-Criteria Decision Analysis, 12:285–298, 2003.

[47] Frank, A. Asuncion, Uci machine learning repository.

[48] J. ALCALÁ-FDEZ, A. FERNÁNDEZ, J. LUENGO, J. DERRAC, S. GARCÍA, L. SÁNCHEZ and F. HERRERA, Keel datamining software tool: Data set repository, integration of algorithms and experimental analysis framework, Journal of Multiple-Valued Logic and Soft Computing 17 (2-3) 255-287, 2011.

CAT5:A Tool for Measuring the Maturity Level of Information Technology Governance Using COBIT 5 Framework

[1]Souhaïl El ghazi El Houssaïni
Systems, architectures and networks team
EHTP
Casablanca, Morocco

[2] Karim Youssfi,
Architectures and system team,
LISER laboratory
ENSEM
Casablanca, Morocco

[3]Jaouad Boutahar,
Systems, architectures and networks team
EHTP
Casablanca, Morocco

***Abstract*—Companies have more and more trends to automate their operational and organizational activities, therefore the investment of information technology (IT) continues to increase every year. However, good governance that can ensure the alignment of IT and business strategy and realized benefits from IT investments has not always followed this increase. Measurement of IT governance is then required as a basis for the continuous improvement of the IT services. This study is aimed at producing a tool CAT5 to measure the maturity level of IT governance, thus facilitating the process of improvement of IT services. CAT5 is based on COBIT 5 framework and the design used is Unified Modeling Language. Through stages of information system development, this research results in an application for measuring the maturity level of IT governance that can be used in assessing existing IT governance.**

Keywords—COBIT5; IT Governance; Process Capability Model; Maturity Model; CAT5; Process Assessment

I. Introduction

In recent years, due to the increase of IT investment, the IT governance has become a center of interest among practitioners and researchers.

Several issues made its contribution to explain this phenomenon [1]: (1) Business activities became largely dependent in IT systems. (2) Therefore, business failure and success are increasingly dependent on IT (3) IT should deliver value to business and be aligned with the organization’s goals. (5) Response to fast changes in business environment. (6) Ensure business continuity.

Further, companies wishing to implement IT Governance must establish a system to measure the maturity of their IT organizations. The purpose of process capability determination is to identify the strengths, weaknesses and risk of each IT processes with respect to a particular specified requirement through the processes used and their alignment with the business need.

It will be interesting to analyze this issue in relation to a largely well-accepted framework as COBIT [4]-currently in its fifth edition- covering the IT activities of the enterprise end to end, therefore, this study has developed a tool CAT5 to measure the maturity level of IT governance: this tool will helps in conducting self-assessments and determining to what extent the implementation of IT governance has been done.

Such assessments, will normally be used as part of an enterprise’s process improvement program and can then be used to report internally to an enterprise’s executive management or board of directors on the current capability of its IT processes and facilitates the development of recommendations and improvements on each weak process, explained in a previous article, that present a design of “roadmap of IT governance implementation” [12]. The measurement results illustrate the current state to facilitate improvement of IT governance. All processes that exist within the framework of COBIT 5 are used and measured based on COBIT 5 attributes and criteria Capability Model is based on ISO/IEC 15504 (SPICE).

This document is organized as follows: first, we introduce the research approach; afterward, we present a summary of requirements on good IT governance maturity assessments; and we explore the evolution of maturity model to COBIT 5 process capability model. Then, we outline the existing tools of IT governance assessment. At the end of this paper, we present the result of the study by showing the design and the layout of the proposed tool CAT5.

II. Research Approach

This research is an engineering study, where the end result is a solution that is used to measure the maturity level of IT governance. For that purpose, the research method used is based on the phases of Software Development Life Cycle (SDLC) and prototyping approach. The final result is still a prototype continues to be developed in line with the needs of on ongoing PhD research which subject is the “Implementation of tools supporting the establishment of IT governance”. The research starts with the analyzing of the needs of the system, and then proceeds with the design and manufacturing system design using UML and the last is the creation of applications using java/j2ee and MySQL tools. The first phase of testing by the developer uses “Whitebox” Testing approach and on the user side uses “Blackbox” Testing. The stage of design uses diagrams provided in UML, which are the Use Case Diagram, Activity Diagram, Class

Diagram, Sequence Diagram, Collaboration Diagram, Component Diagram and Deployment Diagram.

The literature search was performed from two main sources:

- The first source is COBIT guides [6, 7, and 8]: Except for the documentations provided by ISACA to their members, there is a lack of important documentation from other sources regarding the latest version of the framework. For this reason, this paper is based on ISACA documentation.
- The second source is the articles on the subject, Sample search terms include IT governance, COBIT, COBIT5, Assessment Process, Capability Model, Maturity Model, Process Assessment.

III. Literature Review

In measurement theory, the goodness of an assessment is specified in terms of validity and reliability [9]. For practical applications, these benefits need to be traded against the cost of performing the measurement. Table 1 summarized a set of requirements within the domains of validity, reliability and cost.

Process maturity has been a core component of COBIT for more than a decade. Determining the level of process maturity for given processes allows organizations to determine which processes are essentially under control and those that represent potential management challenges [10].

The concept of process maturity in earlier versions of COBIT was adopted from the Software Engineering Institute’s. In the last version COBIT Capability Maturity Model has been replaced by the concept of process capability [6] based on the ISO/IEC 15504 (SPICE) standard ‘‘Information Technology—Process Assessment.’’ the COBIT assessment program [6] is designed to provide enterprises with a repeatable, reliable and robust methodology for assessing the capability of their IT processes.

TABLE I. Requirements on Good It Governance Maturity Assessments [5]

ID	Requirement	Domain	Description
Req1	Consistency with common conceptions	Validity	The method should be based on well-known IT governance sources within academia and practice.
Req2	Descriptive operationalization	Reliability	The method should support unambiguous and objective depiction of IT governance in an organization by means of a precise representation. If two analysts individually face the task of describing the IT governance in an organization, a descriptively operationalized language would result in both obtaining equal models, while a fuzzier language would not.
Req3	Normative operationalization	Reliability	The method should support unambiguous and objective analysis of IT governance. It should clearly state how different IT governance concerns affect maturity scores.
Req4	Support for efficient data collection	Cost	The method should provide an efficient representation of IT governance so that data could be collected with little effort.
Req5	Support for efficient analysis	Cost	The method should support efficient normative judgments of IT governance so that analysis can be made easily and at a reasonably low cost.

A. COBIT 5 Process Capability Model

The Capability Model is based on ISO/IEC 15504 (SPICE):

- ✓ Level 0: Incomplete. The process is not implemented or fails to achieve its purpose;
- ✓ Level 1: Performed (Informed). The process is implemented and achieves its purpose;
- ✓ Level 2: Managed (Planned and monitored). The process is managed and results are specified, controlled and maintained;
- ✓ Level 3: Established (Well defined). A standard process is defined and used throughout the organization;
- ✓ Level 4: Predictable (Quantitatively managed). The process is executed consistently within defined limits
- ✓ Level 5: Optimizing (Continuous improvement). The process is continuously improved to meet relevant current and projected business goals.

The capability of processes is measured using process attributes, except for the first level (Level 0) in which the goal of the process is not achieved; in all the other levels there is at least one attribute

The international standard defines nine process attributes [13]:

1.1 Process Performance

2.1 Performance Management

2.2 Work Product Management

3.1 Process Definition

3.2 Process Deployment

4.1 Process Measurement

4.2 Process Control

5.1 Process Innovation

5.2 Process Optimization.

Each process attribute is assessed on a four-point (N-P-L-F) rating scale:

- Not achieved (0 - 15%)
- Partially achieved (>15% - 50%)

- Largely achieved (>50%- 85%)
- Fully achieved (>85% - 100%)

In COBIT 5 to achieve a given level of capability, the previous level has to be completely achieved.

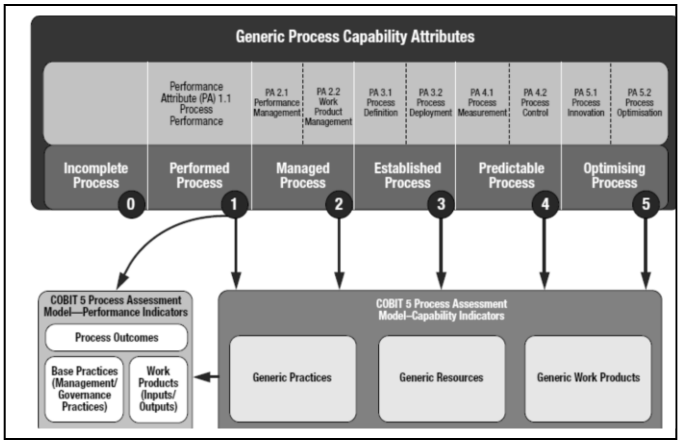

Generic Process Capability Attributes
Performance Attribute (PA) 1.1 Process Performance
PA 2.1 Performance Management
PA 2.2 Work Product Management
PA 3.1 Process Definition
PA 3.2 Process Deployment
PA 4.1 Process Measurement
PA 4.2 Process Control
PA 5.1 Process Innovation
PA 5.2 Process Optimisation
Incomplete Process
0
Performed Process
1
Managed Process
2
Established Process
3
Predictable Process
4
Optimising Process
5
COBIT 5 Process Assessment Model—Performance Indicators
Process Outcomes
Base Practices (Management/ Governance Practices)
Work Products (Inputs/ Outputs)
COBIT 5 Process Assessment Model–Capability Indicators
Generic Practices
Generic Resources
Generic Work Products

Fig. 1. COBIT 5 Process Capability Model[8]

COBIT Process Assessment Model, describe the assessment process activities and an assessment model walkthrough for a proper assessment as shown in figure 2:

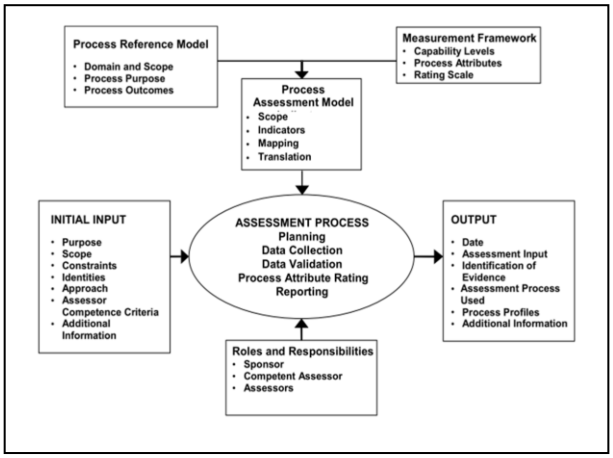


Fig. 2. COBIT 5 Process Capability Model[8]

Initiation: The objective of the initiation phase is to ensure that there is a common understanding with the sponsor on the purpose and scope of the assessment, and to identify the individuals with the appropriate competencies to ensure a successful assessment.

Planning Assessment: The Assessment Planning phase includes such things as: determine the assessment activities, determine the necessary resources and schedule for the assessment, define how the assessment data will be collected, recorded, stored, analyzed and presented and define the planned outputs of the assessment.

Data Collection: The assessor obtains (and documents) an understanding of the process (es) including process purpose, inputs, outputs and work products, sufficient to enable and support the assessment:

- Evidence of process performance for each process within the scope. Evidence includes observation of work products and their characteristics, testimony from the process performers, and observation of the infrastructure established for the performance of the process.
- Evidence of process capability for each process within the scope. Evidence of process capability may be more abstract than evidence of process performance.

In some cases, the evidence of process performance may be used as evidence of process capability.

Data Validation: The assessor ensures that the data collected is correct and objective and that the validated data provides complete coverage of the assessment scope.

Process Attribute Rating: For each process assessed, a rating is assigned for each process attribute up to and including the highest capability level defined in the assessment scope. The rating is based on data validated in the previous activity.

Traceability must be maintained between the objective evidence collected and the process attribute ratings assigned. For each process attribute rated, the relationship between the indicators and the objective evidence is recorded.

Reporting: The results of the assessment are analyzed and presented in a report .The report also covers any key issues raised during the assessment such as:

- Observed areas of strength and weakness
- Findings of high risk, i.e., magnitude of gap between assessed capability and desired/required capability

B. *Available tools*

There are a number of tools:

1) Self-assessment Templates—an Excel file with separate evaluation sheets for all 37 COBIT 5 processes included within the COBIT guidance [13].

2) The COBIT online “Self-Assessment Tool” [14] is a web based tool available online that allows a user registered with a paid member account to enter basic information about the self-assessment, set expected Capability Levels for each process capability before you begin the assessment, and select only those processes you want to assessment. The tool then walks you through the assessment of each attribute and provides a report that details the results of your self-assessment.

Except for the documentations provided by ISACA to their members, there is a lack of important documentation from other sources regarding the latest version of the framework.

3) Process Maturity Assessment in iServer: iServer IT Governance Accelerator is a paid solution released by “Orbus Software” [15], that provides comprehensive toolkit for organizations wishing to reference, adopt and align with COBIT 5 IT governance best practices, the toolkit include:

- A preconfigured iServer repository based on COBIT5

- COBIT5 iServer meta-model, highlighting TOGAF touch-points
- Complete models of all COBIT5 principles and concepts, with relationships and interdependencies easily reported on using iServer's relationship matrix tool.
- Central repository for all IT governance documentation

For the maturity assessment iServer allow to import spreadsheet data from the self-assessment templates into iServer Governance Repository. Once the data is available, there are a number of reports and views that can be generated.

We point out that our research has aim at realization of solution of IT governance; risk and compliance .In this context a first module "roadmap of IT governance implementation" was developed [12]. Our second module will be an IT Governance self assessment tool such assessments will normally be used as part of an enterprise's process improvement program and can then be used to report internally to an enterprise's executive management or board of directors on the current capability of its IT processes and facilitates the development of recommendations

This study showed that existing tools available solutions are not free, and the compatibility with our first module is not provided. Therefore, to overcome these issues, we decide to develop CAT 5 (COBIT 5 Assessment Tool) that helps assessor to conduct measurement of IT process governance maturity level.

IV. CAT5: A TOOL FOR MEASURING THE MATURITY LEVEL OF IT GOVERNANCE USING COBIT 5 FRAMEWORK

The followings are the results of the research in the form of design using UML and the layout of **CAT5,** which is a web-based tool for measuring the maturity level of IT governance. The application design shown includes Use Case, Activity and Class Diagrams.

A. *Use Case Diagram*

Use Case diagram, as in Figure 3, describes the relationship of the functionality contained in the application. Actors within the system are Admin and Auditor. The functions that can be performed by the Admin: Manage Questioner function (this function has sub-functions of Update, Delete Criteria of each process attributes and View Questioner), Manage Score and Analysis function (this function has sub-functions of Input Score and Analysis, Generate Maturity Level and Generate Diagram), Manage Report function (this function has sub-functions of Generate Diagram and View Report). While the auditor has the functions of View Questioner Auditor, Input Score and Analysis, and View Report.

B. *Activity Diagram*

The Activity Diagram, as in Figure 5, describes the business process "Perform an assessment" in the system, covering the following:

Auditor does Login to the system using Username and Password;

System does User Verification, if true, and then the main menu is displayed;

- Auditor chooses the Menu Home: The heat map view show;
- Auditor chooses a specified process;
- Auditor plan an assessment (define start and end assessment dates and participants)
- Auditor inputs Score of each criteria;
- Auditor inputs comment;
- Auditor inputs files as evidences of the given score;
- System processes scoring;
- When completed, system will do Generate Maturity Level;
- When completed, system will do Generate Diagram;
- System will display Report;
- If the Report needs to be printed, Auditor will do Print Report;
- When all the processes have been completed, Auditor can logout from the system.

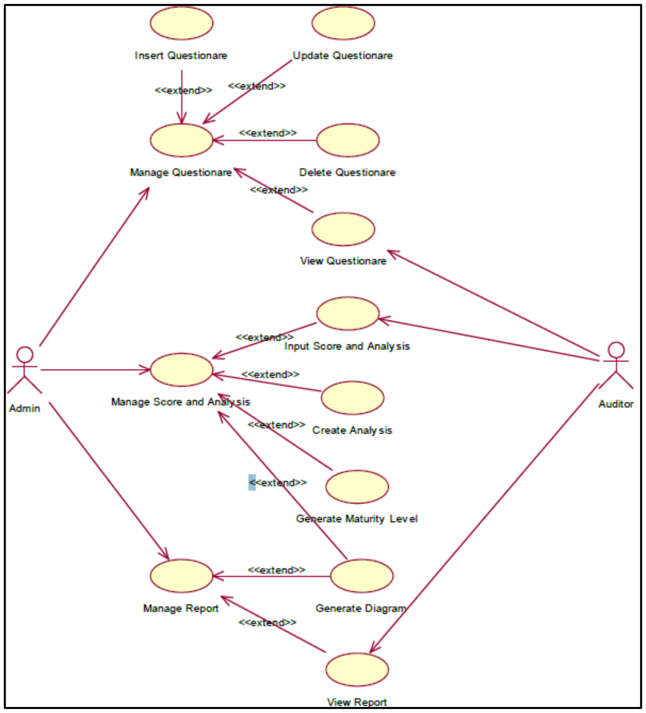


Fig. 3. Use case Diagram

C. *Class Diagram*

This application is an object-oriented information system. Relationships between objects in the system are described using class diagram, as in **Figure 4**.

The figure shows seven classes, and each class has attributes; the classes include:

- Class Process
- Class Maturity Assessment
- Class Assessment Plan
- Class Assessment form
- Attribute
- Attribute scale
- Criteria

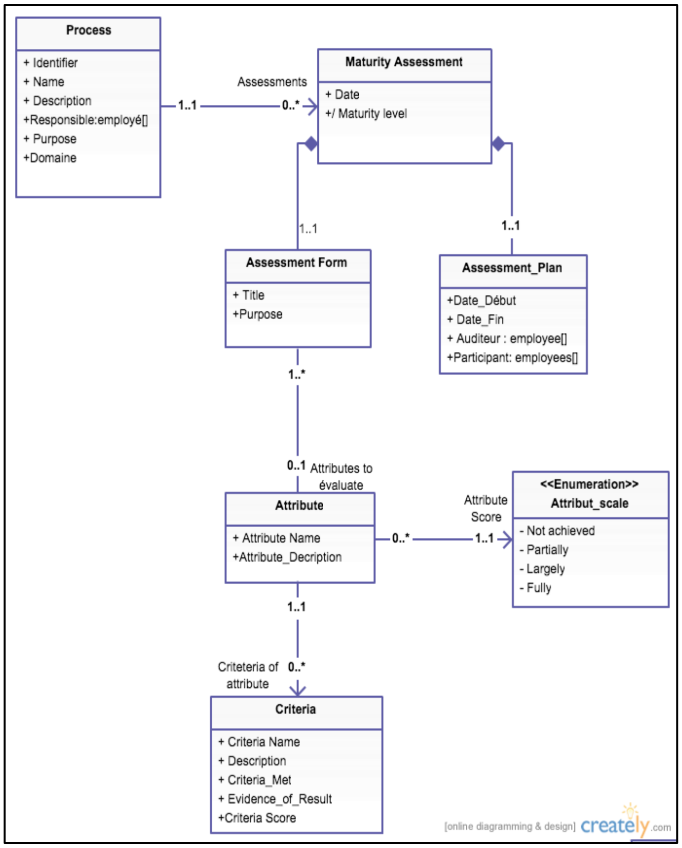


Fig. 4. Class Diagram

D. Application Layaouts

The following is the layout of CAT5 a tool for measuring the IT governance maturity levels IT that builds upon the existing design. After log as a user, the first screen shows a heat map view of COBIT5 process maturity as in Figure 6: the heat map view shows the as-is situation of the of all COBIT5 process, each process is colored according to its current level of maturity, and the color scale is at the top right of the screen.

The button “>” allows access to the history of all performed assessments of a selected process as shown in Figure 7: a user (admin and auditor) can consult or delete an old maturity assessment.

The button “+” allows to perform a new assessment of a selected process, the first step is to plan an assessment, as shown in Figure 8, plan assessment includes defining the start and end date of the assessment and participants.

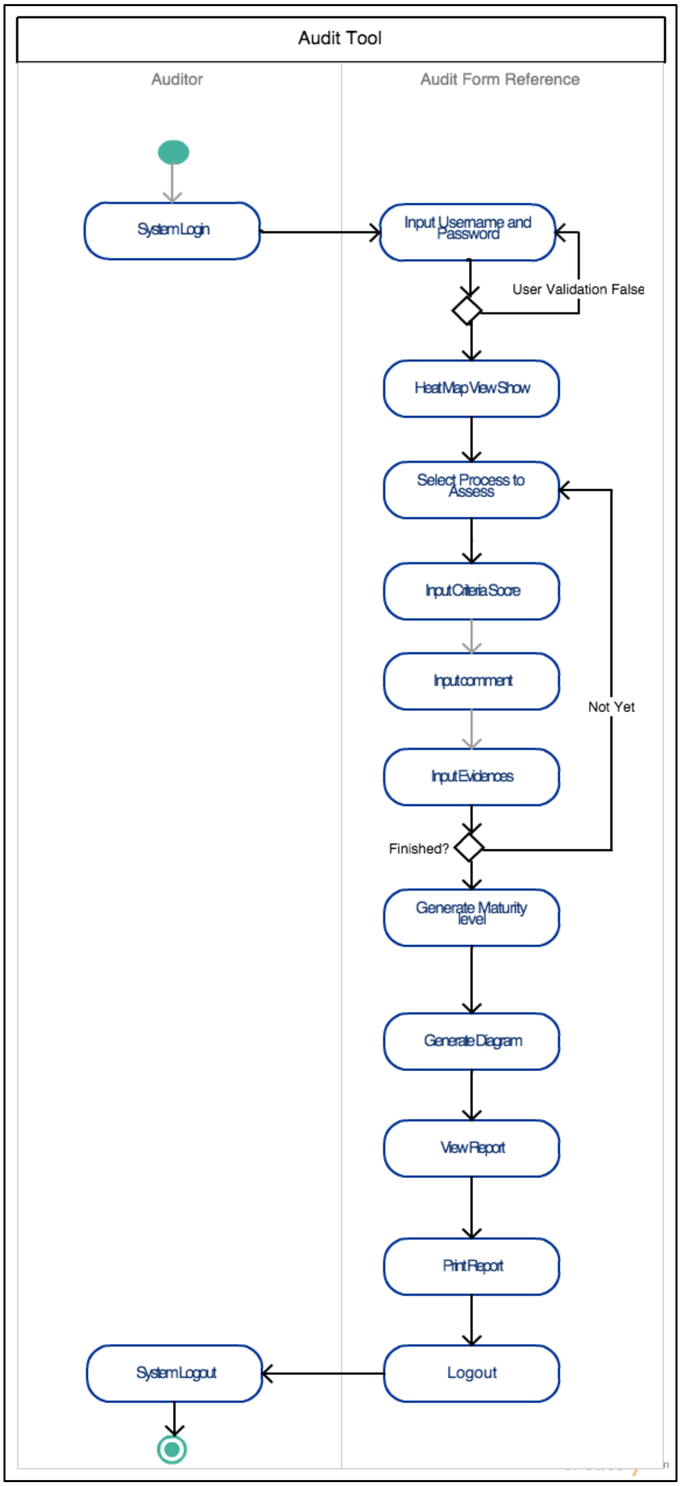


Fig. 5. Activity diagram

The second step is to perform the maturity assessment, as printed excel self-assessment Templates—questionnaire provided by ISACA, auditor will do the assessment based on criteria and attributes of each process, as shown in Figure 9. There are several methods used in conducting the audit, which are Interview and Document Check. Furthermore the auditor fills out a score of each criteria, evidences for the given score

(Tools or equipment used in conducting the audit) and his field findings as a comment.

Afterward, auditor can generate the assessment report, the result of the assessment will look like in figure 10, that shows a dashboard indicating the score for each attribute and the maturity level of the assessed process calculated based on COBIT 5 framework, this new result will impact the heat map view that's shows the latest maturity level of the assessed process.

Admin can customize Update or Delete criteria following the specificity of the organization as shown in figure 11.all new assessments will follow this modified assessment Form

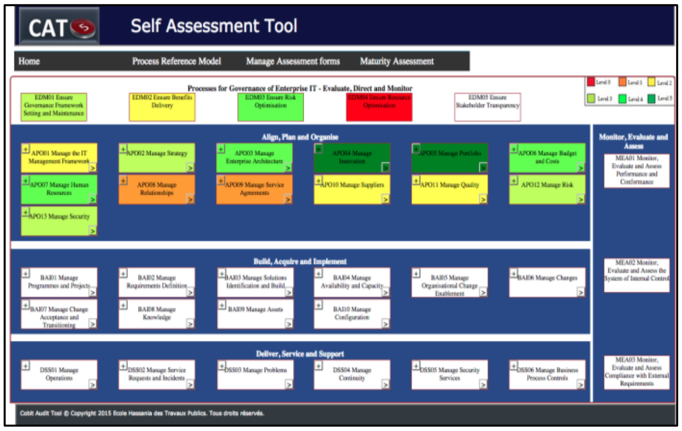
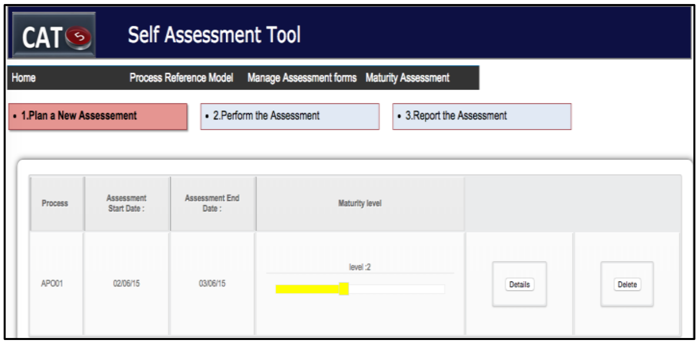
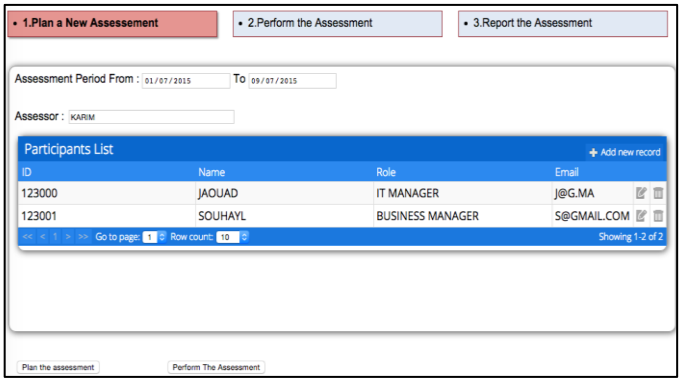

CAT5
Self Assessment Tool
Home
Process Reference Model
Manage Assessment forms
Maturity Assessment
Level 0
Level 1
Level 2
Level 3
Level 4
Level 5
Processes for Governance of Enterprise IT - Evaluate, Direct and Monitor
EDM01 Ensure Governance Framework Setting and Maintenance
EDM02 Ensure Benefits Delivery
EDM03 Ensure Risk Optimisation
EDM04 Ensure Resource Optimisation
EDM05 Ensure Stakeholder Transparency
Align, Plan and Organise
APO01 Manage the IT Management Framework
APO02 Manage Strategy
APO03 Manage Enterprise Architecture
APO04 Manage Innovation
APO05 Manage Portfolio
APO06 Manage Budget and Costs
APO07 Manage Human Resources
APO08 Manage Relationships
APO09 Manage Service Agreements
APO10 Manage Suppliers
APO11 Manage Quality
APO12 Manage Risk
APO13 Manage Security
Monitor, Evaluate and Assess
MEA01 Monitor, Evaluate and Assess Performance and Conformance
Build, Acquire and Implement
BAI01 Manage Programmes and Projects
BAI02 Manage Requirements Definition
BAI03 Manage Solutions Identification and Build
BAI04 Manage Availability and Capacity
BAI05 Manage Organisational Change Enablement
BAI06 Manage Changes
BAI07 Manage Change Acceptance and Transitioning
BAI08 Manage Knowledge
BAI09 Manage Assets
BAI10 Manage Configuration
MEA02 Monitor, Evaluate and Assess the System of Internal Control
Deliver, Service and Support
DSS01 Manage Operations
DSS02 Manage Service Requests and Incidents
DSS03 Manage Problems
DSS04 Manage Continuity
DSS05 Manage Security Services
DSS06 Manage Business Process Controls
MEA03 Monitor, Evaluate and Assess Compliance with External Requirements
Cobit Audit Tool © Copyright 2015 Ecole Hassania des Travaux Publics. Tous droits réservés.

Fig. 6. Home Page of CAT5 :HeatMap View of COBIT5 process



Process	Assessment Start Date :	Assessment End Date :	Maturity level		
APO01	02/06/15	03/06/15	level :2	Details	Delete

Fig. 7. Consult or delete old maturity assessments


• 1.Plan a New Assessement
• 2.Perform the Assessment
• 3.Report the Assessment

Assessment Period From : 01/07/2015 To 09/07/2015

Assessor : KARIM

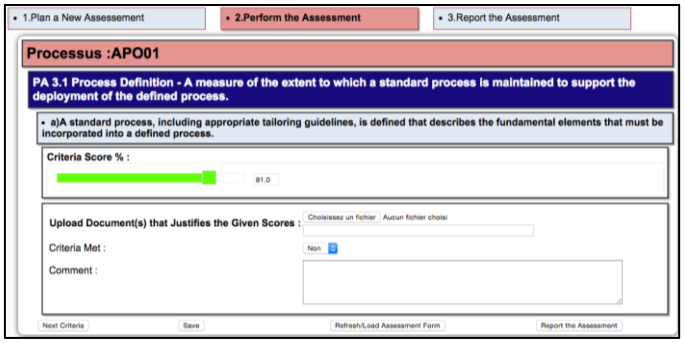
Participants List

ID	Name	Role	Email
123000	JAOUAD	IT MANAGER	J@G.MA
123001	SOUHAYL	BUSINESS MANAGER	S@GMAIL.COM

Showing 1-2 of 2

+ Add new record
<< < 1 > >> Go to page: 1 Row count: 10
Plan the assessment
Perform The Assessment

Fig. 8. Plan a new assessment


• 1.Plan a New Assessement
• 2.Perform the Assessment
• 3.Report the Assessment

Processus :APO01

PA 3.1 Process Definition - A measure of the extent to which a standard process is maintained to support the deployment of the defined process.

- a)A standard process, including appropriate tailoring guidelines, is defined that describes the fundamental elements that must be incorporated into a defined process.

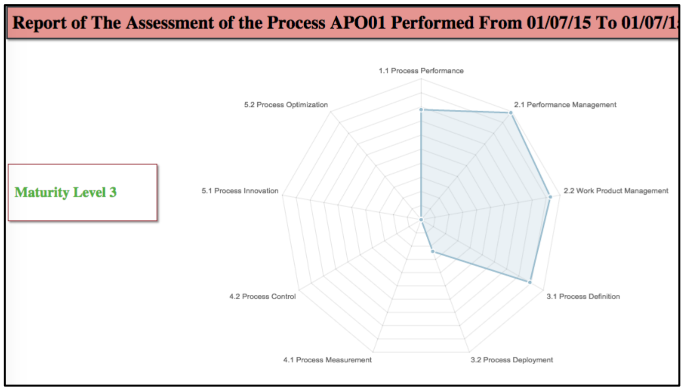
Criteria Score % : 81.0

Upload Document(s) that Justifies the Given Scores : Choisissez un fichier Aucun fichier choisi

Criteria Met : Non

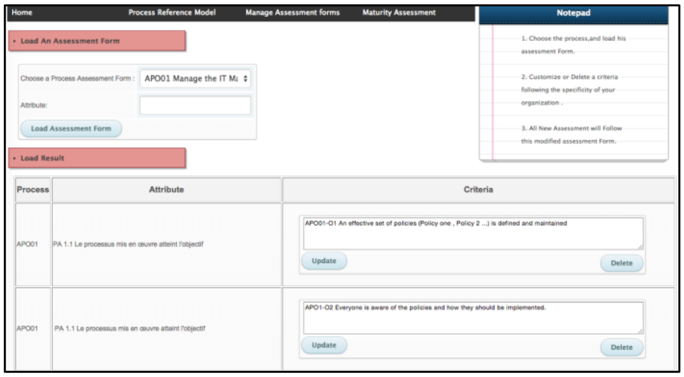
Comment :

Fig. 9. Perform an assessment


Report of The Assessment of the Process APO01 Performed From 01/07/15 To 01/07/1

Maturity Level 3

Fig. 10. Assessment Report



Notepad

1. Choose the process,and load his assessment Form.

2. Customize or Delete a criteria following the specificity of your organization .

3. All New Assessment will Follow this modified assessment Form.

Choose a Process Assessment Form : APO01 Manage the IT Ma

Attribute:

Process	Attribute	Criteria
APO01	PA 1.1 Le processus mis en œuvre atteint l'objectif	APO01-O1 An effective set of policies (Policy one , Policy 2 ...) is defined and maintained Update Delete
APO01	PA 1.1 Le processus mis en œuvre atteint l'objectif	APO1-O2 Everyone is aware of the policies and how they should be implemented. Update Delete

Fig. 11. Update Assessment Form

V. Conclusion and Future Research Direction

This article is part of a research that aims at realization of solution of IT governance, risk and compliance. A previous article [12] present our first module "roadmap of IT governance implementation" was developed [12]. This article present CAT5 as a second module for measuring IT governance by utilizing the COBIT5 framework. Such assessments will normally be used as part of an enterprise's process improvement program and can then be used to report internally to an enterprise's executive management or board of directors on the current capability of its IT processes and facilitates the development of recommendations and improvements on each weak process.

Further research is ongoing to provide a third module "Execution of COBIT5 roadmap action plans".

The implementation guide describes briefly some indicators such potential benefit, ease of implementation (cost, effort, sustainability), and risk; other economic and financial, we will provide a set of key indicators in order to give a widespread support decision- making in the selection and prioritization change scenarios indicators like value creation, and ROI will be considered as evaluation variables.

COBIT 5 management practices, and Other Specific frameworks: such as PMBOK, can also provide guidance through for this step.

References

[1] JoséTomaz,“Using Agent-based Simulation for IT Governance Evaluation”, Proceedings of the 45th HICSS,Doctoral Program on Complexity Sciences, June 15, 2011, pp. 1.

[2] Weill, P. and J.W. Ross, “IT governance – How Top Performers Manage IT Decision Rights for Superior Results”, Harvard Business School Press, 2004.

[3] International Organization for Standardization, ISO/IEC 20000-1 & ISO/IEC 20000-2, 2005.

[4] Information Systems Audit and ControlAssociation, Control Objectives for Information andRelated Technology, 4th Edition, 2005.

[5] M. Simonsson and P. Johnson, “Model-based IT governance maturity assessments with COBIT”, In proceedings of the European Conference on Information Systems, St. Gallen, Switzerland, 2007

[6] ISACA, COBIT 5: A Business Framework for the Governance and Management of Enterprise IT, 2012.

[7] ISACA, COBIT 5: Implementation, 2012.

[8] ISACA, Assessor Guide: Using COBIT®5, 2013.

[9] The GeNIE tool for decision theoretic models,http://genie.sis.pitt.edu/, page visited June 12, 2015.

[10] Weill, P. and J.W. Ross, “IT governance – How Top Performers Manage IT Decision Rights for Superior Results”, Harvard Business School Press, 2004.

[11] IT Governance Capabilities,http://www.orbussoftware.com/governance-risk-and-compliance/it-governance-accelerator/ visited June 12, 2015.

[12] K Youssfi, J Boutahar, S Elghazi, “IT Governance implementation: A tool design of COBIT 5 roadmap”, In proceedings of the Second World Conference on Complex Systems (WCCS), 2014, pp. 115-121

[13] ISACA, COBIT Self-Assessment Guide: Using COBIT 5, 2012.

[14] Self-Assessment,https://cobitonline.isaca.org/publications visited June 12, 2015.

[15] iServer for IT Governance Risk and Compliance, http:// www.orbussoftware.com/governance-risk-and-compliance/ , page visited June 28, 2015.

A Platform to Support the Product Servitization

Giovanni Di Orio
CTS-UNINOVA: Dep. de Eng. Electrotécnica
FCT-UNL
Caparica, Portugal

Dragan Stokic
Institute for Applied System Technology
ATB-Bremen
Bremen, Germany

Oliviu Matei
Technical University of Cluj-Napoca: Dept. of Elect. Eng.
North University Center of Baia Mare
Baia Mare, Romania

José Barata
CTS-UNINOVA: Dep. de Eng. Electrotécnica
FCT-UNL
Caparica, Portugal

Sebastian Scholze
Institute for Applied System Technology
ATB-Bremen
Bremen, German

Claudio Cenedese
GTC – Global Technology Center
Electrolux SpA
Pordenone, Italy

***Abstract*—Nowadays manufacturers are forced to shift from their traditional product-manufacturing paradigm to the goods-services continuum by providing integrated combination of products and services. The adoption of service-based strategies is the natural consequence of the higher pressure that these companies are facing in the global markets especially due to the presence of competitors which operate in low wage region. By betting on services, or more specifically, on servitization manufacturing companies are moving up the value chain in order to move the competition from costs to sophistication and innovation. The proliferation of new emerging technologies and paradigms together with a wider dissemination of information technology (IT) can significantly improve the capability of manufacturing companies to infuse services in their own products. The authors present a knowledge-based and data-driven platform that can support the design and development of Product Extended by Services (PESs) solutions.**

Keywords—Product-Services System; Servitization; Service-Oriented Architecture; Ambient Intelligence; Context Awareness; Data Mining

I. Introduction

In the pursuit of competitiveness, European manufacturing companies are required to create value by designing and producing the so called products of the future that satisfy an heightened costumer awareness and needs, improve their own operational efficiency and effectiveness while enabling market expansion in Europe and abroad [1]. As a response to this need, manufacturing companies are increasingly shifting from pure manufacturing and delivering of physical product to the provisioning of sophisticated integrated solutions where physical products are enhanced by functions and services [2]. This business trend can be designed as servitization that means the process of creating value in products and goods by adding services. The term was initially coined by Vandermerewe & Rada [3], and now is widely recognized and adopted to identify a specific competitive manufacturing strategy as pointed in [4]. The integration of products and services offers innovative and sophisticated solutions that are distinctive and, above all, easier to defend from competition based in lower cost economies [4]. The trend in servitization is confirmed in [5], where it is claimed that European manufacturing industries of today are under high pressure in the global market due to: 1) the presence of competitors which operate in regions with low-wage and are absolving very fast the available technologies; 2) and to the need to keep pace with science-based innovation processes and products that are creating new markets and new business. The fierce competition for key markets share between manufacturing companies are boosting the search for innovation in both processes and products in order to enable the paradigm migration from cost-oriented to High-Adding-Value (HAV) manufacturing (see Fig. 1). Therefore, the key to competitiveness passes through the capability to provide innovative products, i.e. to provide products that encompasses components, consumer goods and capital goods while extending them with services.

Considering this baseline the research challenge addressed by this work is:

“Which tools and methodology should be implemented and included in manufacturing companies to enable the extensions of products by services for global markets?”.

II. Contribution to Cyber-Physical Systems

As stated in [1] and [6], the provisioning of both products and associated services according to the Product-Service System (PSS) approach will incredibly benefit from an increased product/process intelligence and an overall manufacturing enterprise infrastructure to support both the Product Lifecycle Management (PLM) and Service Lifecycle Management (SLM) integration. Current trends and developments in emerging technologies, such as Internet of Things (IoT), service-oriented, high-performance and distributed computing combined with the increasing advances in manufacturing technologies – embedded control and monitoring systems are radically changing the way product and processes are designed to cope with additional features, improved monitoring and performance – could potentially trigger a new generation of systems which main capabilities relies on local on-device distributed intelligence empowered by global accessibility over the cloud.

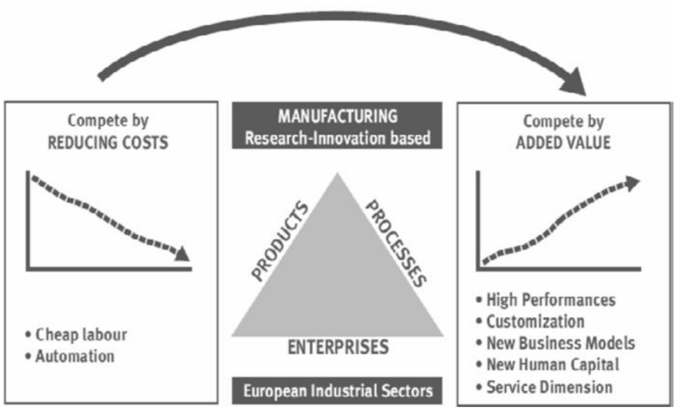


Fig. 1. Competition Shift from reducing costs to High-Added-Value (HAV) [5]

Current work benefits from the CPS research domain that in turn relies on on-device functionalities for monitoring, controlling, optimizing and adapting physical entities, as well as, in-cloud functionalities for delivering advanced features. However, the usage of embedded mechatronics components that combine the physical part with new technologies such as innovative materials, nanotechnologies, and information and communication technology is a necessary but not sufficient condition to allow data extraction from the environment and data processing. As a matter of fact to fully implement the PLM approach, it is necessary to have a totally integrated manufacturing information system for integrating people, data, processes, and business systems and providing product information for manufacturing companies and their extended enterprise , as they are presented in [7]. The combination of smart embedded systems together with a SOA-based infrastructure will allow the extraction of knowledge from products while enabling a fast response in the most different conditions and the design, development and provisioning of services and/or enhanced functionalities associated to these products. In this scenario current work provides such SOA-based infrastructure to fully exploit the capabilities of cyber-physical systems.

III. Related Areas and Supporting Concepts

A. *Product-Services System*

As stated in [8], the first formal definition of PSS has been given by Goedkoop in [9] that identifies its three fundamental elements: i) Product; ii) Service; and iii) System. During the years several definitions for PSS have been given. For instance in [10] a PSS is defined as; "a system of products, services, networks of players and supporting infrastructure that continuously strives to be competitive, satisfy customer needs and have a lower environmental impact than traditional business models". In [11] PSS is "an innovation strategy, shifting the business focus from designing (and selling) physical products only, to designing (and selling) a system of products and services which are jointly capable of fulfilling specific client demands". Although the different definitions given all of them adding some elements to the core definition given by Goedkoop. Despite the particular definition of PSS, all the authors in the literature agree in thinking on PSS as a manufacturing business model and/or strategy for manufacturing companies pursuing competitiveness, customer satisfaction as well as sustainable development [12].

B. *Cloud Manufacturing*

Cloud Manufacturing (CMfg) is a new paradigm where manufacturing resources and capabilities are virtualized as services available in the cloud to users. This concept was firstly proposed by [13] with the intent to transform the manufacturing business into a new paradigm where manufacturing resources (i.e. physical devices, machines, products, processes, etc.) are transformed into cloud entities. This transformation is also called virtualization and enables for full sharing and circulation of virtualized resources that are capable of providing fundamental information about their own status. This information can potentially be used for local and global optimisation of the whole lifecycle of manufacturing.

C. *Service-Oriented Architecture*

Service-Oriented Architecture (SOA) paradigm has emerged and rapidly grown as a standard solution for publishing and accessing information in an increasingly Internet-ubiquitous world. SOA defines an architectural model aiming to enhance efficiency, interoperability, agility and productivity of an enterprise by positioning services as the building blocks through which solution logic is represented in support of realization of strategic goals [14]. The existence of Web Services technology has enabled and stimulated the implementation and development of SOAs. The application of SOA and Web Services in the context of manufacturing layer is still scarce, since a set of persisting technical challenges exists as pointed in [15]. SOA and Web Service are considered promising techniques for integrating all the existing layers within a manufacturing enterprise spacing from business to the physical process. The capability of encapsulating functions and tools as services through standard interfaces and protocols, enables their access and usage by clients without the need to know and control their specific implementations. All these aspects promote the SOA paradigm and its most used implementing technology (Web Services) as the de-facto standards for fast, secure and, above all, easy integration of any new functionality within existing software solutions while electing them as one of the pillars for implementing the CMfg paradigma as also confirmed in [16].

D. *Service Composition*

Services are the building block of a SOA, they provide simple interactions between client and server provider. However, sometimes atomic services need to be straightforwardly combined and/or assembled in order to generate more complex ones rising the service abstraction as referred by [17]. In this scenario as argued in [18], the term service composition is referred to the process of developing a composite service. Moreover, a composite service can be defined as the service that is obtained by the composition of the functionalities of several simplest services. Currently in the domain of SOA-based systems, two main approaches can be used for the service composition, namely [19]: orchestration and choreography. As stated in [20], although there is an available assortment of standards for web services orchestration, the most employed are consistently Web Services Business Process Execution Language (BPEL) [21] and Business Process Modelling Notation (BPMN) [22]. Actually the latter is becoming wider popular than the former as also confirmed by several solutions that adopt it [23].

E. Ambient Intelligence and Context Awareness

Ambient Intelligence (AmI) is about sensitive and thus adaptive electronic environments that actively interact with people to fulfill their needs [24]. However, as pointed in [25], people are still far from being immersed in the envisioned scenarios because of two main factors, namely: digital divide, and security and privacy threats. Even if AmI research has not generated the expected result some concepts and principles that are relevant whenever it is necessary to support human in complex and intricate tasks. AmI is strictly related to the design and development of context awareness applications as confirmed in [26]: *"Ambient intelligence a new paradigm of information and communication technologies […] to realize context-aware environments that are sensitive and responsive to the presence of people"*. Context awareness is widely applied in modern ICT solutions for developing pervasive computing applications that are characterized by flexibility, adaptability and are capable of acting autonomously [27]. As exposed in [28], a system is context-aware if it can extract, interpret and use context information to adapt its functionalities and behaviour to the current context. The development of such kind of applications is inherently complex since typically context information is gathered from a variety of sources that differ in quality of information they produce and are usually failure prone [29]. According to [27], the complexity of engineering context-aware applications can be reduced solely using infrastructure capable of gathering, managing and provisioning context information to the applications that require it.

F. Data Mining in Manufacturing

The process of extracting/discovering knowledge from large quantities of data is also known as data mining (DM) [30], [31]. DM can be defined as the process that starting from apparently unstructured data tries to extract knowledge and/or unknown interesting patterns [32]. During this process machine learning algorithms are used.The discovered knowledge can be used for classification tasks, modeling tasks, and to make prediction about future evolution of the analyzed variables. As stated in [33], the application of data mining techniques is quiet old. As a matter of fact, DM has been used and successfully applied in several areas like banking, insurance, fraud detection, telecommunication data etc. for future strategy and planning [34]. However, there are areas where these techniques are not exhaustively explored such as manufacturing. In this scenario, the current proliferation of new technologies and paradigms and the consequent trends in connectivity (cars, home appliance, etc) is pushing the data availability to another level unreached before while empowering the possibility to collect data from products/processes during their lifecycle. The analysis of this data by using data mining techniques is the basis for gaining competitive advantage.

IV. Research Contribution and Innovation

The research motivation behind this work relates with the strategic objective of allowing the manufacturing companies to enter in a continuous process of upgrading their products along their life cycle in the direction of the Product Service System (PSS) model. A PSS is a function-oriented business model aimed at offering products enriched by services. In this scenario the opportunities to extend products (PES) can be enormously facilitated by the wider dissemination of intelligent devices and, thus, the proliferation of cyber-physical systems. However, these aspects alone are not sufficient if a comprehensive ICT infrastructure – to allow to fully take advantage of intelligent and connected devices – is not provided. Therefore, in line with the research question mentioned in section 1, the research statement that supports the current research work is:

"The extensions of products by services can be facilitated if a service-oriented cloud-based platform that provides a set of tools and services to support the data extraction, collection and analysis is available"

Thereby the presence of an ICT supporting environment is a necessary condition to enable and boost the extension of product and processes by services as also presented in [35]–[37], . Such technical environment is aimed at a lower level to gather information from any available resource (product/process) and at a higher level to provide novel mechanisms (such as AmI monitoring, context extraction, data mining and more in general data analysis) in a unified approach to facilitate the extraction of useful knowledge from data. Finally, the presence of the referred technical environment is only the key enabler for extension of products/processes by services and requires to be supported by a strong methodology for driving the users of ProSEco in the complex process of acquiring necessary data related to product/processes and how to transform such "raw data" in valuable and relevant knowledge that in turn will be used as foundation for the creation of new services associated to the selected product/process.

V. The ProSEco Architecture

The overall aim of the ProSEco project is to implement a novel Cloud-enabled and extensible platform for collaborative design of product/production processes extended by services. To do this a SOA-based software architecture – that describes the ProSEco solution/system and its internal components/modules – has been designed and implemented during the project (see Fig. 2).

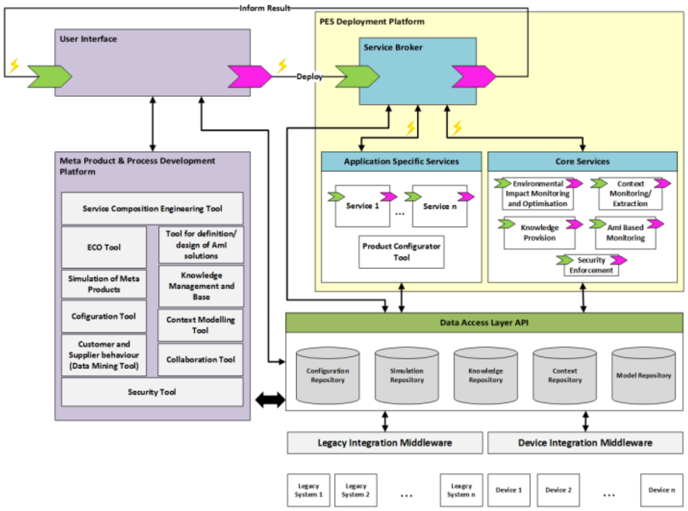


Fig. 2. Collaborative Environment for Product-services design and deployment of PES [48]

The ProSEco solution/system is constituted by two main platform the Meta Product & Process Development platform and the PES Deployment platform. Both the platforms relay on a backbone infrastructure that enables the development, provisioning, and deployment of Product Extensions Services (PES) solutions around the products and their production processes. In particular the Meta Product & Process Development platform provides all the necessary mechanisms to allow the users (PES designers) of the ProSEco system/solution to design/configure their own PES associated to a certain product/process as well as to simulate the meta product offerings in dynamic business ecosystem in order to explore and test the effects of alternative offering designs. In this case a set of engineering tools are provided and are used during these activities. Therefore, the main purpose of the engineering tools could be categorized as in Fig.3. Tools such as simulation of meta-products and user behavior analysis can be used by the industry partners to test the market viability of a new product as well as gathering intelligence from user behavior data to un-derstand how their designs could be further enhanced to be eco-friendly as well as customer friendly. Tools such as context modeling, design of AMI solutions, security and service configuration could be used by the partners to design PES to provide better meta-products and services. On the other hand, tools such as knowledge management, eco-design rules and metrics using LCA techniques and collaboration environments provide the industry partners with capabilities to manage their knowledge share their knowledge and collaborate with range of partners to design innovative PES.

On the other hand, the PES Deployment platform provides all the necessary mechanisms to assure the execution of a PES. In this scenario, several core software modules/functionalities are provided to allow the extraction/gathering of data from the environment, the processing/analysis of the extracted data and the provisioning of the results to the user (the data generated by the ProSEco solution/system is the basis for competitive advantage). According to Fig. 2 the PES Deployment platform is constituted by the following modules:

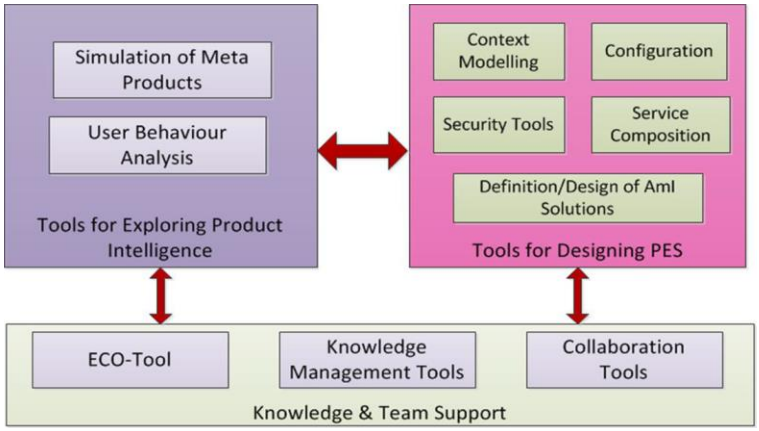


Fig. 3. Categorisation of Engineering tools [48]

- *Service Broker*;
- *Application Specific & Core Services*
- *Data Access Layer*;
- *Integration Layer;*
- *and Security Enforcement.*

Finally, the PES Collaborative Development platform provides the tools for configuring all the necessary elements of the PES Deployment platform whenever a new PES solution needs to be implemented. More details about the components of the ProSEco solution/system can be found in [38] and in [39].

A. *Service Broker*

The *Service Broker* module/component is responsible to execute a PES solution designed by using the Meta Product & Process Development platform. It is responsible to identify the application specifc and core services that compose a PES solution and to orchestrate them according to what has been specified in the the Meta Product & Process Development platform and in particular by the service composition engineering tool.

B. *Application Specific & Core Services*

The *Application Specific & Core Services* are atomic functionalities that are provided by the ProSEco system and can be used for designing PES solutions. The *Application Specifc Services* represent functionalities that are specific for the particular application scenario. On the other hand the *Core Services* represent core functionality that are generic enough to be used and applied in all the application scenarios. Therefore, a PES solution is basically composed by a set of linked *Application Specific & Core Services* and their configurations.

C. *Data Access Layer*

The *Data Access Layer* is responsible to separate the business logic from the knowledge storage. This layer is intended to handle every access (read and/or write) to the repositories. Furthermore, it is responsible to store all the knowledge generated during the ProSEco system runtime and provide this knowledge to other external applications/systems that want use it. In such a way, it facilitates the smoothly integration of the ProSEco system in already existent and deployed IT solution inside a manufacturing company.

D. *Integration Layer*

The *Integration layer* comprises the Legacy Integration Middleware and the Device integration Middleware. The former provides wrapper services to integrate legacy systems into the ProSEco architecture. The latter provides an infrastructure to integrate AmI and other devices (intelligent devices) into the platform.

E. *Security Enforcement*

The *Security Enforcement* module/component is responsible for controlling the access to the system as well as guaranteeing the integrity and confidentiality of data, and the availability of the system to perform its primary functionality.

VI. APPLICATION SCENARIO

To validate the current proposal, four business cases from four industrial partners drove the current research work. Each business case is constituted by several application scenarios extracted form concrete/typical situations allowing experimental validation of the ProSEco platform in order to assure that the proposed solution and cloud-based infrastructure as well as the methodology is generic enough and valid to be applied in distinct industrial environments. Thereby, the objective is to use the developed solution to exploit cyber physical features of modern products and processes for extending them by generating services.

In this paper the focus was put over one of the existing application scenarios to validate current infrastructure supported by early test results.

A. *Modelling of the Consumer Behavior*

The purpose of the application scenario from Electrolux grounds on the continuous monitoring of the home appliances as a necessary condition to enable the modelling of the consumer behaviour during the home appliance lifecycle (see Fig. 4).

The monitoring process is done in completely anonymous way, without collecting sensitive data, to guarantee the full respect of privacy. In this context, Electrolux intends to deeply use the ProSEco solution for studying advanced post sale services for its customers.

In particular Electrolux plans to apply the ProSEco ICT infrastructure and the entire set of engineering tools for extracting and collecting data from the home appliances during their normal use, with the purpose of analyzing them to collect information capable to improve the User-(i.e. misuse or not efficient use of the home appliance) and, in general, the product performances. The possibility of running this type of User-behaviour analysis, based on a large scale, with different level of clustering (e.g. ethnographic, age-based …) can provide a big amount of feedbacks that was not possible to collect with simple physical products.

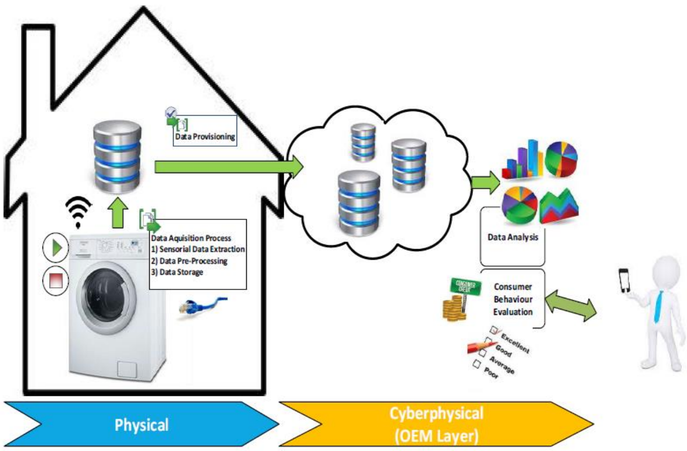


Fig. 4. The concept of the consumer behaviour application scenario

The following functionalities provided by the ProSEco solution are planned to be used during this application scenario:

1) *Meta Product & Process Development platform:*

a) AmI Monitoring Engineering Tool: to configure the sensorial information to be extracted from the home appliance;

b) Context Modelling Engineering Tool: to model the contextual information to be used for extracting and identifying current context from sensorial data;

c) Data Mining Engineering Tool: to define the data source and the machine learning algorithm to be applied.

2) *PES Deployment Platform & Service Composition:*

a) AmI Monitoring Core Service: to extract and collect sensorial data from the environment;

b) Context Extraction Core Service: to extract the context from sensorial data;

c) Data Mining Core Service: to analyze the data provided by sensors plus the context enrichment with the goal of modelling the behavior of the consumers.

B. *Preliminary Experimental Results*

To validate the fitness of the proposed approach offline data about some home appliances – provided by Electrolux – has been used. The data has been recorded over several months from 85 refrigerators installed to some pilot customers in the US. The data is stored as a pair (Timestamp, Door_Close), from which atomic information has been derived, such as day of the week (DW), Date and Hour, as depicted in Table I.

A preliminary inspection of the data, described in depth in [49], shows that the values are consistent (e.g. the number of records with Door_Close = TRUE is equal with the number of records with Door_Close = FALSE). Therefore for the experiments only half of the record may be needed in most of the cases.

The objective was to feed the ProSEco solution/system with the provided data in order to test and understand the capability of the system to find patterns and correlation between the data extracted and the behavior of the consumer.

TABLE I. A SNAPSHOT OF THE DATA SOURCE AND FORMAT USED DURING THE EXPERIMENT

Data extracted from home appliance				
Timestamp	***DW***	***Date***	***Hour***	***Door_Close***
15.11.2013 00:28	***6***	***15.11.2013***	***0***	***TRUE***
15.11.2013 00:28	***6***	***15.11.2013***	***0***	***FALSE***
15.11.2013 01:11	***6***	***15.11.2013***	***1***	***TRUE***
15.11.2013 01:11	***6***	***15.11.2013***	***1***	***FALSE***
15.11.2013 01:40	***6***	***15.11.2013***	***1***	***TRUE***
15.11.2013 01:41	***6***	***15.11.2013***	***1***	***FALSE***
15.11.2013 01:44	***6***	***15.11.2013***	***1***	***TRUE***

For starting to explore the data, some preliminary analysis has been carried out and the results are depicted in the three chart of Fig. 5, Fig. 6, and Fig. 7.

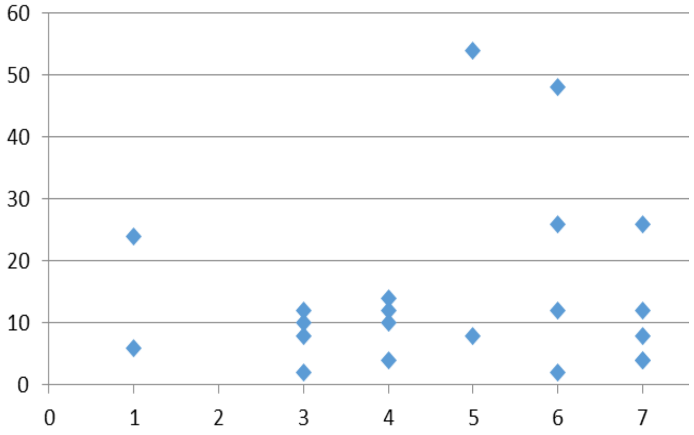


Fig. 5. Results of the preliminary analysis: DW vs Hours of openings

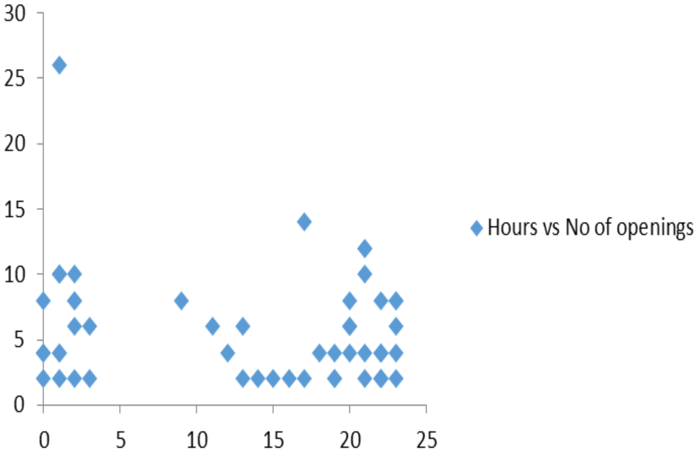


Fig. 6. Results of the preliminary analysis: Hours vs No of openings

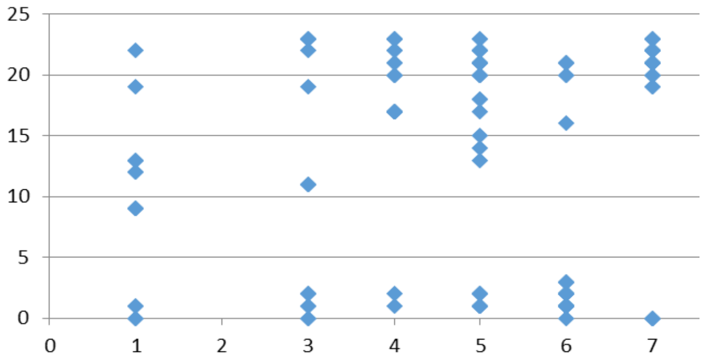


Fig. 7. Results of the preliminary analysis: DW vs No of openings

The chart of the Fig. 5 draws the number of opening vs. the days of the week (1 = Monday… 7 = Sunday). The statistical correlation factor is 0.1097, which shows that there is no significant statistical correlation between the two. But that does not mean that the two data sets are not related in some other way. For instance, it is obvious that on Tuesday the oven is not used at all. The chart of the Fig. 6 displays the number of openings vs. the hours of the day. For relevance of the data, the time consists only of hours, not of minutes, this is all events between 2:00 and 2:59 are recorded on the hour 2. The correlation factor is -0.202 which says that there is no significant statistical correlation (although more than in the previous case – DW vs. number of openings). However, some conclusions can be drawn, such as that the oven is used over the night and is idle from 4:00 to 9:00. The chart of Fig. 7 represents the days of the week ((1 = Monday,…, 7 = Sunday)) vs. the hours of openings. Like in the previous chart, the time contains only the hours, not the minutes. The correlation factor is 0.0247, which means that there is no statistical relation between the two data sets. However, a close look at the plot shows that:

- The oven is not used on Tuesdays;
- The over is not used from 4:00 – to 9:00;
- The time when the oven starts being used goes later over the week:
    - on Mondays, the oven starts working around 9:00;
    - Tuesdays are off;
    - on Wednesdays: starts around 11:00;
    - on Thursdays: starts around 17:00;
    - on Fridays: starts around 13:00;
    - on Saturdays: starts around 16:00;
    - on Sundays: starts around 19:00.

C. Data Mining Results

The previous section shows obvious relationships between openings and days (of the week). However, the next step is to find them out. To do that three sets of experiments have been carried out, namely:

- determining the number of openings per day;
- determining the next hour when the appliance is opened;
- and determining the interval when the appliance is not used at all.

The experiments have been conducted by using the RapidMiner[1] data mining tool. The applied methodology is similar to the one reported in [50].

1) Determining the number of openings per day

This experiment aims to determine how many times the appliance is opened based on the number of openings in the previous days. This would be a good statistical indicator about the usage of the appliance itself. The time window varies from 1 day to 5 days, but only the best results are reported here. The final results of the experiment are summarized in Table II.

2) Determining the next hour whene the appliance is opened

This experiment is intended to determine when the appliance is used again, having the previous openings. This is a difficult task as no clear distribution of the usage can be found over the days (see Fig. 5,6 and 7). However, if possible, this is the best indicator. The results are summarized in Table III.

3) Determining the next hour whene the appliance is opened

As can be seen in Fig. 3, there is a band over the week when the device is not used at all. The process described here aims to find the bandwidth (minimum and maximum hour). The best results are summarized in Table IV.

TABLE II. ACCUREACY OF THE DATA MINING ALGORITHM FOR DETERMINING THE NUMBER OF USAGES

Experiment 1		
Algorithm	**Configuration**	**Accuracy**
Local Polynomial Regression	Window size = 4 Degree = 5 Ridge factor = 10^{-6} Numerical measure = Manhattan distance Neighborhood type = Relative number Relative size = 0.1 Smoothing kernel = Gaussian	45.7%
Support vector machine (SVM)	Window size = 7 Kernel type = polynomial Kernel degree = 10 Kernel cache = 200 C = 0 Convergence epsilon = 0.01	54.3%
Neural networks (multi-layer perceptron)	Window size = 7 Hidden layers = 0 Training cycles = 500 Learning rate = 0.5 Momentum = 0.3	56.8%

TABLE III. ACCURACY OF THE DATA MINING ALGORITHM FOR DETERMINING THE NEXT OPENING HOUR

Experiment 2		
Algorithm	**Configuration**	**Accuracy**
Local Polynomial Regression	Window size = 5 Degree = 2 Ridge factor = 10-9 Numerical measure = Euclidean distance Neighborhood type = minimum distance Distance = 10 Minimum distance = 20 Smoothing kernel = exponential	20.8%

k-Nearest Neighbor	Window size = 10
	k = 5
	Numerical measure = Bregman divergency
	Divergence = generalized divergence
		17.0%

TABLE IV. ACCURACY OF THE DATA MINING ALGIRITHM FOR DETERMINIG THE BANDWITH OF THE NON USAGE HUORS

Experiment 3		
Algorithm	**Configuration**	**Accuracy**
Local Polynomial Regression	Window size = 7 Degree = 2 Ridge factor = 1 Numerical measure = Correlation similarity Neighborhood type = relative number Relative size = 0.5 Smoothing kernel = Gaussian	68.8%
Neural networks (multi-layer perceptron)	Window size = 7 Hidden layers = 0 Training cycles = 500 Lear Momentum = 0.2ning rate = 0.6	62.5%
Support vector machine (SVM)	Window size = 7 Kernel type = polynomial Kernel degree = 15 Kernel cache = 200 C = 0 Convergence epsilon = 0.001	68.8%

Evolutionary SVM	Window size = 7
	Kernel type = dot
	C = 0
	Epsilon = 0.1
		56.2%

k-Nearest Neighbor	Window size = 1 k = 15 Numerical measure = Correlation similarity	68.2%

4) Wrap-up

Finally, comparing the three experiments it is obvious that the most difficult task is the determination of the next opening hour. is the most difficult task and the reason is because of the hazardous distribution of the usage over the days, respectively over the week. Predicting the number of openings per day leads to unsatisfactory results. An accuracy of 56.8% is just a little more than 50%. However, finding the bandwidth brings to the best results (68.8%) and this is just a starting point for further research. A summary of the results is presented in Table V.

[1] https://rapidminer.com/

TABLE V. SUMMARY OF THE APPROACHES USED DURING THE THREE EXPERIMENTS

Summary of the Experiments			
Approch	**Number of worthy algorithms**	**Best Accuracy**	**Average Accuracy**
determining the number of openings per day	3	56.8%	52.3%
determining the next hour when the appliance is opened	2	20.8%	18.9%
determining the interval when the appliance is not used at all	5	68.8%	64.9%

VII. CONCLUSIONS

Current work presents the first implementations of a supporting infrastructure to design PES solutions following the PSS strategy. The proposed solution addresses the extraction and collection of relevant data about the product, the enrichment of data with context and the analysis of the enriched data by suing data mining techniques and statistics. Thereby result of such analysis provide the basis for the implementation of highly personalized advanced services (i.e. maintenance services, post sell, etc.).

The application scenario shows the potential of the proposed solution. The data recorded from the field refrigerators is consistent and relevant, although there is no significant statistical correlation between the datasets compared. However, some heuristics can be drawn, such as:

- The oven is not used on Tuesdays;
- The over is not used at all from 4:00 – to 9:00;
- The time when the oven starts being used goes later over the week.

The data processed covers only seven weeks, which is a reasonable interval for a preliminary analysis. However, for a very thorough research, the data should be gathered over a longer period. Further research will focus also on the utilization of real connected appliances as data source (see Fig. 8). In this case high complexity of data acquisition and real-time data analysis algorithms will be addressed in further research to “fully” utilize the opportunities offered by the proposed solution. In particular, the research presented in this paper is focusing on using the data extracted from home appliances for analyzing and modelling the consumer behavior. However, this is not the only purpose of the proposed research. The next step will be the extraction and analysis of the data to model the behavior of the hardware components installed on the home appliance in order to predict failures and to gather fundamental knowledge that allows the paradigm shift from run-to-failure maintenance to the proactive maintenance strategy. Finally, as future works the proposed reaserch will follow two main paths the *consumer behavior* and the *component behavior* where data mining processes will be adapted and configured for the specific objective of the analysis.

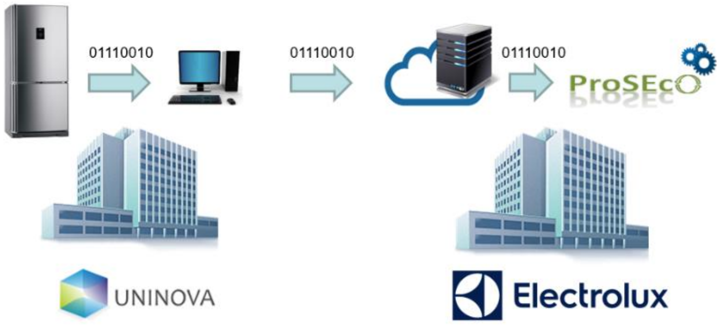


Fig. 8. New experimental setup for further research

ACKNOWLEDGMENT

This work is partly supported by the ProSEco (Collaborative Environment for Eco-Design of Product-Services and Production Processes Integrating Highly Personalized Innovative Functions) project of European Union’s 7th Framework Program, under the grant agreement no. NMP2-LA-2013-609143. This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

REFERENCES

[1] U. européenne D. générale de la recherche, Factories of the Future: Multi-annual Roadmap for the Contractual PPP Under Horizon 2020. Publications office of the European Union, 2013.

[2] A. Gustafsson, S. Brax, L. Witell, G. Lay, G. Copani, A. Jäger, and S. Biege, “The relevance of service in European manufacturing industries,” J. Serv. Manag., vol. 21, no. 5, pp. 715–726, 2010.

[3] S. Vandermerwe and J. Rada, “Servitization of business: Adding value by adding services,” Eur. Manag. J., vol. 6, no. 4, pp. 314–324, 1988.

[4] R. Roy, E. Shehab, A. Tiwari, T. Baines, H. Lightfoot, O. Benedettini, and J. Kay, “The servitization of manufacturing: A review of literature and reflection on future challenges,” J. Manuf. Technol. Manag., vol. 20, no. 5, pp. 547–567, 2009.

[5] F. Jovane, E. Westkämper, and D. J. Williams, The ManuFuture Road: Towards Competitive and Sustainable High-adding-value Manufacturing. Springer, 2009.

[6] S. Wiesner, M. Freitag, I. Westphal, and K.-D. Thoben, “Interactions between Service and Product Lifecycle Management,” Procedia CIRP, vol. 30, pp. 36–41, 2015.

[7] M. Lobonțiu and A. Petrovan, “Product innovation & preparation for technological diffusion of equipment,” in The 6th International Conference on Management of Technological Changes, 2009, vol. 2, pp. 689–692.

[8] C. Sassanelli, G. Pezzotta, M. Rossi, S. Terzi, and S. Cavalieri, “Towards a Lean Product Service Systems (PSS) Design: State of the Art, Opportunities and Challenges,” Procedia CIRP, vol. 30, pp. 191–196, 2015.

[9] “Product Service systems, Ecological and Economic Basics.” [Online]. Available: http://docplayer.net/334668-Product-service-systems-ecological-and-economic-basics.html. [Accessed: 01-Nov-2015].

[10] T. S. Baines, H. W. Lightfoot, S. Evans, A. Neely, R. Greenough, J. Peppard, R. Roy, E. Shehab, A. Braganza, A. Tiwari, J. R. Alcock, J. P. Angus, M. Bastl, A. Cousens, P. Irving, M. Johnson, J. Kingston, H. Lockett, V. Martinez, P. Michele, D. Tranfield, I. M. Walton, and H. Wilson, “State-of-the-art in product-service systems,” Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., vol. 221, no. 10, pp. 1543–1552, Oct. 2007.

[11] E. Manzini and C. Vezzoli, “A strategic design approach to develop sustainable product service systems: examples taken from the ‘environmentally friendly innovation’ Italian prize,” J. Clean. Prod., vol. 11, no. 8, pp. 851–857, Dec. 2003.

[12] A. Tukker and U. Tischner, “Product-services as a research field: past, present and future. Reflections from a decade of research,” J. Clean. Prod., vol. 14, no. 17, pp. 1552–1556, 2006.

[13] B.-H. Li, L. Zhang, S.-L. Wang, F. Tao, J. Cao, X. Jiang, X. Song, and X. Chai, “Cloud manufacturing: a new service-oriented networked manufacturing model,” Comput. Integr. Manuf. Syst., vol. 16, no. 1, pp. 1–7, 2010.

[14] T. Erl, Service-Oriented Architecture: Concepts, Technology, and Design. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2005.

[15] L. Ribeiro, G. Candido, J. Barata, S. Schuetz, and A. Hofmann, “IT support of mechatronic networks: A brief survey,” in Industrial Electronics (ISIE), 2011 IEEE International Symposium on, 2011, pp. 1791 –1796.

[16] J. Jassbi, G. Di Orio, D. Barata, and J. Barata, “The Impact of Cloud Manufacturing on Supply Chain Agility,” presented at the 12th IEEE International Conference on Industrial Informatics (INDIN’14), Brazil, 2014.

[17] N. Kavantzas, D. Burdett, G. Ritzinger, T. Fletcher, Y. Lafon, and C. Barreto, “Web services choreography description language version 1.0,” W3C Candidate Recomm., vol. 9, 2005.

[18] Hazenberg, Meta Products. Building the Internet of Things. Uitgeverij Bis, 2011.

[19] K. Ashton, “That ‘internet of things’ thing,” RFiD J., vol. 22, pp. 97–114, 2009.

[20] G. M. Cândido, “Service-oriented architecture for device lifecycle support in industrial automation,” 2013.

[21] OASIS, “Web Services Business Process Execution Language.” 2007.

[22] S. A. White, “Introduction to BPMN.”

[23] M. Kalauny, “BPM software application with the Industrial Internet of Things (IIoT),” Schneider Electric Blog. [Online]. Available: http://blog.schneider-electric.com/telecommunications/2015/08/07/bpm-software-application-industrial-internet-things-iiot/. [Accessed: 30-Oct-2015].

[24] E. Aarts and R. Wichert, “Ambient intelligence,” in Technology Guide, H.-J. Bullinger, Ed. Springer Berlin Heidelberg, 2009, pp. 244–249.

[25] L. Ribeiro, J. Barata, and P. Barreira, “Is ambient intelligence a truly human-centric paradigm in industry? Current research and application scenario,” Nov. 2009.

[26] R. J. Jiao, Q. Xu, J. Du, Y. Zhang, M. Helander, H. M. Khalid, P. Helo, and C. Ni, “Analytical affective design with ambient intelligence for mass customization and personalization,” Int. J. Flex. Manuf. Syst., vol. 19, no. 4, pp. 570–595, Feb. 2008.

[27] K. Henricksen, J. Indulska, and A. Rakotonirainy, “Modeling context information in pervasive computing systems,” in Pervasive Computing, Springer, 2002, pp. 167–180.

[28] H. E. Byun and K. Cheverst, “Utilizing context history to provide dynamic adaptations,” 2010.

[29] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni, “A survey of context modelling and reasoning techniques,” Pervasive Mob. Comput., vol. 6, no. 2, pp. 161–180, Apr. 2010.

[30] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Mag., vol. 17, no. 3, p. 37, 1996.

[31] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, 2011.

[32] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, 2014.

[33] G. Di Orio, “Adapter module for self-learning production systems,” FCT-UNL, 2013.

[34] M. Shahbaz, S. A. Masood, M. Shaheen, and A. Khan, “Data mining methodology in perspective of manufacturing databases,” J Am Sci, 2010.

[35] G. Candido, C. Sousa, G. Di Orio, J. Barata, and A. W. Colombo, “Enhancing device exchange agility in Service-oriented industrial automation,” in 2013 IEEE International Symposium on Industrial Electronics (ISIE), 2013, pp. 1–6.

[36] G. Di Orio, G. Cândido, and J. Barata, “The Adapter module: A building block for Self-Learning Production Systems,” Robot. Comput.-Integr. Manuf., vol. 36, pp. 25–35, Dec. 2015.

[37] G. D. Orio, D. Barata, A. Rocha, and J. Barata, “A Cloud-Based Infrastructure to Support Manufacturing Resources Composition,” in Technological Innovation for Cloud-Based Engineering Systems, L. M. Camarinha-Matos, T. A. Baldissera, G. D. Orio, and F. Marques, Eds. Springer International Publishing, 2015, pp. 82–89.

[38] D. Stokic, S. Scholze, C. Decker, and K. Stobener, “Engineering methods and tools for collaborative development of industrial cyber-physical based products and services,” in 2014 12th IEEE International Conference on Industrial Informatics (INDIN), 2014, pp. 594–599.

[39] ProSEco Consortium, “D1.3 - ProSEco Concept.” 2014.

On Shear Wave Speed Estimation for Agar-Gelatine Phantom

Hassan M. Ahmed[1, 2*,] Nancy M. Salem[1,] Ahmed F. Seddik[1, 2,] Mohamed I. El Adawy[1]

[1]Biomedical Engineering Department
Helwan University
Cairo, Egypt
[2]Faculty of Computer Science
Nahda University
Beni Suef, Egypt

Abstract—**Conventional imaging of diagnostic ultrasound is widely used. Although it makes the differences in the soft tissues echogenicities' apparent and clear, it fails in describing and estimating the soft tissue mechanical properties. It cannot portray their mechanical properties, such as the elasticity and stiffness. Estimating the mechanical properties increases chances of the identification of lesions or any pathological changes. Physicians are now characterizing the tissue's mechanical properties as diagnostic metrics. Estimating the tissue's mechanical properties is achieved by applying a force on the tissue and calculating the resulted shear wave speed. Due to the difficulty of calculating the shear wave speed precisely inside the tissue, it is estimated by analyzing ultrasound images of the tissue at a very high frame rate. In this paper, the shear wave speed is estimated using finite element analysis. A model is constructed to simulate the tissue's mechanical properties. For a generalized soft tissue model, Agar-gelatine model is used because it has properties similar to that of the soft tissue. A point force is applied at the center of the proposed model. As a result of this force, a deformation is caused. Peak displacements are tracked along the lateral dimension of the model for estimating the shear wave speed of the propagating wave using the Time-To-Peak displacement (TTP) method. Experimental results have shown that the estimated speed of the shear wave is 5.2 m/sec. The speed value is calculated according to shear wave speed equation equals about 5.7 m/sec; this means that our speed estimation system's accuracy is about 91 %, which is reasonable shear wave speed estimation accuracy with a less computational power compared to other tracking methods**.

Keywords—Elasticity Imaging; Acoustic radiation force impulse (ARFI); Shear wave elasticity imaging; Soft tissue stiffness imaging

I. INTRODUCTION

Replacing healthy soft tissues by fibrotic tissues is the pathological change that may cause a malignant or benign tumor. The stiffness of these pathologic tissues is higher than the surrounding [1]. Elastic modulus is a measurand for the stiffness. It is the measure of the material's resistance to deformation in either compression or tension namely the elasticity modulus (E) and in shear namely the shear modulus (μ) [2]. Muscles and fibrous tissue are more resistant to deformation than other compliant tissues such as fat, due to their higher elastic moduli [3-5].

Deformation; is a result of stress over the tissue; formerly, it was manual palpation over the tissue. Nowadays, it is the beat or the push that is caused by the acoustic radiation force generated by the ultrasound probe, a procedure called Elastography procedure. This procedure can be accomplished by many methods. These methods are classified regarding either the obtained images or the source of excitation.

Regarding the obtained images, these methods are either qualitative; revealing relative stiffness differences; or quantitative; leading to an estimate for the underlying tissue elastic modulus using reconstruction methods. Also, classification regarding the source of excitation these methods are either external; which are the dynamic external methods and static external compression methods for mechanical excitation; or internal; like the physiological motion of the tissue itself or an acoustic radiation force (ARF).

In this paper, a model is constructed simulating the soft tissue's mechanical properties. The model simulated the Agar-gelatine phantom, due to its similarity in behavior to soft tissue, where the agar is considered to act as scatterers and gelatin is considered to introduce the elasticity for the model.

A transient ARF is applied to the model to induce scatterers displacements. Then, peak displacements are tracked using Time-To-Peak displacement (TTP) method along the lateral direction starting from the focal zone and moving away towards either of the lateral edges. Speed; of the shear wave generated by the propagation of the peak displacements; is then estimated by calculating the time taken by the wave to reach a specific point on the lateral direction divided by the distance.

This paper is organized as follows, firstly an introduction section about elastography and its major methods, and a background section about the elastography mathematics. Then a literature review section; about the progression in the field of elastography applications starting from its origin till the recent application and the major imaging techniques; is presented. A methodology section demonstrating the FEM generation and ARFI application and acquisition procedure is introduced. Finally, results, discussion, and conclusion sections are given.

II. Background

This section aims to give the reader a proper background about the mathematical equations involved in the field of elastography, the assumptions made to simplify the calculations and the limitations that are found if the tissue is modeled as viscoelastic and non-linear.

Firstly, it is worth speaking about the strain, which is the deformation caused by the stress, and to highlight its behavior in the finite element models and its relationship with displacement induced. Strain (ϵ) is related to the tissue displacement by (1) [1, 2], and shown in Figure 1.

$$\in = (\tfrac{1}{2})(\nabla u^{T} + \nabla \mathrm{u}) \qquad (1)$$

where ∇u is the spatial displacement gradient, and T is the transpose operation.

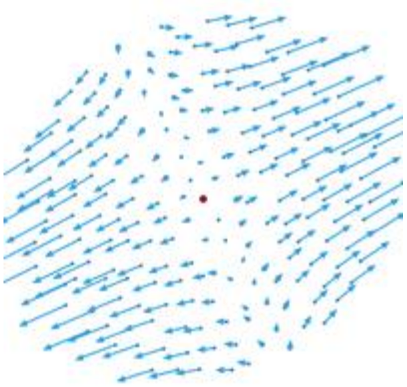


Fig. 1. Spatial displacement gradient

The dynamic displacement response of the soft tissue can be tracked by using the cross-correlation and Doppler-based auto-correlation; such as Kasai's method [6]. The resolution of the tracking methods is not fixed along the lateral and the axial directions. It is better in the axial direction (it is actually fractions of a micrometer) than in the lateral one (a few tens of micrometers) [7]. Axial direction is chosen, because it is the direction of better resolution. Another issue that should be considered is the homogeneity of soft tissues whereas materials constituting them are not homogeneous. Many assumptions are proposed in that field (elasticity imaging field) to simplify the interpretation and the analysis of images obtained. The most common assumptions for the tissue material are [1, 2]:

a) Stress-strain relationship of the tissue is linear.

b) The tissue is elastic.

c) The tissue is isotropic.

Under these assumptions we conclude the relation between stress (σ) and strain (ϵ) to be given by (2) [2].

$$\sigma = E\,\epsilon \qquad (2)$$

where E is the elasticity modulus in KPa.

Another way to estimate the tissue's elastic properties is by tracking the propagation of the shear waves generated inside the tissue. Shear waves propagate orthogonally to the direction of the propagation of the ultrasonic compressive waves; i.e. perpendicular to the direction of the induced tissue displacement.

Under the preceding assumptions of the tissue material, the equation of the shear modulus is given by (3) [1].

$$\mu\nabla^2 u - \rho\left(\frac{\partial^2 u}{\partial t^2}\right) = 0 \qquad (3)$$

where ρ is the material density in Kg/m³, ∇² is the Laplacian operator and t is the time.

The speed of the propagation of the shear wave is also related to the shear modulus according to (4).

$$C_x = \sqrt{\frac{\mu}{\rho}} \qquad (4)$$

Moreover; the shear modulus is related to the elasticity modulus by (5) [1, 2].

$$\mu = \frac{E}{2(1+\upsilon)} \qquad (5)$$

where υ is the Poisson's ratio.

There are two deviations from the above assumptions when modeling the tissue as viscoelastic and non-linear:

a) Considering the viscosity of the tissue results in a dependence of the tissue stiffness on the excitation frequency. Higher excitation frequencies yield a stiffer tissue response compared to lower excitation frequencies. In other words; both the elastic and shear modulus is a function of the frequency (E (f) and μ (f)) [8]. A frequency dependent shear modulus will result in a frequency dependent shear wave speed, a phenomenon called Dispersion.

b) Furthermore, as a result of the introduction of the viscosity term, the tissue would absorb some of the imparted energy.

Nonlinearities imply that the strain in response to an applied stress is dependent on the absolute stress that is applied to the tissue; hence, the elastic moduli are a function of strain (E (ϵ) and μ (ϵ)) [1].

III. Literature Review

Elasticity imaging methods adopt one concept which is applying either a mechanical excitation or stress to tissues. Stress can be from external or internal excitation source. The internal source may be the physiological motion of the tissue itself or an acoustic radiation force (ARF). The resulting tissue deformation (displacement) is measured in response to that excitation using ultrasound, magnetic resonance or optical methods [1].

Based on the Helmholtz model for shear wave propagation; introduced by equations (3), (4) and (5); the measured tissue deformation can be related to tissue stiffness. When the imaging methods were first proposed, the excitation source was obtained and considered as the physiological pulsation of the tissue itself, and ultrasound was used to monitor the tissue response [9, 10]. Then, dynamic methods were introduced, where dynamic external vibration is used to create shear waves inside the tissue to be studied (Sonoelasticity) [11] and methods using external static compression for mechanical excitation (strain imaging) [12].

The excitation using Acoustic Radiation Force Impulse (ARFI) was introduced in the early 90s by Sugimoto *et al.* [13]. This method has an advantage of coupling the source of excitation to the organ under study directly, rather than being coupled through intervening tissues. Greenleaf *et al.* and Parker *et al.* provide efficient reviews of elasticity imaging methods [14, 15]. Nyborg *et al.* [16, 17] introduced a model of the tissue to be acting as a viscoelastic fluid in response to the ultrasonic wave propagation, and under plane wave assumptions, the ARFI is given by (6).

$$F = \frac{2\alpha I}{C} \tag{6}$$

where α (dB/cm.MHz) is the acoustic absorption coefficient of the tissue, I is the temporal average intensity of the wave and C (m/sec.) is the speed of sound in tissue.

Besides, the contribution of scattering is neglected because the majority of attenuation arises from absorption [17]. The relationship between the depth and the frequency should be considered. The attenuation increases by increasing depth for higher frequencies. As a result, there is an optimal frequency for each depth which depends on the attenuation-frequency tradeoff.

The ARFI can be applied for different temporal duration's methods, such as the quasi-statically, transient method, and harmonic method [1, 8]: the Quasi-static method which proposes that; the excitation pulses are applied on the tissue to reach a steady state response, typically longer than one second; the transient method which proposes that the excitation is applied for a very short duration, typically a temporal impulse, faster than the natural frequency of the tissue associated with the dynamic tissue response; and the Harmonic method which proposes the application of the impulse in a harmonic way; a pulsed manner; to achieve a sinusoidal tissue excitation of one or more frequencies.

Applying a specific stress (force) then measuring the corresponding mechanical response is considered as the most common method for elasticity imaging. By having variations in the duration of the applied force and different measuring methods, a variety of imaging methods have been obtained, each having advantages and drawbacks [18]. A brief about the most common methods is given in the following sub-section.

A. Sonoelasticity

Harmonic shear waves are generated mechanically by using external actuators in contact with the skin. Doppler or any other imaging technique is used for measurement of wave propagation [19].

Krouskop *et al.* induced the shear wave inside the muscle tissue of thigh using a motorized actuator placed on the medial side of the thigh. The ultrasound transducer was put on the lateral side of the thigh to measure the wave propagation using Doppler methods [19]. On the other hand, Lerner *et al.* proposed the use of the acoustic horn to generate the waves in the phantom and the use of colored Doppler for measurement [20]. Shear wave velocities distribution is obtained by Sono-elastography for phantoms, human prostates, and skeletal muscles [21].

B. Acoustic Radiation Force Impulse (ARFI) imaging

Acoustic radiation force impulse (ARFI) is defined as the force resulting from the momentum transfer from the propagating ultrasound wave through the tissue due to absorption and scattering mechanisms [16]. Displacements in tissue can be generated using a focused force impulse, these displacements (deformations) are relaxed and the tissue returns to its original position by the removal of the force [22].

In response to this focused force, the tissue within the region of excitation (ROE) is deformed and shear waves are generated and propagate away from the ROE. After this short duration excitation is achieved, an ordinary imaging procedure is done to image these waves departure along the transverse direction of the ROE. Along a single line, excitation is accomplished and measurement is made, then, another adjacent line is investigated and so on till the image of tissue response is constructed. The tissue response is characterized by a set of parameters; the peak displacement, and the time the tissue takes to reach the peak displacement and the recovery time [23].

This method has been used for many applications. It is used with phantom tissue imaging [24], imaging thermally induced lesions [25], abdominal imaging of lesions [26], human prostate imaging [27] and imaging of cardiovascular vessels [28, 29].

C. Transient Elastography

An external source of vibrations is used to generate waves and to provide a single cycle of low frequency; typically 40-50 Hz. Compressional and shear waves are generated together by this sort of excitation. Fortunately, they are separated from each other by the time lag between them, as the compressional wave is much faster than the shear one [30, 31]. Transient excitations avoid biases caused by the sinusoidal excitations of cylindrical source [30, 31].

Motion tracking is the most important part of this method, where cross-correlation is used to locate the time shift between two echo signals, and by knowing the speed of sound in tissue we can estimate how much motion has occurred inside the tissue.

Dutt *et al.* were pioneers to use mechanical actuation and to obtain a measurement of shear waves using this method [32]. Transient elastography can be used to measure stiffness in phantoms, breast, and skeletal muscle [33].

D. Shear Wave Elasticity Imaging

The modulated ultrasound beam that generates acoustic radiation force was not used until Nightengale and Trahey have made their experiments on the proposed theory of Sarvazyan *et al* [34]. The generated force, when applied to the tissue, generates the shear waves that are detected by any other method. It is used to palpate the tissue but from the inside. Thereby, it has replaced the physicians' finger and is used as a virtual finger. Having high localization of induced strain; due to high attenuation after a few micrometers away from the ROE; was the main difference between this method and any other elasticity imaging method.

The high absorption feature of these newly generated waves was the most important reason for the feasibility of using it. Then, an induced vibration is possible to be located in a very tiny portion of the tissue, namely in the focal zone.

In this technique, the major drawback is the small value of deformation. Hence, very complex signal processing methods are required to accurately estimate the motion [35-37]. This method has been used for investigation of phantoms, liver, prostate, and cardiac tissue [38, 39].

E. Supersonic Shear Imaging

In the preceding methods mentioned, the imaging procedure was made along a single line and with single focal point probably the focal point. The supersonic shear imaging involves investigation to be done with multiple focal points for the same line. The focal point is changed axially along the vertical beam with a speed much higher than the propagation speed of the shear wave resulted [40]. The multiple shear waves resulting from many focal locations (axially) constructively interfere to construct a conical shear wave [41].

This results in a Mach cone, where the Mach number of excitation can be adjusted to make the shear wave directionally oriented. It can be used for phantoms, liver, skeletal muscle and breast assessments [42].

It has been noticed that in viscous fluids, applying acoustic radiation force generates acoustic streaming or fluid flow, and this fluid flow has a velocity proportional to the fluid viscosity and the boundary condition. It is worthy to mention that phenomenon as it helped greatly in breast cancer detection and differentiation.

Starritt *et al.* were the first to investigate this phenomenon of generation of acoustic streaming. Nightingale *et al.* were the first to use it to differentiate between the fluid filled and solid lesions in the breast. This was achieved by interspersing pushing pulses with Doppler pulses to detect the resulting fluid flow using Doppler techniques [34].

IV. Methodology

In this search, a FEM is constructed and given the mechanical properties of the Agar-gelatine phantom by a Finite Element Modelling software LISA FEA V8.0. For estimating the mechanical properties of this phantom by ultrasound imaging of the shear wave, a point force is applied for about 0.01 seconds in a transient manner. The resulting shear wave is then tracked by B-mode ultrasound imaging at a high frame rate and its speed is estimated to calculate the stiffness of the phantom; which is a mechanical property of the phantom; using Matlab software version R2010a. Tracking the shear wave propagation off-axis and measuring its velocity is a necessary process to have a quantitative map of elasticity for the tissue under consideration.

Tracking process is performed at several stationary nodes inside the model; where every node gives a curve of its displacement versus time profile from which we estimate the speed of the propagated wave. During its propagation; the shear wave; covers a few meters per second, and a specific frame rate (FR) is needed to appropriately catch the propagating peak leading to a good estimation of the wave's speed. Hence, a frame rate of several kilohertz is needed to have a good estimation for the speed since the conventional ultrasound scanners are not efficient due to their low frame rates; typically reaching about 50 to 60 frames per second [43].

The block diagram of the proposed method is shown in Figure 2.

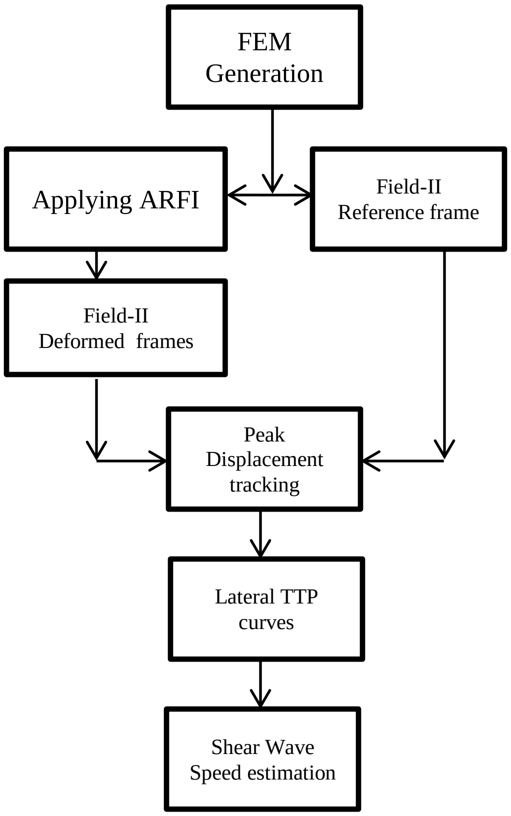


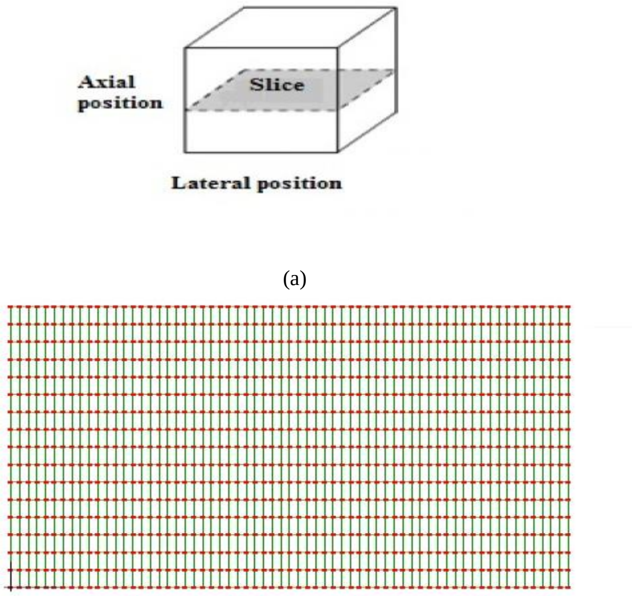
Fig. 2. Block diagram of the methodology

A. FEM Mesh generation

The experiment is performed using an elastic Agar-gelatine model. The choice of this material is due to the ability of the gelatine to maintain the stiffness of the phantom, and the ability of the agar to act as scatterers for the ultrasound waves in the phantom. A FEM mesh for this phantom is generated by a Finite Element Modelling (FEM) software LISA FEA V8.0 to simulate its behaviour at applying the acoustic radiation force impulse (ARFI). The FEM is a square shaped plate and unity in dimensions (1m side length), this square shaped plate resembles the plane inside which the ARFI is generated and the shear wave propagation takes place.

Dividing this model to small squares; by Meshing techniques; provides an accurate displacement calculations' leading to a good estimation of the shear wave speed, each sub-square is attached to its neighbourhood by nodes.

It is at these nodes where the wave tracking takes place, this is achieved by tracking the peak displacements using B-mode ultrasound imaging. Differences in times for reaching the peak displacement at two or more successive nodes are calculated to provide the speed estimation. The mesh consists of 1105 nodes and 1024 elements, where the opposing face to the transducer is constrained completely and the face where the transducer touches is allowed to move in the perpendicular direction; i.e. the direction of the shear wave and other faces are allowed to move freely in all directions. The distance between any two successive nodes in the lateral direction is 0.015625 meters. The phantom is shown Figure 3.



(a)

(b)

Fig. 3. Slice of the phantom, and (b) Square shaped 1m side length of the Agar-gelatine phantom

Assigning the material mechanical parameters of the Agar-gelatine phantom for the simulated model to act as a real material, those properties are listed in Table 1.

TABLE I. MATERIAL MECHANICAL PARAMETERS FOR AGAR-GELATINE PHANTOM

Material parameter	Value	Unit
Young's modulus (E_o)	107585	Pa
Shear modulus (G_o)	35886	Pa
Poisson's ratio (υ_o)	0.499	---
Density (ρ)	1060	Kg/m^3

B. Acoustic Radiation Force Impulse (ARFI) generation

After generating the mesh nodes, the ARFI is generated and applied on the model. A point source force is generated by LISA FEA V8.0 program and applied at the centre of the mesh model; equally at the focal point of the ultrasound beam in the axial direction; this force is applied for a very short duration, typically 0.02 second and less. The force is applied in a gradually increasing manner, and is removed in a gradually decreasing manner as well; each part of the applying and the removal of the force take typically 0.01 second; as shown clearly in Figure 4.

The time of investigation is 50 milliseconds; the force is induced within 10 milliseconds to reach its peak, and another 10 milliseconds to be fully removed and to reach zero magnitude, summing up for 20 milliseconds of excitation. The rest of the time is for the tracking B-mode imaging procedure.

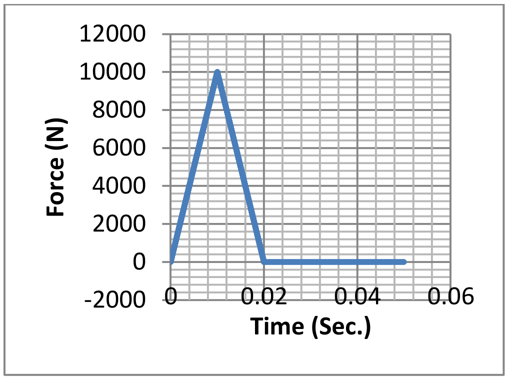


Fig. 4. Force magnitude versus the time of excitation

C. Acquisition Sequence

In the beginning, the medium is investigated by using a plane wave. This results in providing a reference frame. Then, the pushing sequence is sent by focusing the ultrasound beam to the area of interest (the focal zone of the beam is the area of interest) for a very short duration, typically about 20 milliseconds.

Just after the generation of the pushing pulse, an ordinary B-mode imaging procedure is carried out to catch the progress of the wave propagation through the model [43]. The B-mode imaging of the shear wave propagation is carried out by a very high frame rate, almost 1000 frames per second and up [43]. For investigating any other region in the model, the last procedure is repeated after modifying the focal zone of the beam.

For visualization of the propagation of the wave, the peak of the wave is plotted versus time profile for the whole model. The speed can be estimated from this visualization. Converting 3D visualizations to 2D curves of nodes displacements versus time profiles is more efficient for estimating the speed of the shear wave as proposed in [44]. For more feasibility of the calculations, one of the sides of the curve is selected to calculate the shear wave speed from it, by tracking the peak displacement. This is proposed by Ned C. Rouze in [45] and shown in Figure 5.

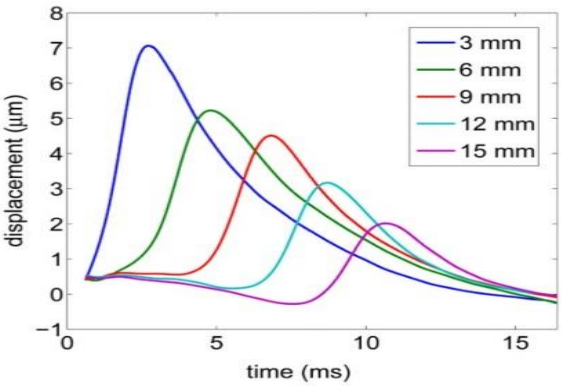


Fig. 5. Single sided curve for estimating the shear wave speed at different lateral positions along the time profile [45]

V. Results

In this section, results for the Agar-gelatine phantom are reported for mechanical properties presented in Table 1. Results are obtained by using both the Finite Element Analysis software LISA in conjunction with Matlab software.

The shear wave speed is estimated using the Time-To-Peak displacement (TTP) method. Time-To-Peak displacement is defined as the time taken by a specific part of tissue to reach its maximum displacement. It is a characteristic for each tissue. Simply, the shear wave speed is estimated as the distance difference between two nodes divided by the time difference at which the TTPs are occurring for these two nodes. Yet, for a good estimation of the shear wave speed, more nodes are involved in the calculation where the average is calculated.

For calculating the TTP, the displacement profile for each node with time must be obtained. In other words, to observe the specific time at which the node reaches its peak displacement. The displacement profile for the central node of excitation is obtained as shown in Figure 6. It is clear from the figure that at t = 0.014 sec. the maximum displacement has taken place. The force is applied at the central node.

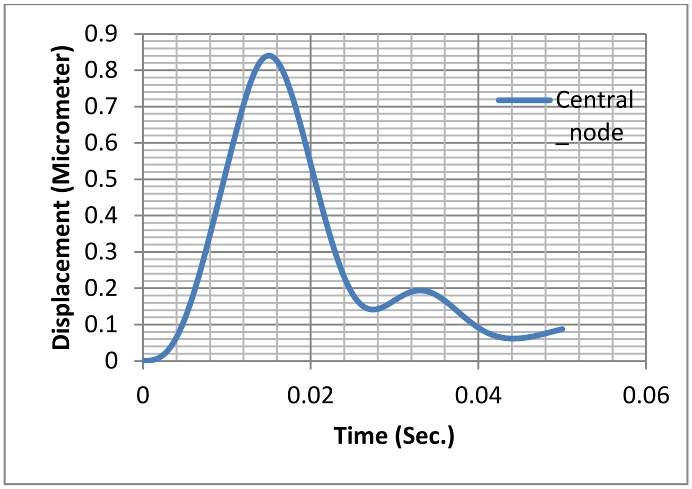


Fig. 6. Displacement magnitude profile

Having displacement profiles for the nodes that are laterally away from the node at which the excitation happens, allows precise estimation of the shear wave speed, by calculating their successive TTPs.

Figure 7 shows clearly the displacements profiles for five successive nodes away from the central node, where the TTPs are calculated from them. Moving outwards from the central node, peaks are observed to be gradually decreasing as the node gets farther.

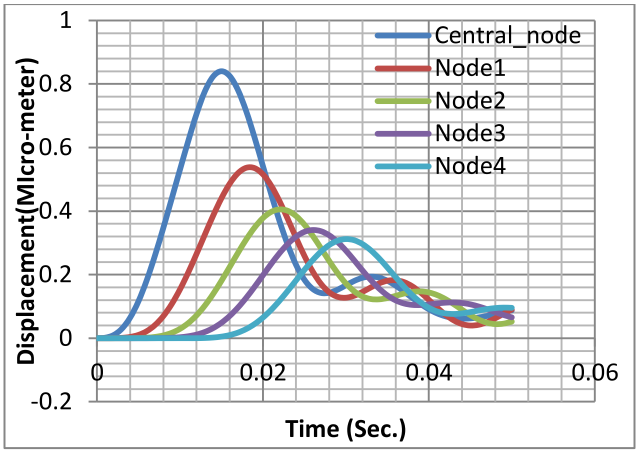


Fig. 7. Displacements profiles for five successive nodes away from the central node

The shear wave speed is calculated as the average value of the quotients, of distance difference between two successive nodes divided by the difference in the TTPs for these same two nodes, for eight nodes, and given by (7) and (8).

$$C_{2,1} = \frac{\Delta x}{\Delta t} \tag{7}$$

$$C_{avg} = \frac{\sum_{i=1}^{n}(c_{n,\,n-1})}{n} \tag{8}$$

The following tables (2), (3), and (4) show different obtained values for the shear wave speeds at each successive two nodes. The average speed is found to be 5.2083 (m/sec) at its optimum case of 0.001 seconds as a time difference between each two successive frames.

TABLE I. Shear Wave Speeds at 8 Different Nodes at 0.001 Sec Difference Between Each Two Successive Frames

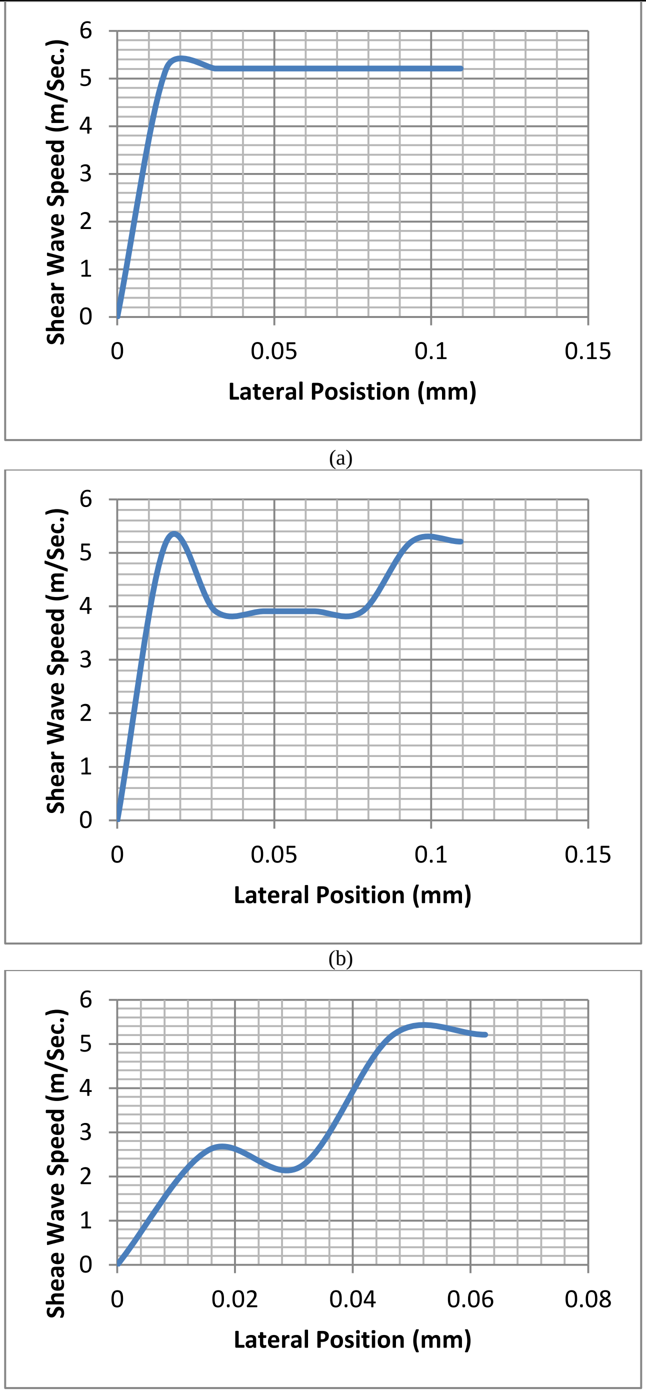
Node no.	Δx (m)	Δt (sec.)	C (m/sec.)
0, 1	0.015625	0.003	5.2083
1, 2	0.015625	0.003	5.2083
2, 3	0.015625	0.003	5.2083
3, 4	0.015625	0.003	5.2083
4, 5	0.015625	0.003	5.2083
5, 6	0.015625	0.003	5.2083
6, 7	0.015625	0.003	5.2083

TABLE II. Shear Wave Speeds at 8 Different Nodes at 0.0009 Sec Difference Between Each Two Successive Frames

Node no.	Δx (m)	Δt (sec.)	C (m/sec.)
0, 1	0.015625	0.003	5.2083
1, 2	0.015625	0.004	3.9063
2, 3	0.015625	0.004	3.9063
3, 4	0.015625	0.004	3.9063
4, 5	0.015625	0.004	3.9063
5, 6	0.015625	0.003	5.2083
6, 7	0.015625	0.003	5.2083

TABLE III. SHEAR WAVE SPEEDS AT 8 DIFFERENT NODES AT 0.0005 SEC DIFFERENCE BETWEEN EACH TWO SUCCESSIVE FRAMES

Node no.	Δx (m)	Δt (sec.)	C (m/sec.)
0, 1	0.015625	0.006	2.6042
1, 2	0.015625	0.007	2.2321
2, 3	0.015625	0.007	5.2083
3, 4	0.015625	0.003	5.2083
4, 5	0.015625	0	NaN
5, 6	0.015625	0	NaN
6, 7	0.015625	0	NaN



(a)

(b)

(c)

Fig. 8. Laterally probed velocities at different frame rates: (a) at 1 KHz, (b) at 1.1 KHz, and (c) at 2 KHz

VI. DISCUSSION

From previous tables, it is clear that estimating the shear wave speed does not give a fixed value for the speed. On other words, there is a tradeoff between the estimated speed and the frame rate used for the estimation process. There is an optimum frame rate that leads to the estimation of the closest speed value, although higher frame rates give closest estimated value.

In our experiments, the optimum frame rate is found to be 1 KHz (each frame of imaging takes about 0.001 sec). Other frame rates were found to give fluctuating velocities around 5 m/seconds as shown in Figure 8. The optimum frame rate in these experiments led to a velocity of 5.2 m/sec. This is a characteristic for this phantom.

On the first laterally probed velocities curve, it is observed that velocities are fixed and independent on the node lateral position inside the phantom. Other laterally probed velocities curves show the instability of the speed and its dependency on the node lateral position.

The calculated shear wave speed for the phantom under study is about 5.7 m/sec. The deviation of our results from the calculated is due to the insufficient number of nodes involved in the phantom construction, yet the speed estimation does not give a precise value as the calculated one. Moreover, the distance between the nodes in the x-direction is predicted to have a role in a good estimation of the speed. As the nodes are the stationary stations from which we monitor the propagation of the wave, less distance means better tracking for the peak displacement.

All methods using external actuators and external vibrating sources are introducing a proper deformation value, which is large enough to be picked up and processed. This facilitates the measurement process of tissue nonlinearities. This also has the drawback of complex hardware to achieve such a function. Amongst these methods, Sono elasticity and transient elasticity.

On the other hand, methods which utilize ARFI have the advantage of using the same ultrasound scanner for excitation and imaging, i.e. the same scanner for probing and measurement. The only disadvantage, here, is that the intensity of the excitation pulse must be kept under certain limit due to the mechanical and thermal considerations when dealing with ultrasound waves [46].

These unwanted bio-effects limits the intensity of excitation, and hence the induced deformation to less than 30 µm. This also limits the shear wave's estimation beyond 6 or 7 cm, due to high attenuation. Furthermore, as mentioned earlier, the high attenuation is considered an advantage, where it produces highly localized shear waves [46].

As far as we know, there is no formula to predict the relationship between the excitation frequency and the elasticity moduli. If we take into consideration that the tissue particles may be modeled as a vibrating pendulum, there will be some excitation frequency that will have no action on these particles, as their moment of inertia will be very high for that frequency. In other words, shear waves will not be generated and hence the tissue will be misestimated to having low modulus of elasticity, but, in reality it is not. This critical frequency or any frequencies approaching it should not be used as they deflect the estimation. This critical excitation frequency will be investigated in our future work.

VII. Conclusion

In this paper, shear wave speed is estimated for the Agar-gelatine. A phantom with specific mechanical properties from literature is used. A model for this phantom is generated by a finite element modeling software to simulate its behavior. A point source force is applied at the focal point, which is the center of the phantom, to stimulate the shear wave propagation. TTP method is used to estimate the speed. The estimated speed is found to be 5.2 m/sec; while the calculated one is 5.7 m/sec. The difference between both of them arises from the shortage of the number of nodes by which the model has been constructed. Further investigation will be carried out to improve the results and to study the effect of the excitation frequency on the estimated speed.

References

[1] M. L. Palmeri, K. R. Nightingale, “Acoustic Radiation Force Based Elasticity Imaging methods”, Interfocus, vol. 1, pp. 553-564, 2011.

[2] W. M. Lai, D. Rubin, E.Krempl, “Introduction To Continuum Mechanics”, M.A. Woburn: Butterworth- Heinmann, 1999.

[3] F. Duck, “Physical Properties Of Tissue a Comprehensive Reference Book”, New York, NY: Academic Press, 1990.

[4] A. Sarvazyan, A. Skovoroda, S. Emelianov, J. B. Fowlkes, J. G. Pipe, R. S. Adler, R. B. Buxton, L. Carson, “Biophysical Bases Of Elasticity Imaging”, Acoust. Imag., vol.21, 223–240, 1995.

[5] A. Sarvazyan, “Elastic Properties Of Soft Tissue, Handbook Of Elastic Properties Of Solids, Liquids and Gases”, eds M. Levy, H. E. Bass and R. R. Stern, London, UK: Academic Press, pp. 107–127, 2001.

[6] C. Kasai, N. Koroku, A. Koyano, R. Omoto, “Real-time Two-dimensional Blood Flow Imaging Using an Autocorrelation Technique”, IEEE Trans. Ultrason., Ferroelec., Freq. Contr., vol. SU-32, pp. 458–463, 1985.

[7] W. Walker, G. Trahey, “A Fundamental Limit On Delay Estimation Using Partially Correlated Speckle Signals”, IEEE Trans. Ultrason. Ferroelec. Freq. Contr., vol. 42, pp. 301–308, 1995.

[8] Y. C. Fung, “Biomechanics: Mechanical Properties Of Living Tissues”, New York, NY: Springer, 2nd edn. 1993.

[9] R. Dickinson, C. Hill, “Measurement Of Soft Tissue Motion Using Correlation Between A-scans”, Ultrasound Med. Biol., vol. 8,pp. 263–271, 1982.

[10] L. Wilson, D. Robinson, ‘Ultrasonic Measurement Of Small Displacements and Deformations Of Tissue”, Ultrason. Imag., vol. 4, pp. 71–82, 1982.

[11] R. M. Lerner , K. J. Parker , J. Holen, R. Gramiak, R. C. Waag, “Sono-elasticity: Medical Elasticity Images Derived From Ultrasound Signals in Mechanically Vibrated Targets”, Acoust. Imag., vol. 16, pp. 317–327, 1988.

[12] J. Ophir, I. Cespedes, H. Ponnekanti, X. Li, “Elastography: A Quantitative Method For Imaging The Elasticity Of Biological Tissues”, Ultrason. Imag., vol. 13, pp. 111–134, 1991.

[13] T. Sugimoto, S. Ueha, K. Itoh, “Tissue Hardness Measurement Using The Radiation Force Of Focused Ultrasound”, IEEEXplore, vol. 3, pp. 1377–1380, 1990.

[14] J. F. Greenleaf, M. Fatemi, M. Insana, “Selected Methods For Imaging Elastic Properties Of Biological Tissues”, Ann. Rev. Biomed. Eng., vol. 5, pp. 57–78, 2003.

[15] K. J. Parker, L. S. Taylor, S. M. Gracewski, D. J. Rubens, “A Unified View Of Imaging The Elastic Properties Of Tissue”, J. Acoust. Soc. Am., vol. 117, pp. 2705–2712, 2005.

[16] W. L. M. Nyborg, “Acoustic Streaming”, In Physical acoustics (ed. W. P. Mason), pp. 265–331, New York, NY: Academic Press Inc, 1965.

[17] G. R. Torr, “The Acoustic Radiation Force”, Am. J. Phys., vol. 52, pp. 402–408, 1984.

[18] A. Sarvazyan, T. J. Hall , M. W. Urban, M. Fatemi, S. R. Aglyamov, B. S. Garra, “An Overview Of Elastography, An Emerging Branch Of Medical Imaging”, Curr. Med. Imaging Rev., vol. 7, pp. 255-282, 2011.

[19] T. A. Krouskop, D. R. Dougherty, F. S. Vinson, “A Pulsed Doppler Ultrasonic System For Making Noninvasive Measurements Of The Mechanical Properties Of Soft Tissue”, J Rehabil Res Dev., vol. 24, pp. 1–8, 1987.

[20] R. M. Lerner, K. J. Parker, J. Holen, R. Gramiak, R. C. Waag, “Sono-elasticity: Medical Elasticity Images Derived From Ultrasound Signals In Mechanically Vibrated Targets”; Acoustical Imaging., vol. 16, pp. 317-327. New York, NY: Plenum Press; 1988.

[21] K. Hoyt, T. Kneezel, B. Castaneda, K. J. Parker, “Quantitative Sonoelastography For The In Vivo Assessment Of Skeletal Muscle Viscoelasticity”, Phys Med Biol., vol. 53, pp. 4063–4080, 2008.

[22] K. R. Nightingale, M. L. Palmeri, R. W. Nightingale, G. E. Trahey , “On The Feasibility Of Remote Palpation Using Acoustic Radiation Force”, J Acoust Soc Am., vol. 110, pp. 625–34, 2001.

[23] M. L. Palmeri, S. A. McAleavey, K. L. Fong, G. E. Trahey, K. R. Nightingale, “Dynamic Mechanical Response Of Elastic Spherical Inclusions To Impulsive Acoustic Radiation Force Excitation”, IEEE Trans Ultrason Ferroelectr Freq Control., vol. 53, pp. 2065–79, 2006.

[24] M. L. Palmeri, S. A. McAleavey, K. L. Fong, G. E. Trahey and K. R. Nightingale, “Dynamic Mechanical Response Of Elastic Spherical Inclusions To Impulsive Acoustic Radiation Force Excitation”, IEEE Trans Ultrason Ferroelectr Freq Control., vol. 53, pp. 2065–79, 2006.

[25] B. J. Fahey, K. R. Nightingale, D. L. Stutz, G. E. Trahey, “Acoustic Radiation Force Impulse Imaging Of Thermally- and Chemically-induced Lesions In Soft Tissues”, preliminary ex vivo results., vol. 30, pp. 321–8, 2004.

[26] B. J. Fahey, K. R. Nightingale, R. C. Nelson, M. L. Palmeri, G. E. Trahey, “Acoustic Radiation Force Impulse Imaging Of The Abdomen: Demonstration Of Feasibility and Utility”, Ultrasound Med Biol., vol. 31, pp. 1185–98, 2005.

[27] L. Zhai, J. Madden, W. C. Foo, M. L. Palmeri, V. Mouraviev, T. J. Polascik, K. R. Nightingale, “Acoustic Radiation Force Impulse Imaging Of Human Prostates E-vivo”, Ultrasound Med Biol., vol. 36, pp. 576–588, 2010.

[28] D. Dumont, J. Dahl, E. Miller, J. Allen, B. J. Fahey, G. E. Trahey, “Lower-limb Vascular Imaging With Acoustic Radiation Force Elastography: Demonstration Of In-vivo Feasibility”, IEEE Trans Ultrason Ferroelectr Freq Control., vol. 56, pp. 931–944, 2009.

[29] J. J. Dahl, D. M. Dumont, J. D. Allen, E. M. Miller, G. E. Trahey, “Acoustic Radiation Force Impulse Imaging For Noninvasive Characterization Of Carotid Artery Atherosclerotic Plaques: A Feasibility Study”, Ultrasound Med Biol., vol. 35, pp. 707–716, 2009.

[30] S. Catheline, J. L. Thomas, F. Wu, M. A. Fink, “Diffraction field Of A Low Frequency Vibrator In Soft Tissues Using Transient Elastography”, IEEE Trans Ultrason Ferroelectr Freq Control., vol. 46, pp. 1013–1019, 1999.

[31] S. Catheline, F. Wu, M. A. Fink, “A Solution To Diffraction Biases In Sonoelasticity: The Acoustic Impulse Technique”, J Acoust Soc Am., vol. 105, pp. 2941–50, 1999.

[32] V. Dutt, R. R. Kinnick, R. Muthupillai, T. E. Oliphant, R. L. Ehman, J. F. Greenleaf, “Acoustic Shear-wave Imaging Using Echo Ultrasound Compared To Magnetic Resonance Elastography”, Ultrasound in Med. & Biol ,vol. 26, pp. 397–403, 2000.

[33] J. L. Gennisson, G. Cloutier, “Sol-gel Transition In Agar-gelatin Mixtures Studied With Transient Elastography”, IEEE Trans Ultrason Ferroelectr Freq Control., vol. 53, pp. 716–723, 2006.

[34] K. Nightingale, S. McAleavey, G. E. Trahey, “Shear-wave Generation Using Acoustic Radiation Force: In-vivo and Ex-vivo Results”, Ultrasound Med Biol., vol. 29, pp. 1715–23, 2003.

[35] M. A. Lubinski, S. Y. Emelianov, M. O’Donnell, “ Speckle Tracking Methods For Ultrasonic Elasticity Imaging Using Short-time

Correlation", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 46, pp. 82–96, 1999.

[36] Y. Zheng, S. Chen, W. Tan, R. Kinnick, J. F. Greenleaf, "Detection Of Tissue Harmonic Motion Induced By Ultrasonic Radiation Force Using Pulse-echo Ultrasound and Kalman Filter", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 54, pp. 290–300, 2007.

[37] G. F. Pinton, J. J. Dahl, G. E. Trahey, "Rapid Tracking Of Small Displacements With Ultrasound", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 53, pp. 1103–17, 2006.

[38] M. L. Palmeri, M. H. Wang, J. J. Dahl, K. D. Frinkley, K. R. Nightingale, " Quantifying Hepatic Shear Modulus In-vivo Using Acoustic Radiation Force", Ultrasound Med Biol., vol. 34, pp. 546–558, 2008.

[39] R. R. Bouchard, S. J. Hsu, P. D. Wolf, G. E. Trahey, "In-vivo Cardiac, Acoustic-Radiation-Force-Driven, Shear Wave Velocimetry", Ultrason Imaging., vol. 31, pp. 201–213, 2009.

[40] J. Bercoff, M. Tanter, M. Fink, "Sonic Boom In Soft Materials: The Elastic Cerenkov Effect", Appl Phys Lett., vol. 84, pp. 2202–2204, 2004.

[41] J. Bercoff, M. Tanter, M. Fink, "Supersonic Shear Imaging: A New Technique For Soft Tissue Elasticity Mapping", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 51, pp. 396–409, 2004.

[42] M. Tanter, J. Bercoff, A. Athanasiou, T. Deffieux, J. L. Gennisson, G. Montaldo, M. Muller, A. Tardivon, M. Fink, " Quantitative Assessment Of Breast Lesion Viscoelasticity: Initial Clinical Results Using Supersonic Shear Imaging", Ultrasound Med Biol., vol. 34, pp. 1373–1386, 2008.

[43] J. Bercoff, M .Tanter, M. Fink, "Supersonic Shear Imaging: A New Technique For Soft Tissue Elasticity Mapping", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 51, pp. 396-409, 2004.

[44] M. L. Palmeri, S. A. McAleavey, K. L. Fong, G. E. Trahey, K. R. Nightingale, "Dynamic Mechanical Response Of Elastic Spherical Inclusions To Impulsive Acoustic Radiation Force Excitation", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 53, pp. 2065-2079, 2006.

[45] N. C. Rouze, M. H. Wang, M. L. Palmeri, K. R. Nightingale, "Robust Estimation Of Time-Of-Flight Shear Wave Speed Using Radon Sum Transform", IEEE Trans Ultrason Ferroelectr Freq Control., vol. 57, pp. 2262-2270, 2010.

[46] A. P. Sarvazyan, O. V. Rudenko, S. D. Swanson, J. B. Fowlkes, S. Y. Emelianov, "Shear Wave Elasticity Imaging: A New Ultrasonic Technology Of Medical Diagnostics", Ultrasound Med Biol., vol. 24, pp. 1419–35, 1998.

A Variant of Genetic Algorithm Based Categorical Data Clustering for Compact Clusters and an Experimental Study on Soybean Data for Local and Global Optimal Solutions

Abha Sharma
Maulana Azad National Institute of Technology,
Bhopal, India

R. S. Thakur
Maulana Azad National Institute of Technology,
Bhopal, India

***Abstract*—Almost all partitioning clustering algorithms getting stuck to the local optimal solutions. Using Genetic algorithms (GA) the results can be find globally optimal. This piece of work offers and investigates a new variant of the Genetic algorithm (GA) based *k*-Modes clustering algorithm for categorical data. A statistical analysis have been done on the popular categorical dataset which shows the user specified cluster centres stuck at local optimal solution in *k*-Modes algorithm even in all the higher iterations and the proposed algorithm overcome this problem of local optima. To the best of our knowledge, such comparison has been reported here for the first time in case of categorical data. The obtained results, shows that the proposed algorithm is better over the conventional *k*-Modes algorithm in terms of optimal solutions and within cluster variation measure.**

Keywords—Clustering*; *Categorical data; k-Modes; Genetic Algorithm

I. Introduction

There is a growing requirement for the way to extract knowledge from the data [1]. Clustering is a descriptive task which partition the dataset based on the predefined similarity measure [2]. Clustering techniques have been widely used in machine learning, pattern recognition, medical etc. Number of clustering algorithms have been proposed for different requirements and nature of the data [3]. Partition based clustering (k-Modes and its initialisation methods) [4], hierarchical clustering (HIERDENC) [5] model-based clustering (EAST algorithm) [6], density-based clustering [7], graph-based clustering, and grid-based clustering are some basic clustering algorithms with their advantages and disadvantages.

It is hard to discover the distance measure between two categorical data objects, greater the distance between the clusters more separated will be the clusters [8]. One of the well-known clustering for categorical data is *k*-Modes algorithm for large datasets. The traditional way to treat categorical data is binary but does not do justice to the large value difference such as for the very low and very high the difference is same.

The major issue in partition clustering is to initialize the cluster centres, since it has a direct influence on the construction of ultimate clusters. This paper focus on the better partitions of all real world categorical datasets on the lowest cost using GA in less space and time.

GA is proposed by Holland [9] and can apply to many optimization problems. Due to the cluster centre initialization problem which affects the proper clustering of data, GA has been used to convert the local optimal solution into global optimal solution in many GA based clustering algorithm for numeric as well as categorical data in the literature [10]. This paper calculates Total Within Cluster Variation (TWCV), time and conversion of local optima to global optima. Fig 1. shows the various operators used in GA based categorical data clustering found in literature.

This paper is organized as follows: Section II presents Literature Review: Section III presents Background: Section IV presents proposed method: Section V shows the Experimental details where we compare basic *k*-Modes algorithm with propose algorithms: Section VI concludes the paper: Section VII tries to put the future work.

II. Literature Review

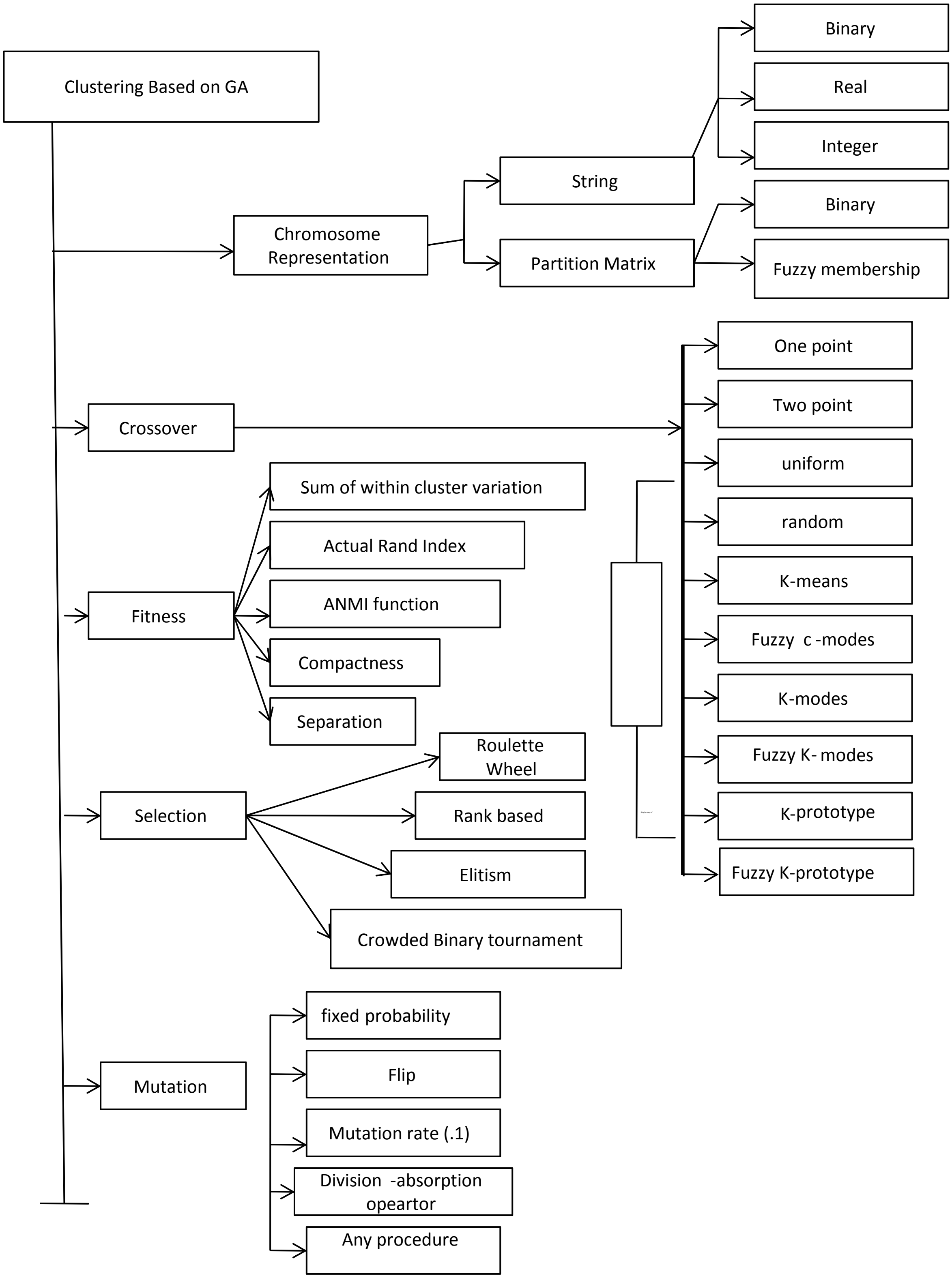


Fig. 1. Many operators used in GA based clustering according to Literature

TABLE I. Comparison of Algorithms

Operators	**NSGA-FMC [10]**	**MOGA**	**G-AMNI [13]**	**Improved G-AMNI [11]**	**AGCUK [12]**	**GA and Simulated annealing based CDC**	**A Genetic *k*-Modes Algorithm [15]**
Chromosome representation	Fuzzy membership matrix kxN matrix. where n=number of data objects K=number of cluster	K x A matrix K=number of clusters A= categorical attributes Chromosomes=set of clustering centres , Comparatively small	-	K x N	The length of the chromosome is K x m, where K =number of clusters and m =number of attributes	Kx N where *K* is the number of clusters and *n* is the number of points.	NxK partition matrix
Population initialization	Code with random numbers	*K* random objects of the categorical dataset of *P(population size)* chromosomes in the population	randomly selected partitions of objects	-		randomly selected partitions of objects	randomly
Chromosome selection	Roulette wheel strategy	Crowded binary tournament selection	Roulette wheel strategy	Roulette wheel strategy		-	-
Fitness of chromosome	Rank based evaluation function	Separation and compactness function	ANMI Function $\Phi^{(ANMI)}=1/r\sum \Phi^{(NMI)}(\lambda^{(q)}, \lambda)$	ANMI function	Davies–Bouldin (DB) index		
Crossover	One step fuzzy k-modes crossover operator	single-point crossover depending on crossover probability μc	Single point crossover, crossover site is selected randomly	-	-	single point crossover with a fixed crossover probability of μc	One step kmode operator
Mutation	.01 mutation	Two step mutation probability μm (a) the gene position is selected randomly (b) the categorical value of that position is replaced by another random value chosen from the corresponding categorical domain.	uniform probability	-	Division–absorption mutation Division operation: the most sparse cluster is determined Absorption operation: determine which cluster is to be merged	fixed probability μm.	The mutation operator changes an allele value depending on the distance between the cluster center and the corresponding data point.
Selection	-	-	roulette wheel strategy, if best chromosome not found use elite selection	-	Elitist operation	Propotional/ roulette wheel	-
Sorting	Elitism non-dominated sorting plus crowded tournament selection is used to evaluate the clustering solution.	-	--	-	-	-	-
Termination criteria	--	-	-	-	In general, two stopping criteria are used in genetic algorithms: based on fixed number of iteration and no further improvement in fitness value of the best individual. This work used fixed number of iteration.	Fixed number of iterations. Elitism at each generation.	-
Tested Dataset	*Soybean* *Zoo*	*Zoo* *Soybean*	*Brest cancer* *Vote*	*Zoo* *Vote*	*Breast cancer* *Wisconsin breast*	*Soybean* *Zoo*	*Mushroom* *Votes*

	Votes	*Breast Cancer* *Vote*	*Zoo* *Mushroom*	*Brest cance* *Mushroo* *m*	*cancer*	*Tic Tac Toe*	*Zoo*
Validity Measure	Compactness, Separation, Time complexity, Actual Rand Index	-	Clustering Error	Accuracy	-	Mincowski value	Corrected Rand Index

III. Background

In many categorical data clustering algorithms the seeds or the cluster centres are not known in advance for example *k*-Modes algorithm is a well-known and widely used clustering technique of this type. However, the major drawback of the *k*-Modes is that it often gets stuck at local minima and the result is largely dependent on the choice of the initial cluster centres.

A. K-Modes Algorithm

a) Dissimilarity measure

Let *A* and *B* be two categorical objects described by *m* categorical attributes. The dissimilarity measure can be defined by the total mismatches of the corresponding attribute categories of the two objects [4]. Formally

$$d_1\left(A,B\right)=\sum_{j=1}^{m}d(a_j,b_j)-----(1)$$

Where

$$d\chi^2(A,B)=\sum_{j=1}^{m}\frac{(n_{aj}+n_{bj})}{n_{aj}*n_{bj}}\delta(a_j,b_j)----(3) \qquad (2)$$

where n_{aj} and n_{bj} are the number of objects in the dataset that have categories a_j and b_j for attribute *j* and $d\chi^2$ $(X, Y)$is Chi-square distance.

This paper work on dataset having frequencies of categories then the distance calculation [4] eq. (2) is used to calculate the distance.

Consider *X* is set of categorical objects described by categorical attributes, $C_1; C_2; : : : ; C_u$

Definition 1. Mode of X is a vector $M=\left[m_1, m_{2,\ldots\ldots,} m_n\right]$ that minimizes $D\left(X,\mathrm{M}\right)=\sum_{j=1}^{m}d_1(X_i,\mathrm{M})$. Where *M* may or may not the element of *X*. [4]

b) Find a mode for a set

Let $n_{t_{k,\,j}}$ be the number of data objects having the k^{th} category $t_{k,\,j}$ in attribute C_j and the relative frequency of $t_{k,\,j}$ in X is frC_j= $t_{k,\,j}$ $|X=\frac{n_{t_{k,\,j}}}{n}$ [4]

Theorem 1. The function D(*X*, *M*) is minimized if $fr\left(C_j=m_j \mid X\right)\geq f_r(A_j=t_{k,\,j}\mid X)$ for $q_j\neq c_{k,\,j}$ for all *j*=1, 2,..,*m*.

c) The k-Modes algorithm

When equation (1) and (2) are used as the dissimilarity measure for categorical objects, the cost function becomes

$$P(W,M)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}w_{i,\,l}\delta(x_{i,\,j},m_{l,\,j})------(3)$$

where $w_{i,\,j}\in W$ and $M_l=\left[m_{l1},\ m_{l1},\ \ldots..,\ m_{lu}\right]\in M$

IV. Proposed Work

An attempt is made in this paper to integrate the effectiveness of the *k*-Modes algorithm for partitioning data into a number of clusters, with the capability of genetic algorithm to bring it out of this local minima. GAs are randomized search and are efficient to provide near optimal solutions of fitness function in an optimization problem.

The Local Search based approach such as *k*-Modes may get stuck at the local optimum solutions. Genetic algorithm based clustering escape from the local optimum, but it is slow and expensive to compute. The Similarity based approaches are not consistent among different inputs and can be context dependent. A small gap between k-modes and proposed GA based algorithm is the assumption of cluster centres on which the clustering is based. GA is an efficient algorithm to solve optimization problems which represented by chromosomes as string encodings and has multiples solutions. GA opts for the best fit solutions in each generation.

Increase in the string length of the chromosome, the search space in GAs increases therefore the whole process becomes more time consuming. When the number of data points and number of attributes are very large then the size of a chromosome which is equal to the number of data points multiplied by number of clusters assumed is difficult to store and manage.

In this paper, a generalized mechanism for all the categorical datasets is presented to identify and ignore the worst cluster centres in a categorical data set. Proposed work utilizes the robustness of genetic algorithm (GA) to optimize the *k*-Modes clustering algorithm that uses searching capability of GAs to determine most appropriate cluster centres which also prohibits the expensive crossover operator by using one step *k*-Modes operator. The associated cost function is defined in terms of the distances between the cluster objects and the cluster centre. This paper presents chromosomes in the form of strings (sequence of data values).

The objective of this work is to find *k* partitions that minimize the cost function and find optimal solution with some GA operators; string representation, population size, selection operator and one-step *k*-Modes algorithm in the place of the crossover operator This paper shows how the conventional k-Modes clustering algorithm may stops at locally optimal solution whereas the proposed hybridized clustering algorithm facilitate the global optimization of the underlying cost as objective function, to construct optimal partition of objects so that the within-cluster dispersion can be minimized and the between-cluster separation can be increased.

a) Encoding for categorical data:

Variant size of matrices are developed for chromosome representation in almost all GA based categorical data clustering. In this paper the chromosomes are encoded as string with *N***k* size [14].

Example 1. Suppose *N*=2 and *k*=4 then the string representation for a chromosome is (Yellow Small Stretch Adult Purple Large Dip Child) from Lenses real world dataset. It embed the two clusters (Yellow Small Stretch Adult) and (Purple Large Dip Child). Each categorical data in the chromosome is a allele.

Consequently the updated cluster mode is (Yellow Small Stretch Adult) (Purple Large Dip Child) with the frequency based method shown in equation (2) later, the cost or within cluster variation is calculated.

b) Fitness calculation

This paper presents fitness function as the sum of within clustering variation, larger the fitness, denser the data in cluster and more separated from the other clusters. The details are described below.

Initially the clusters are formed randomly using the centres encoded in that particular chromosome, then the cluster centres encoded in the chromosome are replaced by the cluster centre (modes) of the respective clusters using frequency method. Therefore assign each point x_i=1, 2, 3,...,n_i with mode m_j such that

$$\|x_i - m_i\| \ < \ \|x_i - m_t\|,\ t = 1,\ 2, 3\ ,....k$$

Frequency method shown in eq. (2) for attribute *j* where C_j is replaced by new C_i.

$$fr\ C_j{*} = t_{k,\ j} \,|\, X = \frac{n_{t_{k,\ j}}}{n}$$

Therefore the fitness is calculated using following equation

$$P(W,M) = \sum_{l=1}^{k} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{i,\,l}\,\delta(x_{i,\ j}, m_{l,\ j}) \quad \text{--------(4)}$$

The fitness function is defined as *f=1/P(W,M)* i.e. less the cost more fit will be the chromosome.

c) Selection

The fundamental selection method for GA based clustering algorithm is spinning the roulette wheel. In this paper after fitness calculation sort the cost of all the chromosomes in the population in the present generation, delete if the highest cost of chromosome in present generation is greater than the average of all cost of the chromosomes in the next iteration else keep that chromosome in that population.

d) Crossover process

Similar to genetic *k*-Modes algorithm, this paper also used one step *k*-Modes algorithm as the crossover operator to exchange of information between the two parent chromosomes to generate two offspring's.

e) Termination criteria

The most popular termination criteria for GA based clustering algorithms are: to run the algorithm based on user defined iterations. In the proposed algorithm iteration stops for the particular chromosome if the constant fitness value persist even before user specified iteration count.

f) Solution of the Empty cluster problem

The Empty cluster formation is the well-known problem in clustering. And the problem becomes big if the optimization techniques are used, this paper try to remove the empty cluster issue using following algorithm:

Algorithm:

If (In any generation for C_i the intermediate clusters in chromosome are found to be null or empty)

```
{
 iteration ++
if(found any empty cluster)
{
delete the chromosome & $M$=cost $C_i$
}
else
go head
else go to next iteration ()
    }
```

Example 2. Suppose *N*=4 and *k*=2 if the intermediate clusters

After first generation:

(Yellow Small Stretch Adult) (Purple Large Dip Child)

After Second generation:

(Yellow Adult Stretch Child) (Purple Large Dip Adult)

......
......
......
After m^{th} generation

(Yellow Small Dip Child) ( )

After n^{th} generation

( ) ( )

After using the above algorithm the updated clusters are:

(Yellow Small Stretch Adult) (Purple Large Dip Child)

In the current implementation of GA this paper used the standard *k*-Modes algorithm for creating multiple partitions of the categorical data for global optimal solution.

B. Flowchart of proposed GA based clustering Algorithm

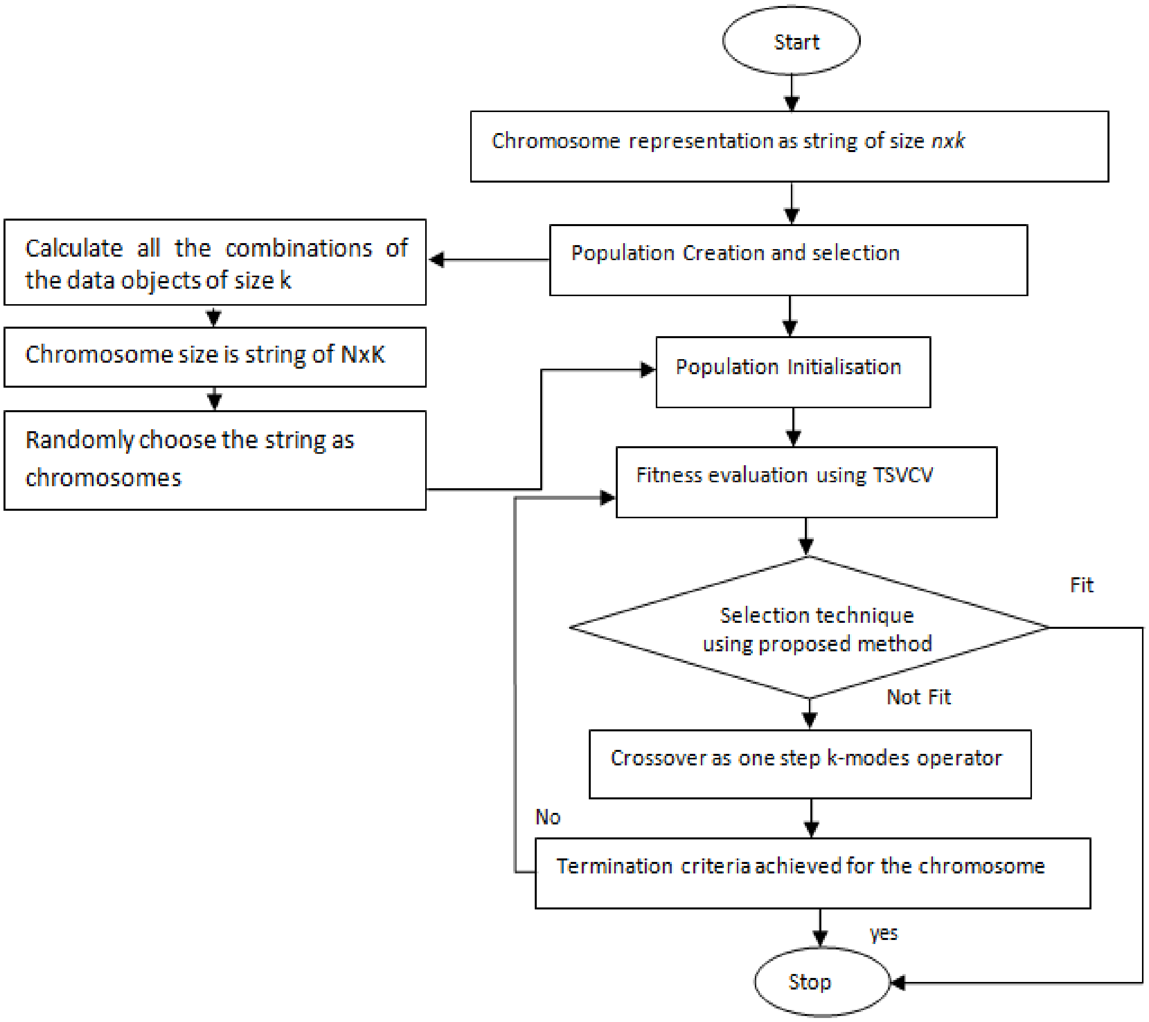

Start
Chromosome representation as string of size *nxk*
Population Creation and selection
Calculate all the combinations of the data objects of size k
Chromosome size is string of NxK
Randomly choose the string as chromosomes
Population Initialisation
Fitness evaluation using TSVCV
Selection technique using proposed method
Fit
Not Fit
Crossover as one step k-modes operator
No
Termination criteria achieved for the chromosome
yes
Stop

Fig. 2. Flow chart Proposed clustering Algorithm

V. Experimental Results

In this work, the proposed GA based clustering algorithm and the standard *k*-Modes method were coded using python language. The experiments has been conducted on a computer laptop with 2.89 GHz CPU and 8 G RAM under a Windows 8.1 operating system. To test the effectiveness of the proposed algorithm on Soybean dataset from UCI [16] has been used.

Soybean dataset: The dataset contains 47 instances, 35 attributes, and 19 classes and four classes are considered in reality. And out of 35 attributes 14 attributes categories are same so we shall use 21 attributes only. Existing *k*-Modes algorithm has been run for 100 iterations with different initialisation say different seeds and different number of *k*. Proposed GA based clustering algorithm were executed 5 times for soybean dataset and *k*-Modes executed approx. 10-10 times for each *k* of the dataset. Proposed algorithm has been run till 100 iterations. To evaluate performance measure computational time (in seconds) has been calculated for algorithm efficiency.

Secondly, the TWCV is an intrinsic validity measure to calculate the sum of within cluster variation for all clusters. The smaller value of TWCV means the dataset are more compact. Therefore, in order to obtain compact clusters or mor separated clusters the value should be minimized for clustering task. If only considering the computational efficiency, the faster algorithm is better. The detailed analysis will be shown in the next sub-sections.

The shaded values shown in tables II-VII are locally optimal and globally optimal in case of *k*-Modes and proposed algorithm respectively.

The detailed clustering results of *k*-Modes algorithm for soybean data on different initialization with different *k* values has been shown in Table II-IV which shows the values are stuck at locally optimal. The proposed clustering algorithm provides the optimal values from table V-VII in all the runs for all the *k*. *K*-Modes algorithm also attains somewhere the optimal value as proposed value of the total runs but the ratio are very less. Table VIII, X, XII, XIV, XVI shows the average cost of different initialisation in 100th iteration for different *k* using *k*-Modes. Table IX, XI, XIII, XV, XVII shows the average cost of different population in 100th iteration for different *k* using proposed algorithm.

Fig. 3 shows the cost gap increases between the *k*-Modes and proposed method which shows the compact clustering of proposed algorithm. Fig. 4. shows the time obtained to cluster *k*-Modes and proposed method show the very less time gap between the proposed algorithm and *k*-Modes.

TABLE II. TWCV USING *K*-MODES ALGORITHM ON VARIOUS ITERATIONS WITH VARIOUS CLUSTER CENTRES WHEN K=2

Initial Configuration	i=1	i=2	i=3	i=4	i=5	i=6	i=100
1	23.89	19.29	16.37	16.34	16.34	16.34	16.34
2	20.96	19.29	18.25	17.7	16.93	16.24	16.24
3	20.99	17.18	17.07	17.07	17.07	17.07	17.07
4	22.22	16.78	16.35	16.34	16.34	16.34	16.34
5	21.42	19.81	18.19	17.41	17.07	16.34	16.34
6	25.3	19.35	19.22	18.89	17.89	17.87	17.87
7	29.13	17.88	17.07	16.34	16.34	16.34	16.34
8	24.28	17.33	16.5	16.5	16.5	16.5	16.5
9	25.033	17.93	17.07	16.34	16.34	16.34	16.34
10	31.53	17.93	17.07	16.34	16.34	16.34	16.34
11	20.29	19.22	19.22	19.22	19.22	19.22	19.22
12	18.23	17.15	17.15	17.15	17.15	17.15	17.15
13	7.35	17.07	17.07	17.07	17.07	17.07	17.07

TABLE III. TWCV USING *K*-MODES ALGORITHM ON VARIOUS ITERATIONS WITH VARIOUS CLUSTER CENTRES WHEN K=3

Initial Configuration	i=1	i=2	i=3	i=4	i=5	i=6	i=100
1	20.33	15.72	15.4	15.4	15.4	15.4	15.4
2	20.58	20.32	19.47	19.47	19.47	19.47	19.47
3	17.77	16.41	15.87	17.48	18.41	16.24	16.24
4	22.16	15.15	14.93	14.93	14.93	14.93	14.93
5	23.56	19.89	18.19	17.41	17.07	16.34	16.34
6	27.72	14.32	14.32	15.26	15.22	15.22	15.22
7	24.35	17.42	18.68	16.35	14.78	15.22	15.22
8	25.28	17.7	17.7	15.64	15.54	15.48	15.48
9	20.63	17.76	15.07	14.41	15.16	14.93	14.93
10	28.69	15.73	15.44	15.44	15.44	15.44	15.44
11	20.87	17.34	16.45	16.42	16.42	16.42	16.42
12	18.04	15.34	16.42	16.42	16.42	16.42	16.42

TABLE IV. TWCV USING K-MODES ALGORITHM ON VARIOUS ITERATIONS WITH VARIOUS CLUSTER CENTRES WHEN K=4

Initial Configuration	i=1	i=2	i=3	i=4	i=5	1=6	i=100
1	19.58	19.16	17.24	15.04	19.79	18.99	18.99
2	20.14	18.1	17.22	16.24	16.24	16.24	16.24
3	17.58	15.44	15.44	15.44	15.44	15.44	15.44
4	21.44	19.89	18.1	17.41	17.04	16.34	16.34
5	23.59	15.92	15.66	16.32	16.22	16.22	16.22
6	28	17	15.29	15.29	15.29	15.29	15.29
7	19.13	12.57	13.87	13.83	13.83	13.83	13.83
8	26.34	19.37	17.94	17.07	16.34	16.34	16.34
9	20.59	12.75	12.75	12.75	12.75	12.75	12.75
10	28.46	16.6	15.43	15.41	15.41	15.41	15.41
11	21.39	17.89	17.49	17.31	17.31	17.31	17.31
12	16.59	16.06	15.31	15.97	15.18	14.93	14.93
13	18.97	15.36	16.42	16.42	16.42	16.42	16.42

TABLE V. TWCV using Proposed Algorithm on Various Iterations with Various Population Size when K=2

Initial Population	i=1	i=2	*i=3*	i=4	i=100
5	16.7	15.44	15.44	15.44	15.44
10	17.25	15.59	15.59	15.59	15.59
15	17.09	15.44	15.44	15.44	15.44
20	17.28	15.44	15.44	15.44	15.44
100	16.92	15.44	15.44	15.44	15.44

TABLE VI. TWCV using Proposed Algorithm on Various Iterations with Various Population Size when K=3

Initial Population	i=1	i=2	i=3	i=4	i=100
5	11.99	11.06	11.06	11.06	11.06
10	15.16	11.06	11.06	11.06	11.06
15	16.88	11.29	11.06	11.06	11.06
20	15.05	11.06	11.06	11.06	11.06
100	11.58	11.14	11.06	11.06	11.06

TABLE VII. Twcv using Proposed Algorithm on Various Iterations with Various Population Size when K=4

Initial population	i=1	i=2	i=3	i=4	i=100
5	15.23	10.97	10.55	10.55	10.55
10	12.29	10.55	10.55	10.55	10.55
15	13.37	8.60	8.60	8.60	8.60
20	11.78	11.15	11.14	11.14	11.14
100	11.40	10.28	10.47	10.47	10.47

TABLE VIII. Average TWCV Obtained Using K-Modes Algorithm for Different Initialisation After 100 Iteration when *K*=2

Initial configuration	*k*-Modes
1.	18.10
2.	19.22
3.	17.07
4.	17.07
5.	16.34

TABLE IX. Average TWCV Obtained Using Proposed Algorithm for Different Population Size After 100 Iteration when *K*=2

Initial population	Proposed algorithm
5	15.44
10	15.44
15	15.44
20	15.44
100	15.44

TABLE X. Average TWCV obtained using k-modes algorithm for different initialisation after 100 iteration when k=3

Initial configuration	*k*-Modes
1.	15.44
2.	19.47
3.	15.48
4.	16.24
5.	16.24

TABLE XI. Average TWCV obtained using proposed algorithm for different population size after 100 iteration when *k*=3

Initial population	Proposed algorithm
5	11.06
10	11.06
15	11.06
20	11.06
100	11.06

TABLE XII. Average TWCV obtained using k-modes algorithm for different initialisation after 100 iteration when k=4

Initial configuration	*k*-Modes
1.	14.80
2.	16.18
3.	10.62
4.	15.22
5.	15.22

TABLE XIII. Average TWCV obtained using proposed algorithm for different population size after 100 iteration when *k*=4

Initial population	Proposed algorithm
5	10.55
10	10.55
15	8.60
20	11.14
100	10.47

TABLE XIV. AVERAGE TWCV OBTAINED USING K-MODES ALGORITHM FOR DIFFERENT INITIALISATION AFTER 100 ITERATION WHEN K=5

Initial configuration	*k*-Modes
1.	18.99
2.	16.24
3.	17.31
4.	16.34
5.	16.34

TABLE XV. AVERAGE TWCV OBTAINED USING PROPOSED ALGORITHM FOR DIFFERENT POPULATION SIZE AFTER 100 ITERATION WHEN *K*=5

Initial population	Proposed algorithm
5	11.35
10	9.70
15	11.14
20	10.39
100	9.24

TABLE XVI. AVERAGE TWCV OBTAINED USING K-MODES ALGORITHM FOR DIFFERENT INITIALISATION AFTER 100 ITERATION WHEN K=6

Initial configuration	*k*-Modes
1.	14.12
2.	12.89
3.	16.22
4.	18.99
5.	16.22

TABLE XVII. AVERAGE TWCV OBTAINED USING PROPOSED ALGORITHM FOR DIFFERENT POPULATION SIZE AFTER 100 ITERATION WHEN *K*=6

Initial population	Proposed algorithm
5	10.87
10	12.46
15	9.45
20	11.93
100	9.05

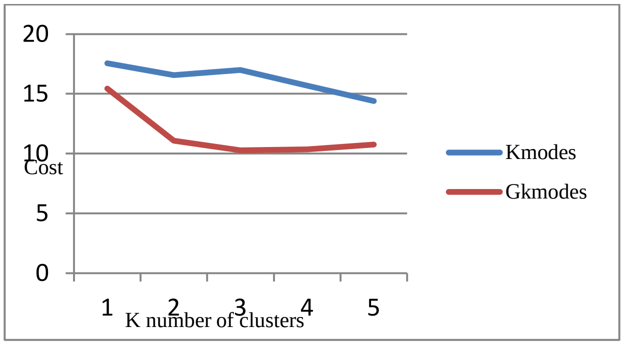


Fig. 3. Comparison of average cost obtained by proposed algorithm and k-modes algorithm for different k after 100 iteration

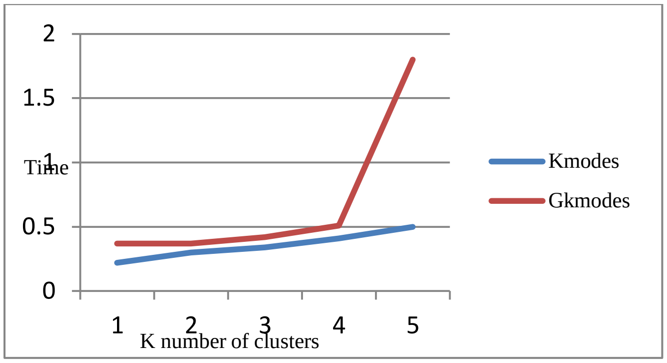


Fig. 4. Comparison of total time obtained by proposed algorithm and k-modes algorithm for different k after 100 iteration

VI. CONCLUSION

Many clustering results are sensitive to the selection of the initial cluster centres as well as gives local optimal solution. The determination of cluster centres in a data set is attracting attention in many research areas. This paper introduced a new variant of GA based clustering for categorical data with the analysis of local and global optimality with *k*-Modes. Existing approaches does not serve as the best method in terms of time and space, Experiments proves noticeably results in terms of cost, within cluster variation, time and initialization of cluster centres.

VII. FUTURE WORK

As this work gives better results for less number of clusters using MATLAB [17]. This can be modify if the number of clusters increased. Proposed method can be compared to more recent algorithms with more number of real world datasets. To discover an algorithm which can perform clustering without knowing cluster number is also a significant work in clustering analysis can be done. And to increase the convergence speed is an important area of future research. Using GA on large number of attributes in datasets need more time and space so latest feature selection techniques [18] can also be applied.

REFERENCES

[1] Han, J. ; Kamber, M. (2001): Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, CA.

[2] Dongxia Chang, Yao Zhao , Changwen Zheng, Xianda Zhang, A genetic clustering algorithm using a message-based similarity measure Expert Systems with Applications, 39, (2012), 2194–2202.

[3] Sneha Antony, Jayarajan J N , T-GEN: A Tabu Search based Genetic Algorithm for the Automatic Playlist Generation Problem, Procedia Computer Science 46 (2015) 409 – 416

[4] Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets inData Mining", 1997.

[5] B. Andreopoulos, A. An, X. Wang, "Hierarchical Density-Based Clustering of Categorical Data and a Simplification", PAKDD, 2007, pp. 11–22.

[6] T. Chen, N.L. Zhang, Y. Wang, "Efficient model evaluation in the search-based approach to latent structure discovery", Proceedings of the Fourth European Workshop on Probabilistic Graphical Models (PGM-08), Vol. 8, 2008, pp. 57–64.

[7] Yinghua Lv, Tinghuai Ma , Meili Tang, Jie Cao, Yuan Tian , Abdullah Al-Dhelaan, Mznah Al-Rodhaan, An efficient and scalable density-based clustering algorithm for datasets with complex structures, Neurocomputing 171 (2016) 9–22.

[8] Arkajyoti Saha, Swagatam Das, Categorical fuzzy *k*-Modes clustering with automated feature weight learning, Neurocomputing 166 (2015) 422–435.

[9] E. David, Goldberg, H. Holland John, "Genetic Algorithms and Machine Learning, Machine Learning", Vol. 3,1988. pp. 95-99.

[10] C. L. Yang, R. J. Kuo, C. H. Chien, N. T. P. Quyen, "Non-dominated sorting genetic algorithm using fuzzy membership chromosome for categorical data clustering", Applied Soft Computing, Vol. 30, 2015, pp 113–122.

[11] H. Qin, X. Ma, T. Herawan , J.M. Zain, "An Improved Genetic Clustering Algorithmfor Categorical Data", In PAKDD Workshops, LNAI, Vol. 7769, 2013, pp. 100–111.

[12] Y. Liu, X. , Y.Shen, "Automatic clustering using genetic algorithms", Applied Mathematics and Computation, Vol. 218, 2011, pp. 1267–1279.

[13] S. Deng, Z. He, X. Xu, "G-ANMI: A mutual information based genetic clustering algorithm for categorical data", Knowledge-Based Systems, Vol. 23, 2010, pp. 144-149.

[14] U. Maulik, S. Bandyopadhyay, "Genetic algorithm-based clustering technique", Pattern Recognition, Vol. 33, 2001, pp. 1455-1465.

[15] G. Gan, Z. Yang, J. Wu, "A Genetic *k*-Modes Algorithm for Clustering", In ADMA, LNAI Vol. 3584, 2005, pp. 195–202.

[16] UCI Machine Learning Repository (2011). http://www.ics.uci.edu_/mlearn/MLRepository.html

[17] Abha Sharma, R. S. Thakur, Cluster analysis for categorical data using MATLAB, International Journal of Research in Management, Science & Technology 2014, Vol2, No. 2, pp 65-68.

[18] Hari Seetha; M. Narasimha Murty; R. Saravanan, Effective feature selection technique for text classification, Int. J. of Data Mining, Modelling and Management , 2015, Vol.7, No.3 , pp.165 - 184.

Improving Throughput and Delay by Signaling Modification in Integrated 802.11 and 3G Heterogeneous Wireless Network

Majid Fouladian
Department of Electrical Engineering, Science and Research Branch, Islamic Azad University,
Tehran, Iran

Mohammad Ali Pourmina
Department of Electrical Engineering, Science and Research Branch, Islamic Azad University,
Tehran, Iran

Faramarz Hendessi
Department of Electrical and Computer Engineering, Isfahan University of Technology
Isfahan, Iran

***Abstract*—Current trends show that UMTS network and WLAN will co-exist and work together to support more users with higher data rate services over a wider area. However, this integration invokes many challenges such as mobility management and handoff decision making. Vertical handoff is switching process between heterogeneous wireless networks in a 3G/WLAN network. Vertical handoffs often fail due to the abrupt degrade of the WLAN signal strength. In this paper, proposed a new vertical handoff method for to decrease the number of signaling and registration processes, to examine the Special conditions such as WLAN black holes and to eliminate disconnecting effects. By estimating user locations related to WLAN position, interface on-times will be reduced which makes the power consuming of the system decreased. To indicate the efficiency of proposed approach the performance and the delay results are simulated.**

Keywords—Registration; WLAN; 3G network; Vertical handover

I. Introduction

Non-homogenous networks are integrated of several networks with various technologies. Vertical handover in these networks are unavoidable because of transparent services to end users. The two wireless networks categories are low bandwidth services over a large geographical region such as 3G and 2.5G and high bandwidth services over a small geographical area like WLANs. Universal Mobile Telecommunications System (UMTS) is a common 3G technology. Quality of service QoS) is one of the most important challenges in these networks which should not experience remarkable changes when handover process is being done. Therefore an efficient algorithm must be employed in order to initialize and complete handovers [1, 2].

Vertical handover are either upward or downward cases. In the downward case the mobile node (MN) enters from a large cell like 3G to a smaller one like WLAN while in the upward case the MN enters from a smaller cell like WLAN to a larger one like 3G [1, 3]. In the upward handover, since the mobile node switches to a cellular network before disconnecting from the WLAN, delay is an important problem while in the downward case delay sensitivity is low. However, in the upward handover, the mobile node should keep its connection with the WLAN to achieve more effective QoS and lower cost. Identifying the decline in signal strength is of high importance when the MN departs a WLAN. In this case, before disconnecting from the WLAN, the mobile node has to take decision when to start the handover to a 3G network. In the downward case, handover is often done in order to promote QoS, but the mobile node should not start handover too soon because it might depart the WLAN quickly and disconnect from UMTS as a consequence of inappropriate handover [4, 5].

The handover process is fundamentally dependent to the network architecture and its software and hardware elements. Various protocols and algorithms have been suggested for this process. They could be classified based on the existing network layers like the application, transport, and network layer. MIP (Mobile IP), which is a handover protocol in the network layer, represents the FA (Foreign Agent) and HA (Home Agent) to address mobility difficulties [7,8]. Synchronization, authorization and updating of user connections and CoA (Care of Address) of the foreign networks are provided by FA and HA.

The most important mobility management protocol in the transport layer is Stream Control Transmission Protocol (SCTP). It was designed to control the signaling in voice over IP (VoIP) networks. SCTP was improved by the Internet Engineering Task Force **(**IETF) in order to support dynamic address reconfiguration during connections. Reliability (as with TCP) and Multi-Homing providing one SCTP session on various interfaces with distinct addresses are supported by this protocol. In SCTP, the primary IP is set as the source address of the transmitted packet if this IP is available, while the secondary IP is set as the source address in situations where the primary IP is unavailable. Therefore SCTP hosts can use primary addresses for primary routing and secondary addresses for alternative routing [10]. In the original version of SCTP, end peers exchanged all IP addresses before starting the connection. Thus it was impossible to change one of these addresses during a session in a non-homogeneous network such as UMTS. In a WLAN, a host may have a fixed and known IP address, so this version of SCTP cannot be employed for vertical handover. However, Dynamic Address Reconfiguration (DAR) in newer versions of SCTP allows IP addresses to be modified during a session via mobile SCTP (mSCTP). Session Initiation Protocol (SIP) is the most pivotal protocol in mobility management through the application layer. As vertical handover (as opposed to horizontal handover), is between different networks with different capabilities, applications must be aware of these changes. Thus application requirements must be considered in the handover process [11].

II. None-Homogeneous Integrated Networks

The European Telecommunications Standards Institute (ETSI) suggested two approaches for connecting 3G networks and WLANs, tight and loose, as shown in Fig. 1. The GPRS network and the packet switched network are connected by the Gateway GPRS Support Node (GGSN). When connection is loose, a distinct path is supplied for 3G traffic, therefore WLAN traffic is not passed through the 3G network and routing to the Internet is through the WLAN gateway. Thus, various protocols for accounting and authentication might be employed by WLANs and 3G networks [12, 13]. When connection is tight, a WLAN is linked to the 3G network in the same way as other radio access networks. It becomes a part of the 3G core and all WLAN traffic is passed through 3G networks. Thus, WLAN gateways use 3G protocols and both the WLAN and 3G network employ similar Authentication, Authorization Accounting (AAA) mechanism. The pivotal merit of the procedure is that security, QoS and 3G mobility mechanisms could be employed in the WLAN. Loose connection is preferred over tight connections Due to the fact that it demands fewer reconfigurations and a simpler structure. Yet tight connections are more desirable in terms of security, performance and QoS. [14].

III. The Proposed Method

Consider the system as shown in Fig. 2. In the proposed method, when the MN gets closer to the WLAN the locations of the MN in the UMTS network are approximated. In a suitable time, the MN gets a message to turn on its WLAN interface. With this technique, the WLAN APs in each SGSN are mentioned as Node Bs for that SGSN, and the UMTS Service GPRS Support Nodes (SGSNs) are mentioned as FAs. Meanwhile the MN enters a WLAN of a UMTS network; the MN is given a new IP address from the WLAN. When the new IP address is allocated, all packets from the Corresponding Node (CN) with the old valid IP address are to be tunneled to the MN by using this new IP address. On the other hand, all packets from the MN should be sent to the CN by using the old IP address, routed to the AP, afterwards via the UMTS network, and eventually forwarded to the MN. First, the process of turning on the WLAN interfaces is discussed and then the protocol will be explained in detail. The black holes influences are also mentioned. Black Holes (BH) are defined as tiny areas where the RSS at an MN declines suddenly so that the link to AP will be broken very quickly. When a BH occurs, in case the MN does not change the connection into the 3G network, it cannot maintain connectivity. By using 3GPP methods categorized into four major groups (cell identifier, Observed Time Difference of Arrival (OTDOA), Uplink Time Difference of Arrival (UTDOA), and Global Positioning System (GPS)) the location of the MN is approximated [16]. Environmental conditions play a critical role in the accuracy of every approach and in the optimum case is as little as 10 meters. Any of these approaches or a combination of them can be applied. During connecting to the MN, the CN demands the MN location from the SRNC through the Iu interface regardless of the employed method. The Iu Interface connects the Radio Network Controller (RNC) with the 3G SGSN. The demanded information from the Location Measurement Unit (LMU) will be gathered by the serving RNC (SRNC) and then the MN location will be assessed. Eventually LMU responds to the CN [16]. Denoted the WUD to RNC unit, a database of user and WLAN information is employed in proposed scheme. The geographical location and general characteristic of the networks like delay, bandwidth and cost are kept in The WUD while for users store an access and priority list of those WLANs they can connect to. When a CN requests the location of a MN from the SRNC, the SRNC calculates the approximate location of the MN and compares it with the stored one in the WUD. In case the difference is lower than a threshold (S_{th}), a message will be sent to the MN in order to turn its interface on. The evaluation is performed according to the following inequality:

$$(x-x_0)^2+(y-y_0)^2<(R-R_{th})^2 \tag{1}$$

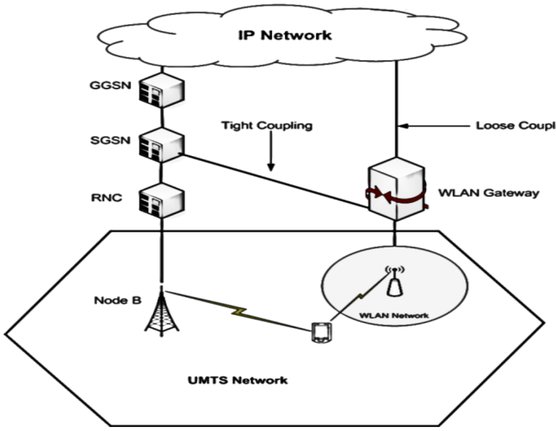


Fig. 1. None-homogeneous networks architecture

Where (x, y) , (x_0, y_0) are the last MN location, WLAN center, and *R* is WLAN radius. Since it is impossible to apply 3GPP estimation when the MN is in a WLAN network, upward and downward handovers have to use different approaches. Therefore in this paper employ two strategies for turning on the UMTS interface, always keep it on, or turning it on while the RSS falls lower a threshold. With the first approach, there is no problem with black holes. With the second approach, black holes cannot be overcome because of increasing handover delay. In the following sections the upward and downward handover techniques are explained.

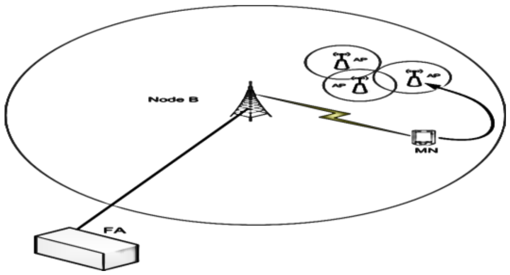


Fig. 2. locating, approximating and turning the WLAN interface on

IV. Downward Handover

It is supposed that the MN is linked to the CN through NODE B and is within the UMTS network. The MN location is assessed periodically by the SRNC and the data will be relayed to the CN. In case the MN gets closer to a WLAN boundary, the MN will be signaled by the SRNC. The MN turns its WLAN interface on and seeks for advertisement messages from closest WLAN APs. Afterwards, the MN determines whether there is any better quality AP than the UMTS network based on a weighted function of the AP parameters. Thus, a network cost function is to be defined. It is noted that the cost function given in [17] is appropriate for proposed protocol which is described as:

$$f^n = W_n = \sum_s \left[\prod_i E^n_{s,i} \right] \sum_j f_{s,j}(w_{s,j}) N(Q^n_{s,j}) \quad (2)$$

where $N(Q^n_{s,i})$ is the normalized QoS, $Q^n_{s,i}$ illustrates the cost in the *j*th network parameter to carry out service "*s"* on network *n*, $f_{s,j}(w_{s,j})$ is the *j*th weighting function for service s and $E^n_{s,i}$ is the *i*th network eradication parameter for service s. The elimination parameter is either one or infinity to show if the current network situations are appropriate for the services which are demanded by the MN. The multiplication over *i* eliminates networks that are not qualified for service *s*, while the summation over *s* considers all services carried by one user. Eventually, the summation over j provides the total cost to network n.

In case the MN cannot find an appropriate AP with a cost function more effective than UTMS it neglects the handover process and keeps the current connection. Otherwise, the shown handover process in Fig. 3 is commenced. The registration request that involves the new CoA address to the FA is sent by AP, and then the FA registers the MN and responds to the AP. Furthermore, the FA sends an update message of binding to NODE B to clean up the dedicated resources. Then, new IP address of the MN that is valid in the AP domain informs the FA. The gained packets of the CN that have the previous MN IP address (that is valid in the UMTS domain) are tunneled to the new IP address by the FA. Since the source address of the packets from the MN to the CN is the previous MN address, these packets should first be routed to the AP through the new IP address and then forwarded to the CN. This process is shown in Fig. 3where W_{ON} is the previous network weight and W_{NN} is the new network, and Fig. 4 indicates the corresponding signal flow where P: 4 is Target network information sending, P: 5 is MN signaling flow, P: 11 to P: 13 are Data path for the neighboring WLAN, and P: 14 to P: 15 are Data path from the MN to CN.

B. Upward handover

Two various cases for upward handover are considered. In the first case, in the WLAN network the MN is linked to the CN. In the second case, the connection is in the UMTS network. Just one upward handover is demanded in the first case whereas two handovers (downward from UMTS to WLAN and upward from WLAN to UMTS) are obligatory in the second case. Afterwards, the handover process is described in detail.

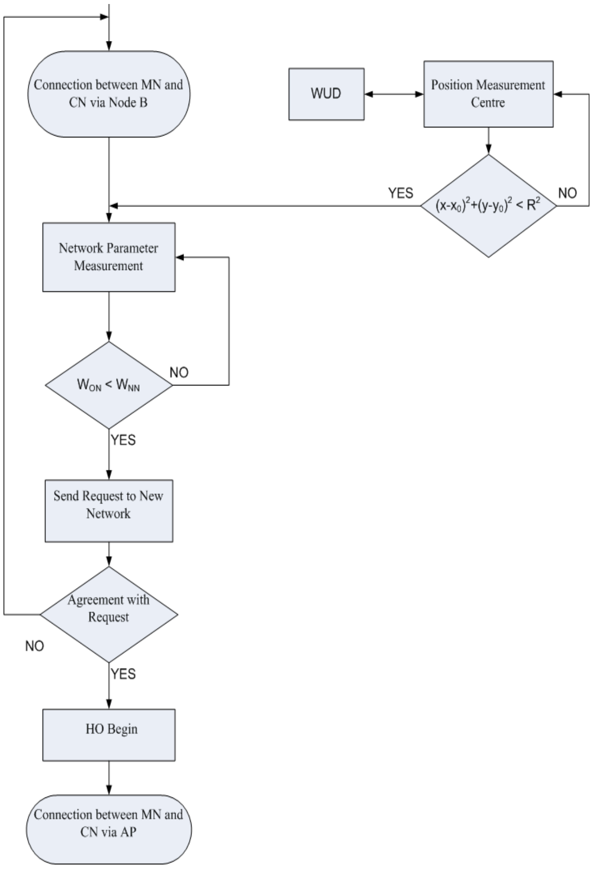

Connection between MN and CN via Node B
WUD
Position Measurement Centre
YES
$(x-x_0)^2+(y-y_0)^2 < R^2$
NO
Network Parameter Measurement
$W_{ON} < W_{NN}$
NO
YES
Send Request to New Network
Agreement with Request
NO
YES
HO Begin
Connection between MN and CN via AP

Fig. 3. The handover process of UMTS to WLAN

1) Upward handover process while the link starts in the WLAN

It is an obligation to continuously assess the RSS and compare it to the threshold S_{th} when the MN is in the WLAN network and linked to the CN through the AP. In case the measured strength is under S_{th} for at least a threshold interval (T_{th}), the MN will send a handover request to NODE B. In case the request is not accepted by the system, the MN is to find another network, otherwise it will be disconnected. Initiating the process, NODE B forwards the request to the FA, which registers the MN and then informs NODE B. Afterwards, all packets sent between the MN and CN are transmitted via NODE B. In addition, the FA sends a binding update message to the AP to clean the dedicated resources up. Fig. 5 indicates the process and the Fig. 6 shows the corresponding signal flow where P: 1 to P: 4 are Data path between the MN and the CN before handover.

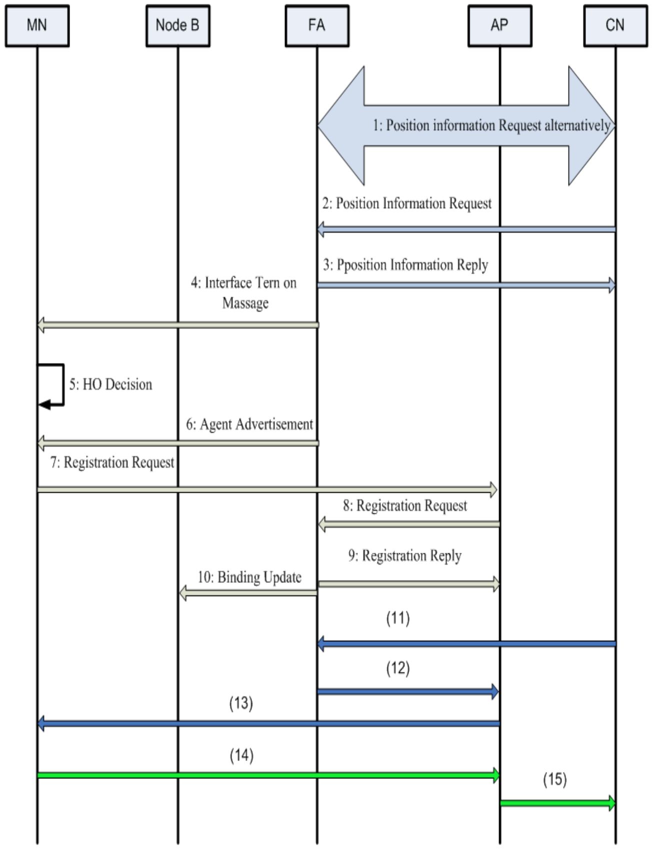


Fig. 4. Signal flow in the proposed downward handover

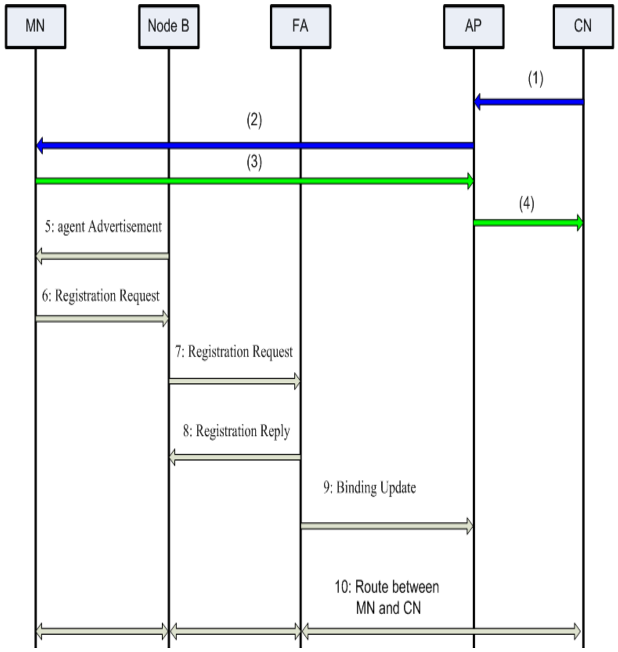


Fig. 6. Signal flow in upward handover while the link starts in the WLAN

2) Upward handover process while the link starts in the UMTS network

In contrast to the previous section, the MN does not require a new IP address and thus the MN is able to resume its connection with the CN applying the old IP address that is valid in the network of UMTS. Two various methods might be used to clean up the resource. After the downward stage the resources might be released instantly, or they might be retained for a particular time to reuse in the following upward handover. This creates a tradeoff between dedicated resources and handover delay. There should be a determination for an appropriate interval for resource reservation in relation to effective reuse by an MN and releasing through an acceptable time for a user who terminates the CN connection in the WLAN. Fig. 7 indicates the signal flow where P: 1 to P: 4 are data path through the CN and the MN before handover.

The above process is held for a WLAN network with a CN connection through the AP as the MN always assesses the RSS and compares it to S_{th}. In case the evaluated strength is under S_{th} for at least a threshold interval (T_{th}), the MN will send a re-association request to NODE B. Handover has the lowest delay in the second method because resources are reserved for the MN whereas there is no significant difference in situation for resource releasing as in Section B.1 Besides, a binding update message is sent to the AP for de-allocation the MN resources by the FA.

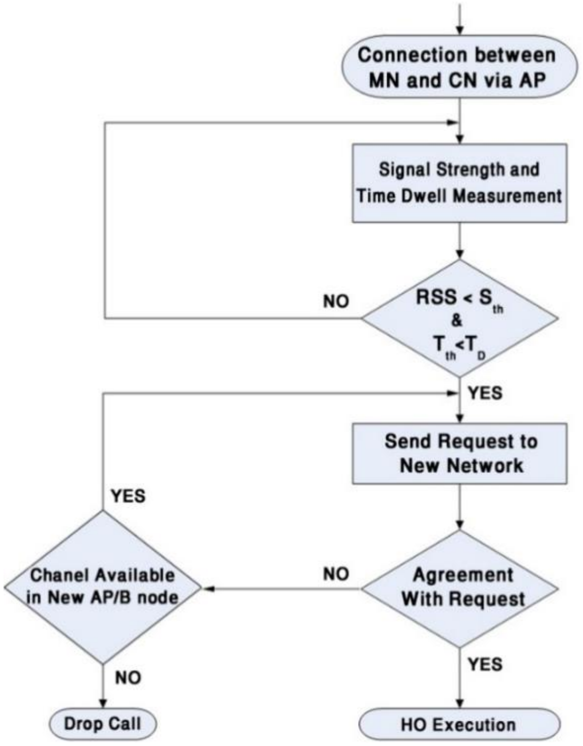

Connection between MN and CN via AP
Signal Strength and Time Dwell Measurement
NO
RSS < S_th & T_th<T_D
YES
Send Request to New Network
YES
Chanel Available in New AP/B node
NO
Agreement With Request
YES
NO
Drop Call
HO Execution

Fig. 5. The WLAN to UMTS handover process

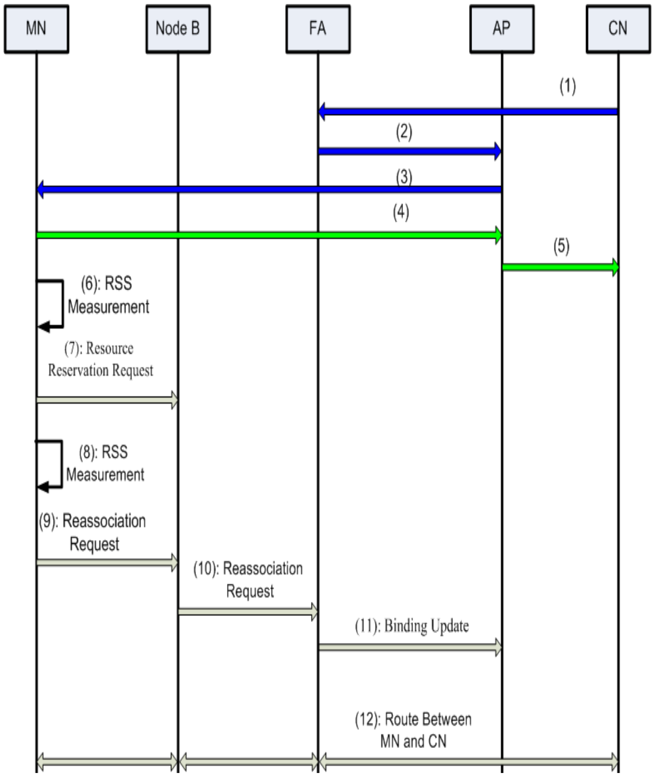

MN
Node B
FA
AP
CN
(1)
(2)
(3)
(4)
(5)
(6): RSS Measurement
(7): Resource Reservation Request
(8): RSS Measurement
(9): Reassociation Request
(10): Reassociation Request
(11): Binding Update
(12): Route Between MN and CN

Fig. 7. Signal flow in upward handover while the link starts in the UMTS network

V. Investigating Boundary Conditions

We have supposed the case that the WLAN is not covered by only one NODE B. When two NODE Bs cover the WLAN two cases are identified according to the number of FAs related to these NODE Bs. These cases are divided based on this issue that the NODE Bs might have a common FA or different FAs. For both cases, downward handover and upward handover when the connection initiates in the WLAN are the same as in section **Ш-B** and **IV-A** section respectively. Yet when the connection initiates in a UMTS network the handover process is different. These cases are discussed in detail in the following sections.

A. Case 1: Two NODE Bs are covered by one FA

As shown in Fig. 8, the MN links to the CN through the first NODE B. Afterwards, it enters to the WLAN by a downward handover and connects to the CN via the AP. Then the MN leaves the WLAN and enters the second NODE B domain. Despite the existence of two NODE Bs, the procedure is the same as the case that there are two FAs. Besides, it is not necessary to have the new IP address of the MN because the protocol is fully transparent to the HA and CN. The signal flow for this protocol is shown in Fig. 9, where P: 1 to P: 5 are Data path through the CN and MN before handover. It should be noted that the simulation results for this part of the protocol illustrate a progress for handover delay which increases the probability of avoiding black holes. The increased success in handover to the 3G network is because of decrease in upward handover delay.

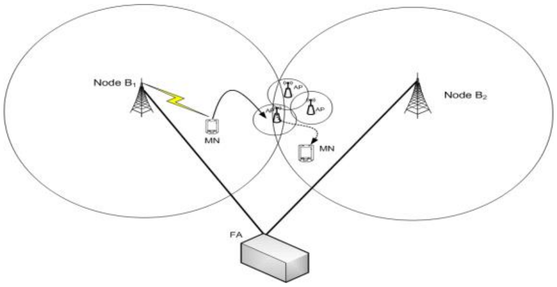


Fig. 8. Handover architecture while the destination and source are covered by one FA

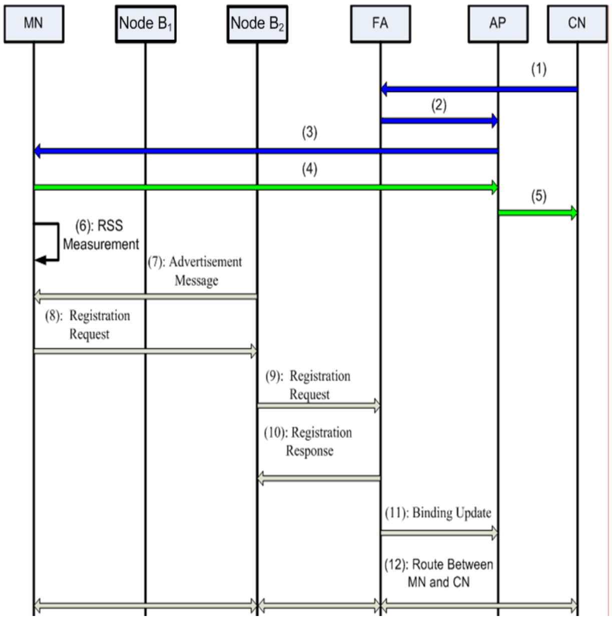

MN
Node B$_1$
Node B$_2$
FA
AP
CN
(1)
(2)
(3)
(4)
(5)
(6): RSS Measurement
(7): Advertisement Message
(8): Registration Request
(9): Registration Request
(10): Registration Response
(11): Binding Update
(12): Route Between MN and CN

Fig. 9. Flow of signal in an upward handover while the destination and source are covered by one FA

B. Case 2: Two NODE Bs are covered by two different FA

As shown in Fig. 10, MN is covered by FA_1 before starting handover whereas MN is covered by FA_2 after handover process. When the MN is in the WLAN and its RSS is under S_{th}, the MN gets an advertisement message from NODE B_2 and afterward responses a handover request to NODE B_2. It is nessecary that the message be forwarded along a path NODE B_2-FA_2-FA_1. Since the mentioned process needs a registration in CN, delay is much more than the previous handovers. Therefore, in order to reduce packet loss and increase delivery rate, it is proposed that FA_1 routes the packets with the MN as the destination through NODE B_1 and then a copy of packets is sent to FA_2 and the packets are stored until new connection is fully established. In case the RSS received from AP goes under the threshold, a registration request is sent by the MN to NODE B_2 which in turn forwards to $FA_{2,}$ and then forwards to the HA. After registration in the HA, the HA returns the responses to FA_2 and then FA_2 forwards them to NODE B_2. At the same time in order to correct the MN address in packets transferred from the CN, the HA is to send an update message to the CN. After getting this message, the FA_2 Node get the new MN address. Next, an update message is sent to the FA_1

by the FA_2 which in turn forwards it to NODE B_1 in order to disallocate the assigned resources of the MN in NODE B_1. After receiving the update message, the FA_1 sends packets with the MN as the destination to FA_2 until the CN registers the new MN address. In order to avoid packet loss due to handover process delay, a message containing the ID of the last received packet is sent to the FA_2 through NODE B_2 which enables the FA_2 to forward packets to the MN. The corresponding signal flow is shown in Fig. 11, Where P: 1 to P: 5 are data path between the CN and the MN before handover process.

VI. Black Holes

As mentioned previously, two strategies are possible for turning on the UMTS interface, always have it on or turning it on when the RSS is under a threshold. In the former case, the challenge is power consumption while in the latter power is saved in favor of accepting the black holes problem (greater packet loss and possible disconnection from the CN). In WLANs, there are often blind spots as a result of obstacles such as elevators and walls which can greatly attenuate signals. This causes packet loss and disconnections during the handover process. Thus by turning the interface on in proposed protocol during the connection, found that the negative effects of black holes are greatly reduced during the handover process.

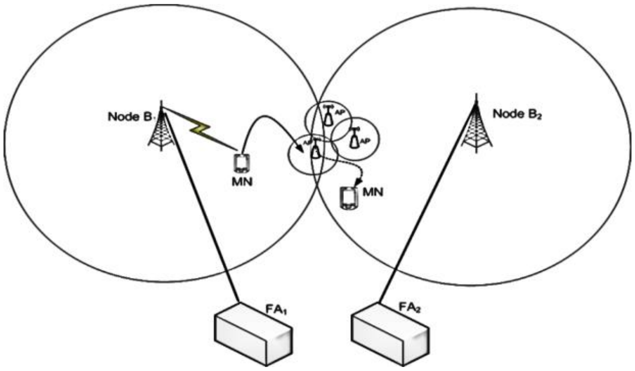


Fig. 10. Handover architecture while the destination and source are covered by two different FA

VII. Performance Evaluation

For simulation the Multi-interface Crass Layer Extension for NS2 (NS-MIRACLE) is employed [18]. For the WLAN network IEEE 802.11b is chosen. A different number of MNs were used in each simulation trial. For each MN, an initial and a final position were randomly selected, and the MN moves along a path between these points with a random speed between 0 to 10 m/s. Every 100 ms, the speed and final position are updated. In order to evaluate delay, 50 MNs were used in the UMTS network and 50 MNs were used in the WLAN network. The S_{th} was set to -75 dBm. Computation of handover delay is the differentiation in time among last packet receipt in the former network and the first packet in the new network.

In figures 13 to 20 simulate proposed scheme handover and Traditional MIP handover and then a comparison is done for them. In all the figures the proposed handover results are shown by green colure and the result of traditional MIP are shown by red colure.

Figs. 13 and 14 illustrate the downward handover delay. Fig.13 is for a tight architecture between the SGSN and AP. There has been a decline in the number of signaling and registration processes because there has been a reduction in utilization of home agents and correspondent nodes. In addition, for mobile users the handover delay and processing time were declined, with a significant improvement relative to the MIP protocol. As the number of MNs increases, utilization of home agents and correspondent nodes, the number of signaling and registration processes will be increased. Fig. 14 indicates the case in which there is no direct wired connection between FA and AP; thus packets among them will be routed via the Internet. Now, a random Internet delay from 0 to 55 ms was applied.

Both loose and tight connections were considered for upward handover. In the first case suppose that when the MN is in WLAN, it connects to CN, afterwards it enters to the UMTS network. Figs. 15 and 16 show the results of simulation for loose and tight connections. In the second case, suppose that the MN connects to the CN through the SGSN address and enters to the WLAN. Then it starts handover to UMTS while there is a connection with the CN. In this case, registration is not necessary and it is sufficient to send a request of re-connection to the SGSN. Fig. 17 shows the results in the case of tight connections and Fig.18 indicates the results in the case of loose connections, respectively.

The results in Figs. 15-18 are for a WLAN that is completely situated within a UMTS network. Fig. 19 indicates the results of simulation when the WLAN is situated between two NODE Bs under one FA for tight connection whereas Fig. 20 shows the results of simulation when the WLAN is situated between two NODE Bs under one FA for loose connection. The maximum WLAN data rate with 802.11b is about 11Mbps which is not accessible by the end users due to overhead such as packet payload and checksum fields. Fig. 21 shows the network throughput for 50 MNs during a 500 ms period.

For the case of black holes, in this paper consider two scenarios when a black hole is passed as indicated in Fig. 12. First, the MN is in the WLAN and passes a black hole which initiates an upward handover to UMTS. This requires registration and obtaining a new address, so the handover delay is greater and network throughput lower than previously, as shown in Fig. 22. Second, the MN connects to the CN through NODE B, and then the MN enters the WLAN via a downward handover. The MN maintains its connection with the CN before entering the black hole, so once the signal strength decreases sufficiently, the MN immediately returns to the UMTS network. This is shown in Fig. 23.

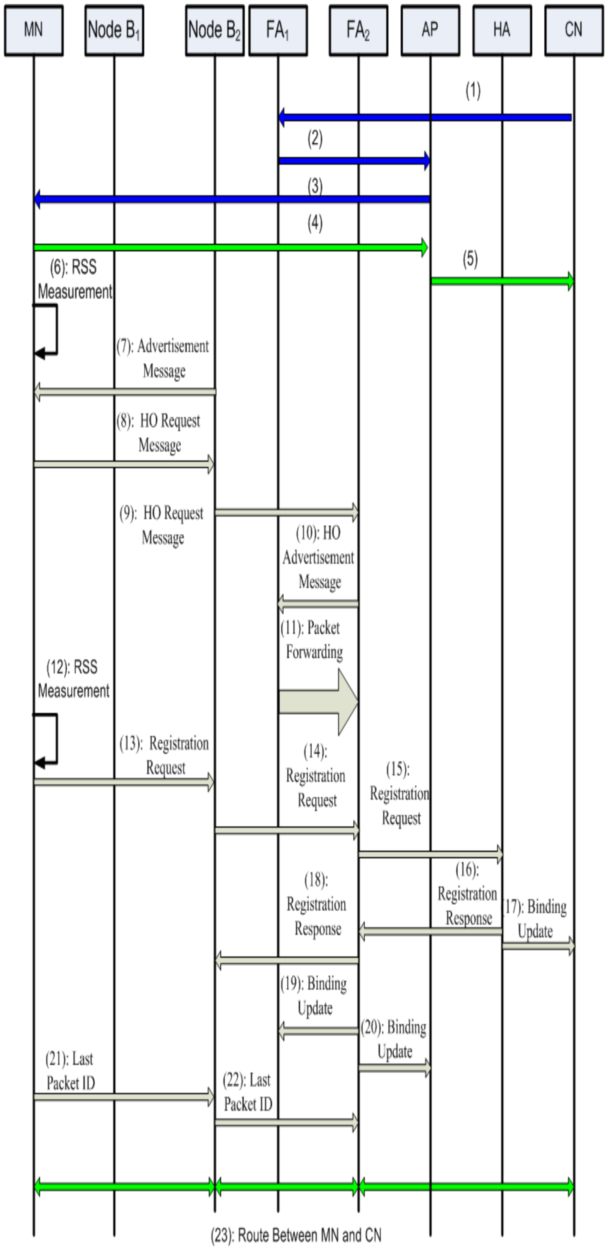


Fig. 11. Upward handover signal flow while the destination and source are covered by two different FA

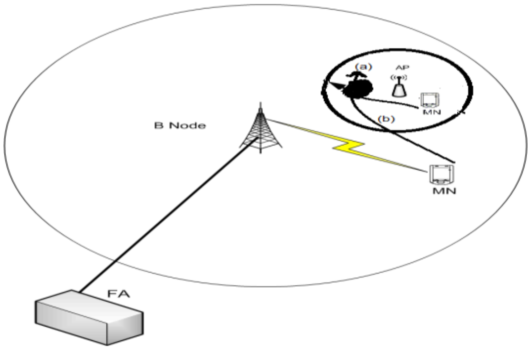


Fig. 12. (a) and (b) Black hole conditions

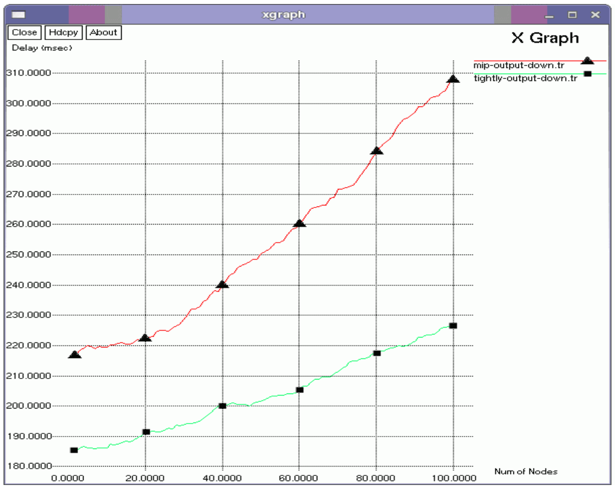


Fig. 13. Delay of downward handover in tight coupling

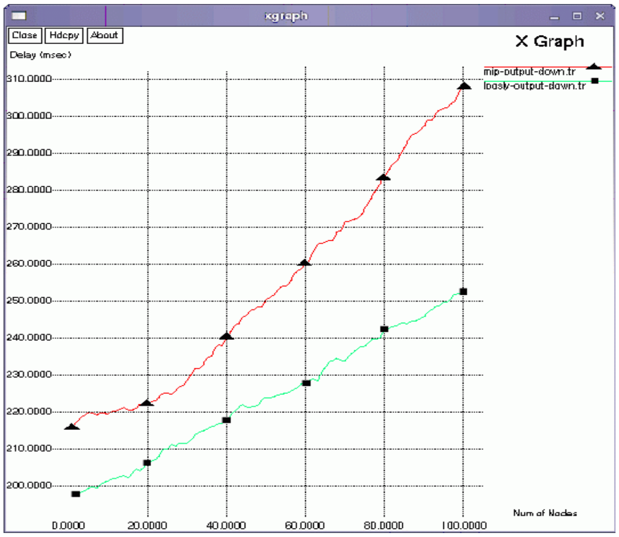


Fig. 14. Delay of downward handover in loose coupling

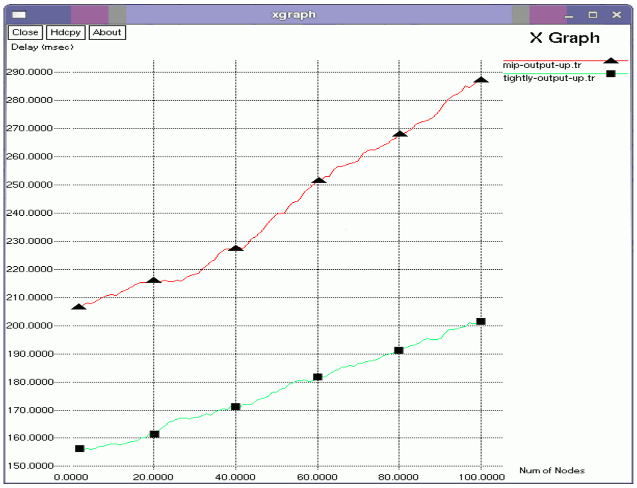


Fig. 15. Delay of upward handover by tight coupling while the connection commences in WLAN

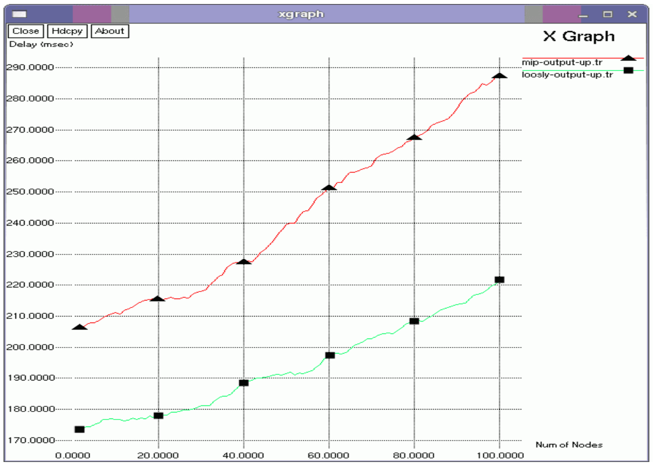


Fig. 16. Delay of upward handover by loose coupling while the connection commences in WLAN

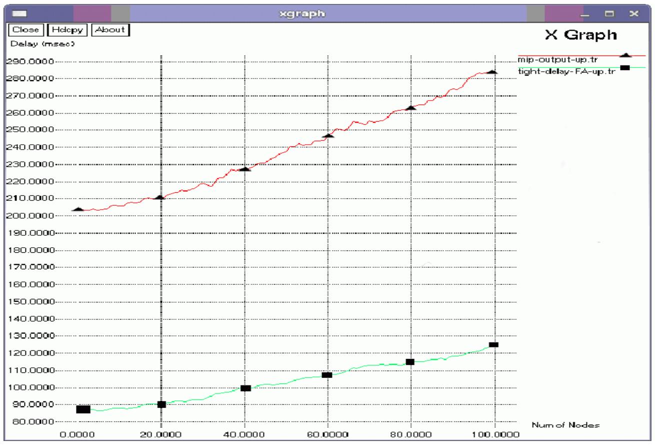


Fig. 17. Delay of upward handover by tight coupling while the connection commences in UMTS

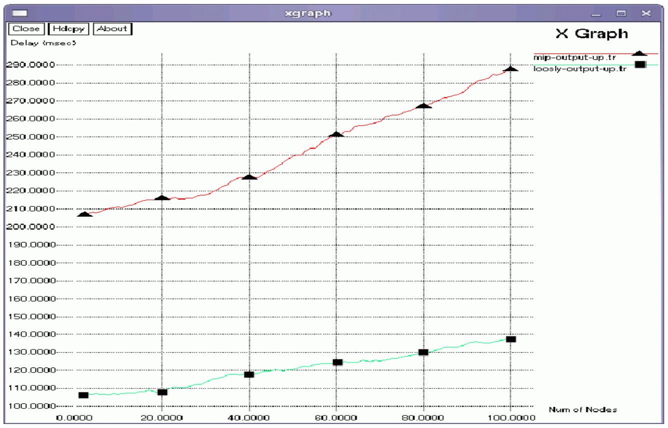


Fig. 18. Delay of upward handover by loose coupling while the connection commences in UMTS

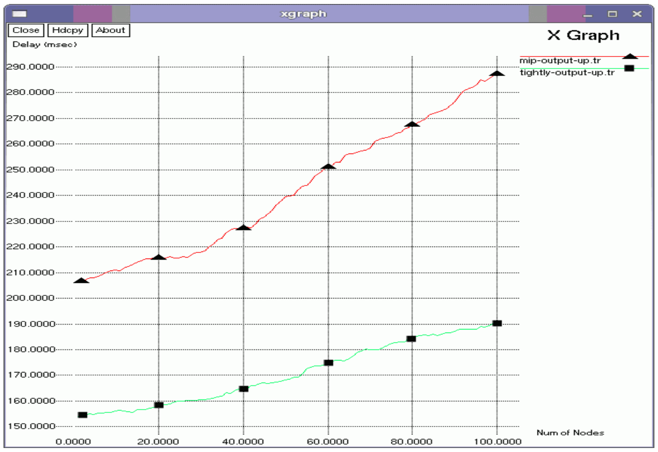


Fig. 19. Upward handover delay with tight coupling and Two NODE Bs are covered by one FA

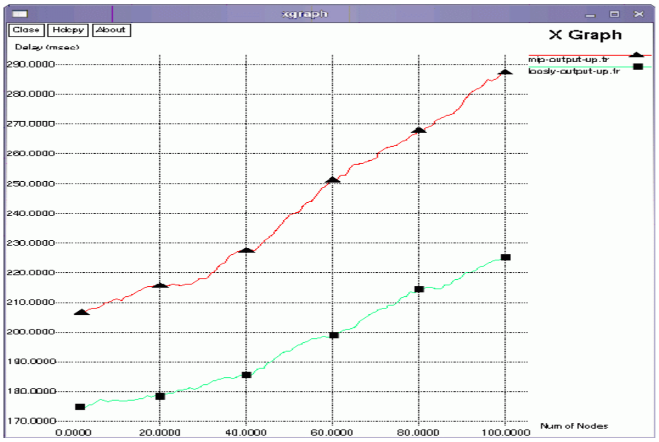


Fig. 20. Upward handover delay with loose coupling and two NODE Bs with one FA

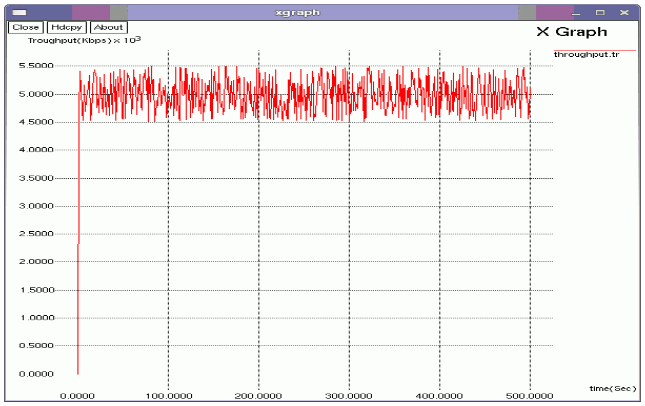


Fig. 21. Throughput with a combination of UMTS and WLAN networks

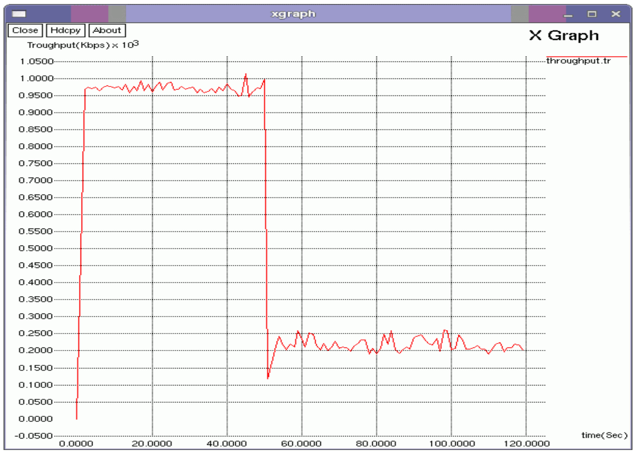


Fig. 22. Blackhole throughput during the existence of WLAN connection

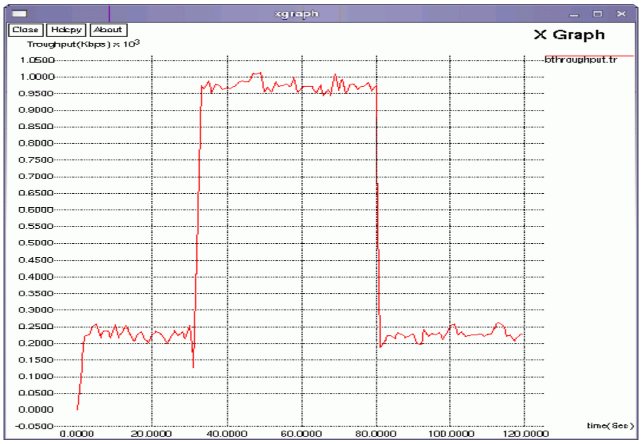


Fig. 23. Black hole throughput during the existence of UMTS connection

VIII. Conclusion

An innovative vertical handover method is proposed that there is no need to modify the WLAN structure. There is just a slight change in UMTS where the SGSN has to tunnel packets to the MN after receiving the new MN address. In the WLAN two MN addresses are used: the first address is assigned to UMTS address that is employed as the source and destination of the transmitted packets and the second one is assigned to CoA address belonging to the WLAN that is employed for routing packets to the CN through the AP during being in the WLAN. The interaction between the HN and CN is declined by this approach. For loose and tight coupling the processing time and delay of handover were declined. There has been a significant reduction in the case of tight coupling where there is a direct wired connection between AP and SGSN. This reduction has led to 38.45% improvement network throughput.

Two general cases were considered for upward handover. In the first case, UMTS keeps turning on while the MN is in a WLAN network. This prevents disconnection due to black holes, but also increases power consumptions, so this approach is not suitable when power is low. In the second case, UMTS turns on only when the WLAN signal strength falls below a given threshold. In this case, there is a higher probability of disconnection due to black holes, but the energy consumption is lower. Reducing energy consumption is one of the most important design criteria of in mobile nodes. Energy of the mobile nodes is supplied from battery. Since, the battery capacity is limited.

References

[1] Ju K., Chen L., Wei H,, and Chen K., “An Efficient Gateway Discovery Algorithm with Delay Guarantee for VANET-3G Heterogeneous Networks” Wireless Pers Commun 77:2019–2036, 2014.

[2] F. Siddiqui and S. Zeadally, “Mobility management across hybrid wireless networks: Trends and challenges,” Computer Commun., vol. 29, no. 9, pp. 1363–1385, May 2006.

[3] S. Kunarak and R. Suleesathira, “Algorithmic Vertical Handoff Decision and Merit Network Selection across Heterogeneous Wireless Networks” WSEAS TRANSACTIONS on COMMUNICATIONS, Issue 1, Volume 12, January 2013.

[4] N. Sattari, P. Pangalos, and H. Aghvami, “Seamless handover between WLAN and UMTS,” Proc. IEEE Vehic. Tech. Conf., pp. 3035-3038, May 2004.

[5] A. Argyriou and V. Madisetti, “A soft-handover transport protocol for media flows in heterogeneous mobile networks,” Computer Networks, vol. 50, no. 11, pp. 1860-1871, Aug. 2006.

[6] S. Busanelli, M. Martal`o, G. Ferrari, and G. Spigoni, “Vertical Handover between WiFi and UMTS Networks: Experimental Performance Analysis“ International Journal of Energy, Information and Communications, Vol. 2, Issue 1, February 2011

[7] 5C. Perkins and R. Glenn, “IP Mobility Support for IPv4”, RFC 2404, Aug. 2002.

[8] K. El Malki, et al., Low Latency Handovers in Mobile IPv4, Internet Draft, Aug. 2005.

[9] T.R. Henderson, “Host mobility for IP networks: A comparison,” IEEE Network, vol. 17, no. 6, pp. 18-26, Nov.-Dec. 2003.

[10] L. Ma and F. Yu, “A new method to support UMTS/WLAN vertical handover using SCTP,” Proc. IEEE Vehic. Tech. Conf., pp. 1788-1792, Oct. 2003.

[11] N. Banerjee, K. Basu, and S.K. Das, “Hand-off delay analysis in SIP-based mobility management in wireless networks,” Proc. IEEE Int. Parallel and Distributed Processing Symp., 8 pp., Apr. 2003.

[12] H.-Y. Hsieh, C.-W. Li, S.-W. Liao, Y.-W. Chen, T.-T. Tsai, and H.-P. Lin, “Moving toward end-to-end support for handovers across heterogeneous telephony systems on dual-mode mobile devices,” Computer Commun., vol. 31, no. 11, pp. 2726-2738, July 2007.

[13] A.K. Salkintzis, “Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks,” IEEE Trans. Wireless Commun., vol. 11, no. 3, pp. 50-61, June 2004.

[14] F. Tansu, M. Salamah “Vertical Handoff Decision Schemes for Heterogeneous Wireless Networks: An Overview” Recent Trends in Wireless and Mobile Networks Communications in Computer and Information Science, Volume 84, 2010, pp 338-348

[15] X. Yan, Y. A. Sekercioglu, and S.Narayanan. “A survey of vertical handover decision algorithms in 4G heterogeneous wireless networks” Elsevier Computer Networks, 54(11):1848–1863, August 2010.

[16] 3rd Generation Partnership Project. 3GPP. Website: http://www.3gpp.org

[17] J. McNair and F. Zhu, “Vertical handover in fourth-generation multinetwork environments,” IEEE Trans. Wireless Commun., vol. 11, no. 3, pp. 8-15, June 2004.

[18] N. Baldo, M. Miozzo, F. Guerra, M. Rossi, M. Zorzi, "Miracle: The Multi-Interface Cross-Layer Extension of ns2" , EURASIP Journal on Wireless Communications and Networking, Vol. 2010, pp,16, Jaunary 2010 .

Constructing Relationship Between Software Metrics and Code Reusability in Object Oriented Design

Manoj H.M
Research Scholar
Jain University, Bangalore.
Karnataka, India

Dr. Nandakumar A.N
HoD
Information Science & Engineering,
New Horizon College of Engineering,
Bangalore, India

***Abstract*—The role of the design pattern in the form of software metric and internal code architecture for object-oriented design plays a critical role in software engineering regarding production cost efficiency. This paper discusses code reusability that is a frequently exercised cost saving methodology in IT production. After reviewing existing literature towards a study on software metrics, we found that very few studies are witnessed to incline towards code reusability. Hence, we developed a simple analytical model that establishes a relationship between the design components of standard software metric and code reusability using case studies of three software projects (Customer Relationship Management project, Supply Chain Management project, and Enterprise Relationship Management project). We also testify our proposal using stochastic based Markov model to find that proposed system can extract significant information of maximized values of code reusability with increasing level of uncertainties of software project methodologies.**

***Keywords*—Analytical Modeling; Code Reusability; Design Pattern; Software Methodologies**

I. Introduction

In today's world, every sector of industry or services is dependent on the computer-based applications. To improve performance and gain a competitive edge, quality of the software has become a crucial factor. Developing and outsourcing of software service is a major and rapidly growing industry in many parts of the world [1]. The process of software development is described through the term software engineering that refers to the usage of a systematic procedure in context to a standard set of goals for performing analysis, designing, implementation, and testing as well as maintenance of the software. The software thus developed must be reliable, usable, efficient, modifiable, testable, maintainable, interoperable, portable and accurate [2]. In the process of software development, object oriented design is considered as one of the important features to evaluate the quality of software [3]. Irrespective of the size of the organization, the object-oriented design methodology is majorly adopted for software development in any organization. Therefore, object oriented designs are considered as standard through which system objects can have particular features and also necessary characteristic. The reason behind the adoption of object-oriented methodologies is that it allows to visualize the problem and acquire a solution in all macro and micro level in related to objects and also ensuring better reliability, adaptability, flexibility and reusability [4]. Currently, software engineers use software metrics to evaluate design component and necessary resources of a certain software project. The advantage of software metric is that it allows the evaluation of design pattern through the better platform as well as assistance in performing the testing of application in a quantitative manner. Through such testing, the reliability of the software can be demonstrated. In general, when a company receives a new requirement from a client, initially they will formulate a design of the requirement. This confirmed design from the architect was sent for production. On completion of coding, the product is dispatched to the customers. Although it is unethical to reuse the code of earlier client to develop an application for the new client [5], code reusability also deals with security of intellectual property. A production team has to go for a fresh development starting from the scratch, which not only requires effort but also a considerable amount of time and money. The majority of the large organization now use design pattern which is subjected to reuse without the ethical issues. The main objective of the design reuse is to assist the developer to use it in the new production, which helps in cutting down the new development cost starting from scratch. However when reusing the design care should be taken so that the design is optimally reused for the current as well as the future client. Adopting design reuse will also help in ensuring the timely production and delivery process. Design reuse does not imply that the entire design is used; it might be like some percentage of the current design is used in the new or future similar project. Hence, the focus of the design team should be such that, the design is not only focused on existing client, but it should be able in providing a minimum proportion of reusable design for future clients also. But in reality, it is not easy since the future requirement of the client is unpredictable.

In Section II gives an overview of the software metrics. Section III highlights about the CK Metric key points. Section IV discusses the related work. Followed by Problem identification in Section V. Section VI discusses proposed system. Section VII discusses the research methodology. Section VIII provides the outcome and result in the analysis of the proposed system, and Finally Section IX provides conclusion and future work.

II. About Software Metrics

In this era of IT revolution, software development as emerged has a crucial requirement in every phase of the day to day life. From academics to public service, healthcare to banking, entertainment to sports, the use of the software is involved in one or other way contributing to the advancement of the domain. Various factors that have contributed to the advancement of software development are its key features like flexibility, portability, design reusability, software metrics and so on. Design reusability plays a vital role in software development, since it is not only involved in the development of current work but also lays a blueprint for future requirement of the current as well as new projects. Design reusability will also help in reducing the cost of production and reducing required manpower and development time by providing a framework that consists of the design which can be reused over a period. Another significant factor in software development is the software metric, which defines the standard degree of measure to which the process or software system will possess certain property. In software engineering, different metrics are available to measure different parameters such as process is measured using process measurement, a project using project measurement and product metric to measure product metrics.

In our work, since object –oriented deign is involved the paper will highlight few things related to Object-oriented metrics. To develop metric for the object-oriented design, seven different measurable qualities are listed below [6].

- Complexity: Analyzed by assessing the way classes are related to each other.
- Coupling: It is the physical connection between the object-oriented elements.
- Sufficiency: Its defines the degree to which the abstraction should possess the features needed by it.
- Cohesion: It is determined by analyzing the group of properties posses by the class being the part of problem domain or design domain.
- Primitiveness: Used to indicate the degree of atomic level of operation.
- Similarity: It is used to identify the similarity between the classes in the term of behavior, structure, purpose or function.
- Volatility: used to define the probability of happening of change in object oriented design.
- Size: Using four different perspectives such as volume, length, population and functionality.

The measurement of the population is done by evaluating the total number of OO entities, which is in the form of classes or operations. Measurement of the volume is achieved dynamically at any instance of time. Functionality denotes the value provided to the user by the object oriented application. Using inter connected design like the depth of inheritance tree length is measured. Object oriented design metrics concentrates on measurement that applies to class and design characteristic. Through these measurements designers are permitted to access software in early process phase, allowing making changes that minimize complexity and enhances the continuing ability of design [7].

III. CK METRIC

In 1994, Chidamber and Kemerer introduced standard software metric for object-oriented programs. CK metrics plays a vital role in knowing the design aspects of software and improving the software quality. The main objective of the CK metric is to provide an in detail measurement of cumulative quality of the software programs to every class level. Metric is associated with each and every tiny segment of the software providing the overall information of the software quality. In CK metric, six classed based metric for object-oriented programs as follows:-

1) *Weighted Methods per Class (WMC)*
2) *Reponses for a Class (RFC)*
3) *Depth of Inheritance Tree (DIT)*
4) *Number of Children(NOC)*
5) *Coupling between Object Classes (CBO).*
6) *Lack of Cohesion of Methods (LCOM)*

1) Weighted Methods per Class: Used to define, the sum of complexity in class. In whole it represents the complexity of the class and this measure can be utilized for indicating development and maintenance effort for class.

2) Response for class: This metric represents a number of growing methods within a set, which can be called in response to a message sent to an object performing the certain task.

3) Depth of Inheritance Tree: This is one of the frequently used metrics; it is used to estimate the extent of depth in the hierarchy of class. It is also used in evaluating maintainability and reusability.

4) Number of Children: This is a measure of number classes that are associated with a particular class with the assistance of inheritance relationship. Class with many children implies a bad class with bad design.

5) Coupling between Object Classes: This defines the number of all another set of classes for which it is coupled. CBO is advantageous in determining the complexity in testing and reusability.

6) Lack of Cohesion of Methods: It is the difference between the number of methods in which the similarity is zero and the number of methods in which the similarity is non-zero. The similarity between the two methods is the number of features that is being used in common. A zero in LCOM does not signify cohesiveness as well a high value does not represent any inference. In LCOM, it is difficult to define a unit and to measure quality. LCOM is not recommended for accurate measurement since it does not quantify quality properly.

Object-oriented metrics are being successfully used in different domains and programming languages in different parts of the world. These metrics have consistently illustrated the relationship to quality factors like reuse, defects, cost as well as maintainability and relationship which may go beyond the size. The evaluation of these metric is achieved through certain tools like metal mill [8], metric 1.3.6 [9], Analyst 4J [10], OOmeter [11] and Dependency Finder.

IV. Review of Literature

This section discusses the prior literature where various approaches of metric software design and its contribution have been introduced. Many approaches have been developed over the years to address the problem of detecting and correcting design flaws in an Object Oriented (OO) software system using metrics. Moreover, with the ever-increasing number of software metrics being introduced, the project managers find it hard to interpret and understand the metric scores. As Object Oriented Metrics require a very good understanding of Object-Oriented concepts and moreover, there is no single metric present which gives all features of Object-Oriented Software System. Table.1 shows the existing survey on a design flaws in an object-oriented (OO) software system using metrics. Table 1 will highlight the tabulated discussion of various problems and respective techniques used to enhance the performance of software metrics in software engineering.

TABLE I. Existing Studies Towards Software Metrics

Authors	**Problem Focused**	**Techniques used**	**Performance parameters**
Basili et al. [12]	examined the suite of Object-Oriented proposal metrics presented by Chidamber & Kemerer.	Empirical validation	C&K OO metrics gives better predictor than conventional metrics.
Anna et al. [13]	to measure the reusability and maintainability degrees of aspect-oriented systems.	Empirical and quantitative analysis, Aspect-oriented software development (AOSD).	Degrees of complexity, diverse domains.
Kaur et al. [14]	To exploring structural factors for software components.	Software metrics using neural networks.	exhibit an efficient model targeted for software programmers.
Kaur et al. [15]	classification and assessment of various reusability metrics.	K-Nearest Neighbor-based technique.	
Kumari et al. [16]	To compare C++ and Java programs.	Statistical inference techniques, Object-oriented metrics.	More realiability.
Linda et al. [17]	To analyze the different metric suites for object oriented schems.	Development of a software prototype like "Class Break point Analyzer (CBA)"	Software quality, realiability.
Wu Et al. [18]	To do comparative analysis on on C++, C#, and Java programs by using object-oriented Metrics.	Comparative study on software metric reusability in software engineering.	Reusability and Realiability.
Srinivasan et al. [19]	To analyze and reviews the most referred object-oriented metrics in software measurement.	Done Comprehensive Review on object oriented metrics using for software measurement.	
Scotto et al.[20]	To evaluate the effectiveness of the metrics.	Web Metrics for SQL queries.	
Singh et al.[21]	To estimate the reuse of software matric.	LMT (Logistic Model Trees).	reusability
Subramanyam et al. [22]	To reduce the computational complexities in object-oriented programs for identifying defects	object-oriented programming	Good enhancement, complexities.
Shaik et al. [23]	To itemize and maintenance of software.	Object Oriented Software Systems	The effort, different metrics
Zahara et al. [24]	To examine the competence and effectiveness of machine learning regression techniques.	Multi-linear regression, "Standard instance-based learning IBk with no distance weighting"	Software crisis, software quality, productiviey.
Manoj et al. [25]	To contemplate about software metrics.	An extensive literature survey on ranking code reusability in software engineering.	Cost effecitve, quality, design complexity.
Oberoi et al.[26]	Analyzing CK metric values of component-based software	Software component design patterns, Self-Organizing Map and empirical evaluation.	Optimized metric values.
Goyal Et al. [27]	To find out the reusability value for software.	Unsupervised neural network.	Reuability, complexity,
Jayalakshmi Et al. [28]	To measure quality in terms of software performance and reliability.	Functional parameters and non-functional parameters.	complexity, reliability and robustness of the

			software
Hudly Et al. [29]	To evaluate the main features of object oriented like Polymorphism, Encapsulations, Data abstraction, Inheritance and classes.	Two kinds of metrics: classes and to measure the class design configuration of the program.	complexity, reliability
Liu Et al. [30]	The gap between quality measurement and design of these metrics.	Object Oriented Designs software.	Safety, complexity.
Paliwal Et al. [31]	To increase the productivity and maintainability of any software.	Reusable module.	Module Reusability, Dependencies, Class size.
Chauhan Et al.[32]	to assess software quality at design level.	14 Java Parser, Eclipse with Metrics 1.3.6	quality of software.
Gupta et al. [33]	a comparative study of many software quality estimation methods.	Fuzzy Logic techniques, artificial neural network (ANN),	Better quality of software
Alcalá et al. [34]	Improving the performance and flexibilyzing the model structure.	Linguistic model structure using weighted double-consequent fuzzy rules.	complexity, reliability

V. Problem Identification

Software engineering has played a significant role in successfully delivering the quality-oriented project. The existing literature discussed in prior section discusses various techniques for enhancing the crude performance of software metrics. It is found that majority of the techniques uses quantitative, empirical, tree-based techniques. Some of the unique evolutionary techniques e.g. neural network, fuzzy logic etc. has also been used. All these above techniques have possible advantages as well as disadvantages too. This section

We will discuss the problems being identified after reviewing the existing system as follows,

A. Lack of Benchmarked Research

Except CK-metrics introduced during the 90s, we have not come across any robust discussion of software metrics, especially for object-oriented programming. Although there is presence of massive volumes of papers in internet media, few authors have been found to introduce any novelty in their ideas. Some of the ideas were to implement CK metric or introduce a new mathematical formula over the old equations of parameters e.g. DIT, WMC, NOC, etc. Also, we have not come across any research model which was found to use or followed by other researchers much towards code reusability.

B. No studies towards CK Metric Relationship

100% of all the papers introduced towards software metrics suites have their formulations. The authors normally check for problems associated with CK metrics and attempt to introduce new software metric suite. However, there was no significant attempt in the past to investigate the relationship among the CK metrics towards code reusability, which is very common operation in any IT industry or any production company.

C. Lack of focus on Code Reusability

Code reusability is the common practice in any IT production. However, it has received very less attention among the research communities worldwide. Studies towards code reusability on its possible relationship with the software metrics are a less-explored topic.

These above problems are addressed in this paper in the form of a simple formula with an aid of case study. The next section discusses the proposed system that enhances the performance of the code reusability and enables the user to visualize enough statistics between software metrics and code reusability.

VI. Proposed System

The proposed study aims to develop an analytical framework which establishes a relationship in between different type of CK-Metrics components with code reusability. The proposed system includes two different type of experimental prototyping which are i) modeling for the estimation of code reusability and ii) frame work to evaluate the impact of design metrics on code reusability. The proposed study has been highly motivated by all the studies that have been carried out in past which represents that the improvement in the software quality can be achieved through performing efficient code reusability, and implementation of various measurement is driven modules. The modeling for estimation of code reusability index considers a flow of processes where CK metric data from design and code artifacts and domain knowledge and experience are further processed through empirical analysis. The CK-Metrics data can be acquired from UML (Unified Modeling Language) class diagrams or the equivalent Java codes. There various processing tools like Rational Rose etc, which are used to extract the metric data from code artifact's and anticipated to bridge a relationship in between the CK-Metrics components such as WMC (Weighted Methods per Class), DIT (Depth of Inheritance Tree), NOC (Number of Children), CBO (Coupling Between Object Classes), RFC (Response Set for Class), LCOM (Lack of Cohesion in Methods) etc. The empirical analysis generates data for the influence of property on code reusability and establishes a relation between various multifunctional estimation equations belong to the different type of CK-Metrics components and maps them with code reusability. Finally, a software framework for estimation of code reusability has been implemented using GQM paradigm and weighted factors from design metrics. The code reusability model has been further processed through a framework which calculates the different type of CK-Metrics components using static class diagram and dynamic sequential diagrams. The proposed model also evaluates a data related empirical analysis and the analysis further generates the influence of individual metrics on code reusability, the linear combination of coefficients on individual design metrics. The proposed system also takes CK-Metrics data as input and evaluates the code reusability index which can be further mapped into CK-Metrics components during a software development lifecycle. There are various existing conventional studies which are based on code reusability design metrics using empirical prototyping. The relationship between various CK-Metric components and the code reusability has been developed based on framing following conventional myth as follows,

1) It can be seen that classes with higher Depth of Inheritance Tree values will have a higher probability of code reusability.

2) Classes with low cohesion result better software design and code reusability.

3) Class which consists of higher values of WMC and NOC extends the code reusability in the dynamic scenario of software development.

4) Classes with higher CBO and RFC values increase the computational complexity and maximize the code reusability.

The proposed model also introduces a framework for code reusability which has been evaluated using deep empirical analysis and data modeling. The Empirical model considered two different types of medium-high-level projects where an experimental analysis has been carried out considering a huge number of classes to investigate the code reusability of the designed metrics. The classes associated with each project configured and grouped regarding different metrics values to avoid the intellectual property issues. The proposed study developed an effective and computationally efficient framework. The contribution of the proposed study includes i) ensuring the estimation of code reusability on heterogeneous object-oriented software modules, ii) calculating the linear combination of weighted polynomial equations, iii) formulating an efficient relationship in between the CK-Metrics components and the code reusability. The performance metrics associated with the empirical model has been evaluated which ensures the effectiveness of the proposed system. The next section will discuss the research methodology which formulates the relationship between design quality metrics and code reusability in detail.

VII. Research Methodology

The proposed system is designed with an aid of analytical modeling approach as a standard of research methodology. The design of the proposed system is based on code reusability concept using the CK-metric suite. Fig.1 highlights the modeling of the code reusability where the extraction of CK metric data is done from UML. Multiple forms of Java-related cases can be used for obtaining CK metric. Various industry-related tools like Metamill, Metric 1.3.6, rational rose, etc. can be utilized in this regards to get the components of CK-metrics. However, class diagram can be manually used for estimating the number of classes. There are various parameters that can be evaluated with an aid of static classes e.g. DIT, WMC, and NOC, whereas various forms of the sequential diagram can be used for evaluation other metrics e.g. RFC and CBO. However, it is quite imperative that LCOM couldn't be assessed or evaluated from the design patterns e.g. UML directly. However, sophisticated industry-based automated tools can be used for the same reason.

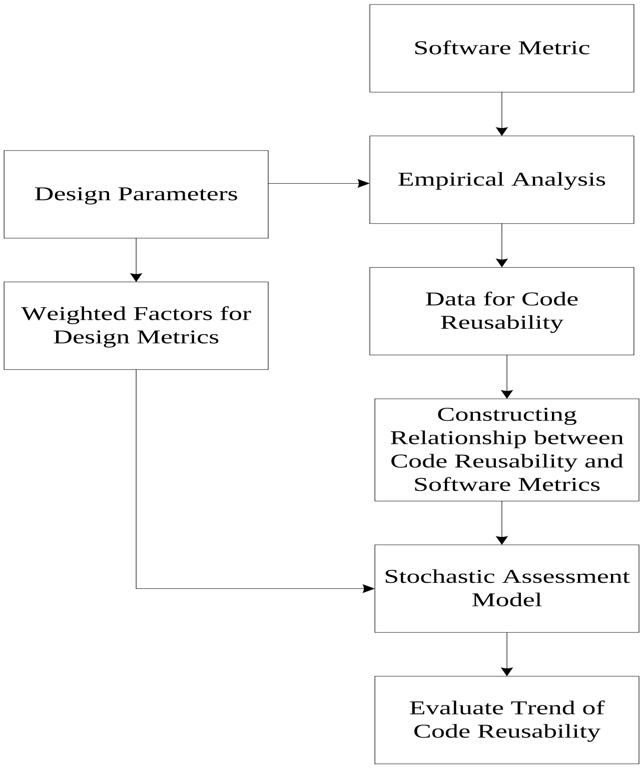

Software Metric
Design Parameters
Empirical Analysis
Weighted Factors for Design Metrics
Data for Code Reusability
Constructing Relationship between Code Reusability and Software Metrics
Stochastic Assessment Model
Evaluate Trend of Code Reusability

Fig. 1. Schematic Diagram of Code Reusability

To assess such design-related issues i.e. code reusability, we consider sample projects developed in Java with a significant number of classes. The proposed system targets to understand how individual components of CK-metrics affect code reusability. We develop a simple function to establish a relationship between metrics and code reusability. We use the concept of weighted coefficient as well as a linear approach for assessing the possible impact analysis of CK-metrics over code reusability. We also use the concept of GQM (Goal-Question-Metric) as the core part of the research methodology that allows formulating the conceptual level of code reusability based on operational level and quantitative level.

VIII. Hypothesis Design

As discussed earlier that proposed study intends to understand the underlying relationship between components of CK metrics with code reusability, hence, an appropriate hypothesis is constructed for this purpose. The study performs analytical assessment on various ERP (Enterprise Resource Planning) and SCM (Supply Chain Management) related software projects on Java and following null hypothesis is being constructed.

- H_{o1}: Better code reusability can be retained by moderate value of DIT in every class.
- H_{o2}: The complexity in code design and reusability decreases for maximized values of RFC.

- H_{o3}: Code reusability degrades for increasing values of NOC present in class.
- H_{o4}: The hypothesis is increased CBO values doesn't have much impact over code reusability
- H_{o5}: Code reusability declines for increasing values of WMC in every class.
- *Developing Analytical Modelling :* The development of the proposed system is carried out using analytical modelling approach for testing the hypothesis. The system also applies simple mathematical estimation techniques to investigate the possible relationship between the components of CK metrics and code reusability. The proposed system considers a case study of ERP and SCM related software projects developed in Java. The total number classes considered for ERP software project is 220. There are around 180 bases classes in it. Similarly, there are around 570 total classes and approximately 380 maintainable classes for SCM software project. Although, our technique could include number of software projects, we choose to consider using only ERP and SCM software projects. The components of the CK metric have multiple values which can be arranged or structured more appropriately. The proposed system performs simple analytical modeling for code reusability in the form of,

$$C_r = \frac{\beta \times 100}{\alpha} \quad (1)$$

The above equation (1) represents mathematical representation of code-reusability, where α is considered as total amount of classes available in every CK metric, whereas β is the component of CK-metric that corresponds to amount of classes newly designed by incorporating code reusability. This will mean that higher the value of C_r, higher is the extent of code-reusability. The evaluation of the α and β parameter is carried out manual as well as recording the same over spreadsheet. However, the values considered for discussion in result analysis is approximated to get contrastive outcome for investigating the effect of components of CK metrics over code reusability in software engineering. Finally, the analytical modelling is also testified with respect to presence of uncertainty.

IX. Result Analysis

The analysis of the proposed study was carried out over SPSS [35] tool. We perform both numerical analysis as well as graphical analysis to assess the effectiveness of the outcome. Table 1 shows the numerical outcomes of the considered case study of ERP and SCM software projects, where the necessary CK metric components were closely observed and computed for code reliability using the simple equation (1) illustrated in prior section. Following are the discussion of the graphical outcomes of proposed system.

A. *Effect of DIT on Code Reusability:-*

The outcome shown in Fig.2 highlights that there is a significant improvement of code reusability for the DIT values. A closer look at the numerical values of DIT shows that with increasing trends in the value of parameters α and β, the code reusability enhances significantly. The increasing number of levels of DIT will represent a little bit of complexity in computation; however, it is productive to make the code more reusable. It is also suggested that for massive objective oriented design, it is quite possible that a number of classes drastically increases, which also increases the possibility of DIT values. However, using the proposed system, using moderate values of DIT can retain better trends in code reusability. Therefore, the null hypothesis stating that "Better code reusability can be retained by the moderate value of DIT in every class" is found to be accepted and true."

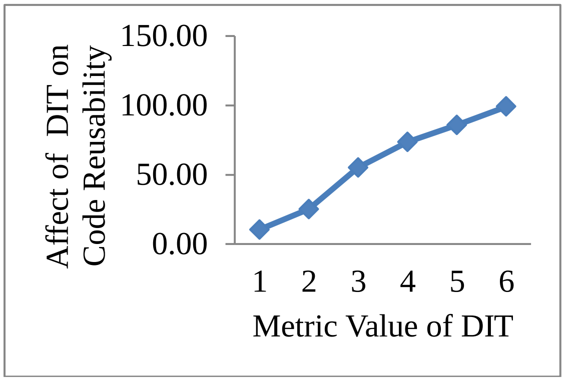


Fig. 2. Affect of DIT on Code Reusability

B. *Effect of RFC on Code Reusability:-*

Fig.3 shows some interesting trends of code reusability. Although, the trend of Cr is found to be increasing the trend is not smooth enough in the preliminary values of RFC metric. The basic trends explored here is that with an increase of α and β parameters has witnessed enhanced code reusability factor. However, it also brings complexity in the major ranges of metric values (1-7), which will mean that although code reusability increases it also brings significant complexity over design. For our analysis purpose, we use two forms of data to generate the graphical outcomes. The first data is synthetic data and can be seen in the RFC row in Table 1, where the parameters α and β are maintained in a combination of both increasing and decreasing order. The outcome shows an increase of code reusability. However, it is less likely that for complex ERP and SCM projects, the values of α and β are usually in increasing order. Hence, we plot a graph by using the increasing order to α and β to investigate the possible impact on code reusability.

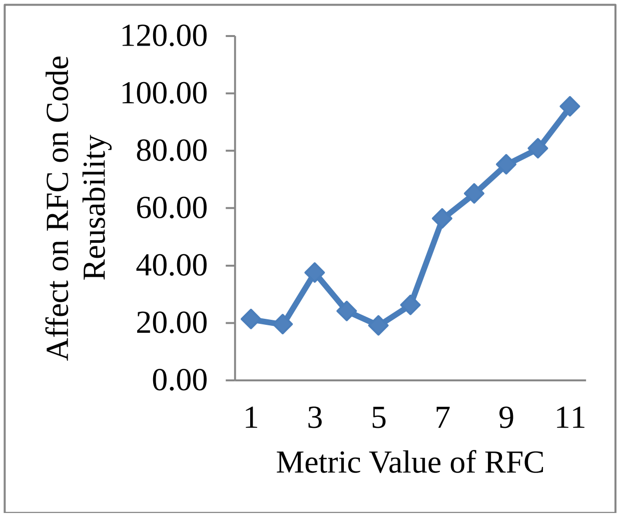


Fig. 3. Affect of RFC on Code Reusability (Synthetic)

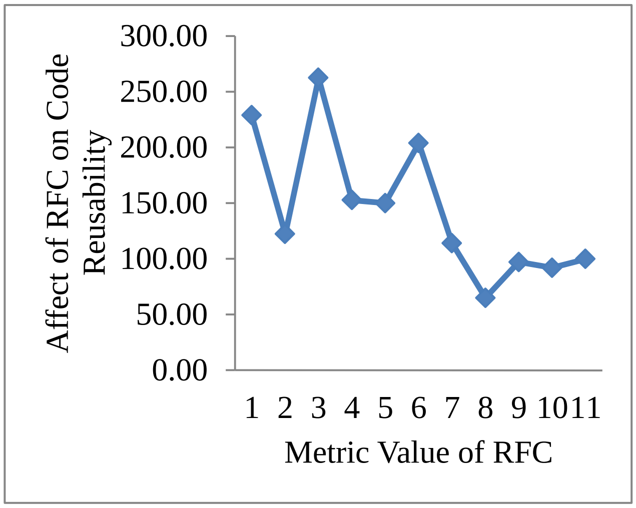


Fig. 4. Affect of RFC on Code Reusability (Original)

Fig.3 shows the outcome of the synthetic data where the parameters are kept in random state, which is less likely to happen in complex software projects. Although the outcome shows an increase in code reusability, there could be possibly latent perspective of the outcome. Hence, we use original data from our ERP and SCM project where RFC parameters (α and β) were structured in increasing order. The outcome shows that there is a significant drop in code reusability. Therefore, the outcome stated in Fig.4 is within the agreement of the hypothesis that “The complexity of code design and reusability decreases for maximized outcomes of RFC.”

C. Effect of NOC on Code Reusability:-

Fig.5 highlights the effect of NOC over code reusability. The outcome shows that code reusability has significantly improved with the numerical values mentioned in Table 1. A closer look at the numerical values will show that parameters (α and β) are in the same trend of maximization order. Therefore, overall it can be said that increase in NOC has resulted in improved code reusability. The prime reason behind this is larger values of NOC will represent maximized amount of base classes that results in significant code reusability. Therefore, the null hypothesis stating “Code reusability improves for increasing values of NOC present in class” is found accepted.

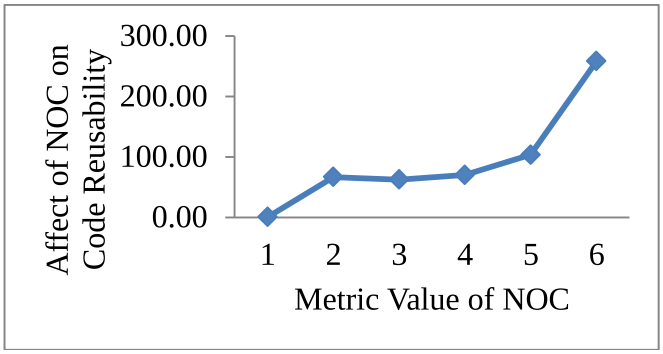


Fig. 5. Affect of NOC on Code Reusability

D. Effect of CBO on Code Reusability:-

It is said that higher values of CBO are not good for software engineering as excessive coupling can turn the design evaluation quite difficult and complex. But, software projects using object-oriented may more likely have CBO values than expected. Hence, it is found that proposed system can drastically enhance the code reusability even if the software projects do have higher CBO values. Fig.6 shows that with increasing CBO values for any forms of object-oriented codes, it is feasible to get more values of code reusability. Hence, the hypothesis stating that “Increased CBO values doesn’t have much impact over code reusability” is accepted.

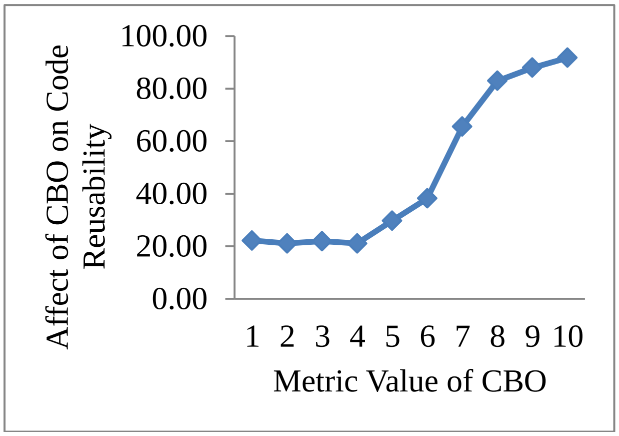


Fig. 6. Affect of CBO on Code Reusability

E. Effect of WMC on Code Reusability:-

Theoretically, it is expected that increased values of WMC may result in design complexity. However, when we performed our evaluation using proposed technique, we found no much adverse effect on code reusability for certain initial rounds. A closer look at the outcomes shown in Fig.7 will highlight that although code reusability increases due to exposure to various methods used to the classes, code reusability increases, at the same time, the performance is found to be degrading for further increment in (α and β) parameters of WMC. Hence, the null hypothesis stating “Code reusability declines for increasing values of WMC in every class” is found accepted in proposed system.

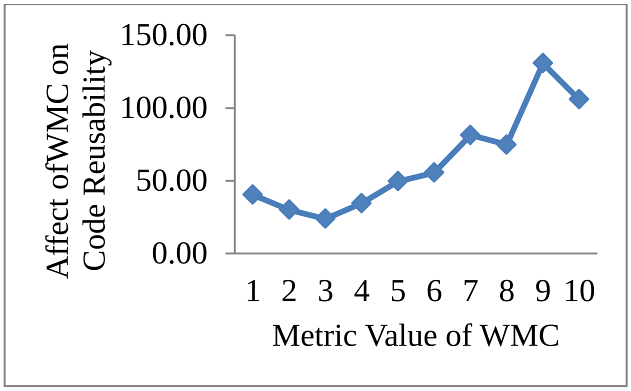


Fig. 7. Affect of WMC on Code Reusability

A closer look at the cumulative outcome of the study will show that code reusability increases for some CK metric parameters and degrades for some other CK metric parameter. Our observation encounters more processing and analysis time to carry out testing. Our work was in the direction of enhanced cohesion between the significant methods as well as to ensure that there is a minimal coupling between the potential objects. This condition has ensured to retain better code reusability by ensuring higher cohesion among the classes. If the intensity of the coupling is more than we found such cases to be quite non-supportive of code reusability. Moreover, it is also explored that inclusion of number of methods in ERP projects causes the design class to incur more computational complexity and thereby leads to inferior design patterns. Internal callbacks and communication for message more than 150 cause degradation in the code reusability.

From the result discussed in this section, it can be seen that there are various factors of CK metric that directly impacts the code reusability process. During working on the new set of the code, it is essential that new components and classes designed should have to be within the anticipated outcomes to claim for reusable codes. Hence, our proposed technique can present a framework that can be used to measure the relationship between the CK metric and code reusability. The design to be incorporated into the new set of the code must have a permissible limit of code reusability, which the designer can easily set up during the formal verification process. The proposed system is highly extensible for various forms of software projects other than ERP and SCM.

TABLE II. NUMERICAL OUTCOMES OF STUDY

ERP Project (C: 220, MC: 120)					**SCM (C:570, MC: 380)**			
Metric	**Value**	**α**	**β**	**C_r**	**α**	**β**	**C_r**	**Avg**
DIT	1	26	3	11.54	32	3	9.38	10.46
	2	36	9	25.00	59	15	25.42	25.21
	3	63	35	55.56	84	46	54.76	55.16
	4	33	24	72.73	121	90	74.38	73.55
	5	36	31	86.11	132	113	85.61	85.86
	6	18	18	100.00	112	110	98.21	99.11
RFC	5	11	1	9.09	3	1	33.33	21.21
	10	12	2	16.67	9	2	22.22	19.44
	20	8	2	25.00	4	2	50.00	37.50
	34	27	4	14.81	6	2	33.33	24.07
	48	7	1	14.29	21	5	23.81	19.05
	65	5	1	20.00	34	11	32.35	26.18
	85	13	7	53.85	63	37	58.73	56.29
	100	41	26	63.41	81	54	66.67	65.04
	140	25	18	72.00	97	76	78.35	75.18
	165	31	24	77.42	119	100	84.03	80.73
	200	33	31	93.94	92	89	96.74	95.34
NOC	0	209	1	0.48	530	1	0.19	0.33
	1	3	1	33.33	2	2	100.00	66.67
	2	4	1	25.00	3	3	100.00	62.50
	3	5	2	40.00	4	4	100.00	70.00
	4	5	2	40.00	3	5	166.67	103.33
	6	6	1	16.67	1	5	500.00	258.33
CBO	1	3	1	33.33	9	1	11.11	22.22
	3	9	2	22.22	10	2	20.00	21.11
	4	7	2	28.57	13	2	15.38	21.98
	5	18	4	22.22	15	3	20.00	21.11
	8	42	11	26.19	21	7	33.33	29.76
	10	23	8	34.78	60	25	41.67	38.22
	14	27	18	66.67	76	49	64.47	65.57
	19	24	21	87.50	88	69	78.41	82.95
	21	34	31	91.18	131	111	84.73	87.95
	24	31	29	93.55	119	107	89.92	91.73
WMC	3	13	4	30.77	2	1	50.00	40.38
	5	20	4	20.00	5	2	40.00	30.00
	8	22	5	22.73	16	4	25.00	23.86
	10	15	5	33.33	28	10	35.71	34.52
	14	22	11	50.00	57	28	49.12	49.56

	18	23	13	56.52	88	48	54.55	55.53
	30	15	10	66.67	51	49	96.08	81.37
	50	37	28	75.68	69	51	73.91	74.79
	80	24	30	125.00	44	60	136.36	130.68
	100	31	38	122.58	184	164	89.13	105.86

F. Outcome Assessment

To validate the proposed system, we have developed a simple assessment model. The prime theme of this validation model is to understand the effect of increasing uncertainty in the three types of software projects towards code reusability (Fig.8). The assessment model is designed over Matlab, which takes the empirical values of metric software suite from Customer Relationship Management project, Supply Chain Management project, and Enterprise Relationship Management project. Using the introduced equation of code reusability, the system performs the computation. The outcome of the code reusability will be subjected to various levels of uncertainties. We define uncertainties as various hidden parameters like skill gap, requirement volatility, ignorance of complying with software development life, project slippage, etc., which is beyond the control of any human. We like to understand that in the case of hidden or unforeseen circumstances in any project management team, what is the performance of code reusability in that case? Hence, we apply the mathematical approach in this regard by using Markov model [36]. Applying Markov modeling, it becomes possible to map all the real-time uncertainties into an empirical parameter and apply it to the proposed code reusability model to understand its behavior.

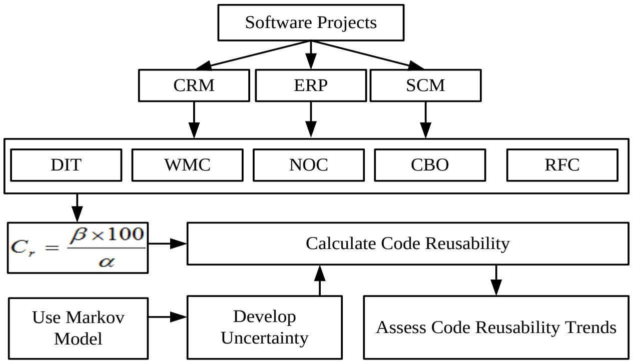

Software Projects
CRM
ERP
SCM
DIT
WMC
NOC
CBO
RFC
$C_r = \frac{\beta \times 100}{\alpha}$
Calculate Code Reusability
Use Markov Model
Develop Uncertainty
Assess Code Reusability Trends

Fig. 8. Model for Assessing Proposed System

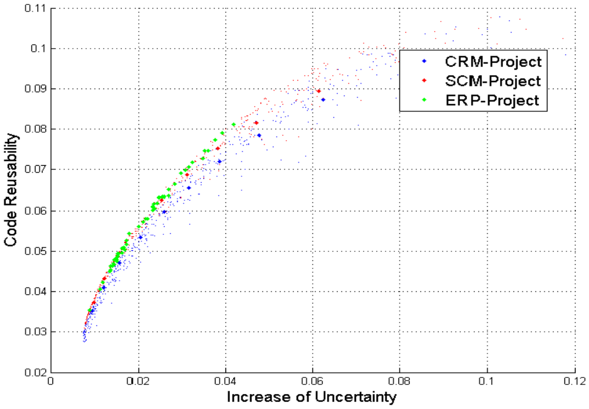


Fig. 9. Comparative Analysis of Software Projects

Fig.9 shows the comparative analysis of the three different software projects based on object-oriented designs using a scattered diagram. The outcome shows dominancy of ERP projects, where the code reusability is found to be quite significantly increase with the increase of uncertainty. However, the adverse side of this outcome is that such dominant result is only visible till 0.01-0.04 levels of uncertainty values. Better than ERP project, SCM project was found with sparse but increasing values of code reusability. A closer look at the scattered plot will show that there is a significant increase in code reusability from 0.01-0.08 values of uncertainties. However, owing to the inclusion of a maximum number of classes and methods, the values of DIT and WMC increases resulting in design complexities. This is the main reason code reusability for SCM can be evaluated in a sparse manner and slower pace, but with better accuracy compared to ERP projects. We have also testified the assessment model with a middle-sized CRM project. Normally, the amount of design complexities CRM projects is highly increased in multifold. However, using the proposed equation, we testified our hypothesis and found that code reusability to be significantly enhanced for CRM project in spite of massive design complexities involved in project architecture.

X. Conclusion

At present, we have drawn a relationship between the most standard software metrics and code reusability. We have testified it on three complex software projects of object-oriented designs and found that our model can significantly calculate code reusability for any extent of complexities even it is very much uncertain. Using mathematical and stochastic approach of Markov Modelling, we proved that our model can extract more data of code reusability on increasing uncertainties. Design pattern plays an important role in software engineering. With the increasing demands of the customers, the IT industries and software project developers are increasingly seeking consultation to minimize the cost of production from more than a decade. In the form of various cost-cutting procedures, code reusability is the most prominent one and requires a highly skilled technical architecture to take a decision. A code reusability deals with two challenging aspects i.e. i) deciding which part of the code to be retained same and ii) deciding which part of the code will need to be designed from scratch. In the first challenging aspect, a developer can easily decide on what part of the code will be required to be retained based on the client's requirement. However, the difficult part is to make a decision related to the new code that is required to be built from the base. Normally, depending on an experienced architecture, the new set of the code that needs to be programmed is designed in such a way that it should posses a certain level of code reusability for the future client, which is unpredictable. An unpractical design, in this case, will go to complete loss of production and may not meet the reusability factor for new projects. Hence, our future direction of study will focus on estimating the level of code reusability for complex software projects. We anticipate that our design concept will highly encourage and motivate the stakeholder to consider it as most cost-effective tool to date.

References

[1] J. Mishra, A. Mohanty , "Software Engineering", Pearson Education India, Electronic books, pp. 387, 2011

[2] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, "Experimentation in Software Engineering: An Introduction", Springer Science & Business Media, pp. 204, 2012

[3] S. K. Dubey, A. Rana, A Comprehensive Assessment of Object-Oriented Software Systems Using Metrics Approach, International Journal on Computer Science and Engineering, vol. 02, no. 8, pp.2726-2730, 2010

[4] N. Mohammed, A. Govardhan, Comparison between Traditional Approach and Object-Oriented Approach in Software Engineering Development, International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, 2011

[5] B.Jalender, A.Govardhan, P.Premchand, Designing code level reusable software components, International Journal of Software Engineering & Applications (IJSEA), vol.3, no.1, 2012

[6] "Classification of Software Metrics in Software Engineering", http://ecomputernotes.com/software-engineering/classification-of-software-metrics, Retrived, 10th Dec, 2012

[7] M. Sarker, "An Overview of Object Oriented Design Metrics", Master Thesis Department of Computer Science, Umea University , Sweden, 2005

[8] "Metamil", http://www.metamill.com/, Retrived, 10th Dec, 2015

[9] "Sourceforge", metrics.sourceforge.net, Retrived, 10th Dec, 2015

[10] "Codeswat Custom Solutions", http://codeswat.com/, Retrived, 10th Dec, 2015

[11] J. Alghamdi, R. Rufai, and S. Khan. Oometer: A software quality assurance tool. Software Maintenance and Reengineering, 2005. CSMR 2005. 9th European Conference, pp. 190–191, 2005

[12] V. R. Basili, L. Briand and W.L. Melo, "A Validation Of Object-Oriented Design Metrics As Quality Indicators", Technical Report, Univ. of Maryland, Dep. of Computer Science, College Park, MD, 20742 USA. April 1995.

[13] C.N.S.Anna, A.F.Garcia, C.V.F.G. Chavez, C.J.P.d. Lucena, A.V. Staa, "On the Reuse and Maintenance of Aspect-Oriented Software:An Assessment Framework", PUC-RioInf.MCC26/03 Agosto, 2003.

[14] P.S,Kaur, and A. Singh.,"Modeling of Reusability of Object Oriented Software System", World Academy of Science, Engineering and Technology, vol. 56, pp.162. 2009.

[15] M. Kaur, M. Mahajan, P.S. Sandhu, "A k-NN based approach for Reusability Evaluation of Object-Oriented Based Software Components, International Conference on Information and Communications Security, 2011

[16] U. Kumari, S. Bhasin. Application of object-oriented metrics to C++ and Java: A comparative study. ACM SIGSOFT Software Engineering Notes, vol. 36(2), pp.1-10, 2011

[17] P. Edith Linda, E. Chandra and J. Sharmila, "An Approach to Evaluate Object Oriented Class Structure using Score Carding Framework", International Journal of Software Engineering and Its Applications, vol. 9, No. 3, pp. 9-16, 2015.

[18] D. Wu, L.Chen, Y. Zhou and B. Xu, "A metrics-based comparative study on object-oriented programming languages", State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing, China, DOI reference number: 10.18293/SEKE2015-064, 2015.

[19] K.P. Srinivasan And T. Devi, "A Comprehensive Review And Analysis On Object-Oriented Software Metrics In Software Measurement", International Journal on Computer Science and Engineering (IJCSE), vol. 6, no.07, 2014..

[20] M. Scotto, A. Sillitti, G. Succi, T. Vernazza, "A relational approach to software metrics", ACM Symposium on Applied Computing, pp.1536-1540, 2004.

[21] S. Singh, P. Singh, N. Mohan, P.S. Sandhu, "Logistic Model Trees based Approach for Prediction of Reusability of Object Oriented Software Components", International Journal of Research in Engineering and Technology, vol. 1, No. 3, 2012

[22] R. Subramanyam, M.S. Krishnan, "Empirical Analysis of CK Metrics for Object-Oriented Design Complexity: Implications for Software Defects", IEEE Transactions on Software Engineering, vol. 29, no. 4. 2003.

[23] A. Shaik, C.R.K. Reddy, B. Manda, C. Prakashini and K. Deepthi, "Metrics for Object Oriented Design Software Systems: A Survey",Journal of Emerging Trends in Engineering and Applied Sciences (JETEAS), vol. 1(2), pp.190-198, 2010.

[24] S. I. Zahara, M. Ilyas and T. Zia, "A Study of Comparative Analysis of Regression Algorithms for Reusability Evaluation of Object Oriented Based Software Components", International Conference on Open Source Systems and Technologies (ICOSST), 2013.

[25] H.M. Manoj and A.N. Nandakumar, "A Survey on Modelling of Software Metrics for Ranking Code Reusability in Object Oriented Design Stage", International Journal of Engineering Research & Technology (IJERT), vol. 3, Issue. 12, 2014.

[26] A. Oberoi and D. Arora,"Quality Model For Analysis And Implentation Of CK Metrics Through Neural Networks: International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622.National Conference on Advances in Engineering and Technology, AET, 2014.

[27] N.Goyal and D. Gupta, "Reusability Calculation of Object Oriented Software Model by Analyzing CK Metric",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 3 Issue. 7, 2014.

[28] N.Jayalakshmi and Nimmati Satheesh," Software Quality Assessment in Object Based Architecture",International Journal of Computer Science and Mobile Computing, vol.3, issue.3, pg. 941-946, 2014.

[29] A.V. Hudli and R.V. Hoskins: "Software metrics for OOD", IEEE International conference, 2002.

[30] H.Lilu, K.Zhou and S.Yang: "Quality metrics of OOD for Software development and Re-development", First Asia-Pacific Conference on Quality Software, 2002.

[31] N. Paliwal, V.Shrivastava and K. Tiwari, "An Approach to Find Reusability of Software Using Objet Oriented Metrics", International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, issue 3, 2014.

[32] N. Chauhan and M. V.Gupta, "Evaluation Of Metrics And Assessment Of Quality Of Object Oriented Software", IJRET: International Journal of Research in Engineering and Technology, vol. 03, special issue: 14, 2014.

[33] D. Gupta, V. K. Goyal and H. Mittal, Comparative Study of Soft Computing Techniques for Software Quality Model, International Journal of Software Engineering Research & Practices, vol.1, issue: 1, 2011.

[34] R. Alcalá, J. Casillas, O.Cordón, and F. Herrera, "Linguistic modeling with weighted double-consequent fuzzy rules based on cooperative co-evolutionary learning", Integrated Computer-Aided Engineering, vol. 10, no. 4, pp. 343-355, 2003

[35] "SPSS software", http://www-01.ibm.com/software/analytics/spss/, Retrieved 10th Dec, 2015

[36] M. Stamp "A revealing introduction to hidden Markov models", Department of Computer Science San Jose State University, 2004.

An Ensemble of Fine-Tuned Heterogeneous Bayesian Classifiers

Amel Alhussan
Computer Science Department
Princess Nourah Bint Abdulrahman University
King Saud University
Riyadh, Saudi Arabia

Khalil El Hindi
Computer Science Department
King Saud University
Riyadh, Saudi Arabia

***Abstract*—Bayesian network (BN) classifiers use different structures and different training parameters which leads to diversity in classification decisions. This work empirically shows that building an ensemble of several fine-tuned BN classifiers increases the overall classification accuracy. The accuracy of the constituent classifiers can be achieved by fine-tuning each classifier and the diversity is achieved using different BN classifiers. The proposed ensemble combines a Naive Bayes (NB) classifier, five different models of Tree Augmented Naive Bayes (TAN), and four different model of Bayesian Augmented Naive Bayes (BAN). This work also proposes a new Distance-based Diversity Measure (DDM) and uses it to analyze the diversity of the ensembles. The ensemble of fine-tuned classifier achieves better average classification accuracy than any of its constituent classifiers or the ensemble of un-tuned classifiers. Moreover, the empirical experiments present better significant results for many data sets.**

***Keywords*—Ensemble classifier; Bayesian Network (BN) classifiers; Fine-tuned BN classifiers; Stacking; Diversity**

I. Introduction

Bayesian network classifiers are probabilistic models that encode the conditional independence relationships between the attributes in different ways. There are many learning algorithms to build TAN and BAN structure, such as TAN search, K2 search, Tabu Search, Hill Climber Search and Repeated Hill Climber Search. These search algorithms yield different TAN and BAN classifiers.

Building ensembles of classifiers is a powerful method for obtaining better classification accuracy through combining the classification of multiple classifiers [1]. Boosting [2] [3] and bagging [2] [3] are the two most commonly used methods for building ensembles of homogenous classifiers. On the other hand, stack generalization (stacking) [4] and ensemble selection [5] are suitable for building ensembles of heterogeneous classifiers.

Diversity and the accuracy of the base classifiers are important factors to achieve a powerful ensemble of classifiers. It would be meaningless to combine several classifiers that make the same predictions. The intuition is that if many classifiers make errors on different instances, the combination of these classifiers can reduce the overall error and improve the performance of the ensemble system [6]. The main advantage of ensemble different BN classifiers is that it is unlikely that all classifiers will make the same mistake. It would also be meaningless to combine classifiers that are too weak. Therefore, in order to build ensemble of classifiers with better accuracy, we need to combine relatively accurate and diverse classifiers.

The diversity of classifiers is achieved by using single learning algorithm with different in data sets (using sampling), training parameters, or subset of features [1] [7] [8]. These methods are considered homogenous methods because they use the same learning algorithm. On the other hand, an ensemble might consist of a group of classifiers, each built using the same training data but a different learning algorithm [7] [9]. Ensembles of heterogeneous classifiers might be more suitable if the learning algorithms are stable in the sense that a small change in the training data does not lead to a substantially different classifier. Heterogeneity might be more suitable at achieving diversity in this case. Naive Bayes (NB) and Tree Augmented Naive Bayesian (TAN) are known to be stable algorithms [10] [11].

We empirically show that ensemble several fine-tuned BN classifiers, namely: fine-tuned Naive Bayesian classifiers (FTNB) [12], fine-tuned TAN (FTTAN) [13] and BAN (FTBAN) classifiers, achieves better classification accuracy for many data sets, than an ensemble of un-tuned classifiers or any of its constituent classifiers. We also propose a Distance-base Diversity Measure (DDM) and use it to analyze our results. Since the error rate of different BN classifiers is below 50%, we expect that the ensemble classifier will yield better classification accuracy over the constituent classifiers. In this research, we achieve diversity by using different types of BN classifiers by using NB classifier, TAN classifier, and BAN classifier to construct an ensemble of classifiers. Moreover, we use different models of TAN and BAN by using different search algorithms to build its structure. Also, by using fine-tuned classifiers [12] [13], we are constructing an ensemble of relatively accurate classifiers. This work improved the classification accuracy of BN classifiers by building three different ensemble classifiers: 1) the original un-tuned 10 BN classifiers (NB, five models of TAN and four models of BAN), 2) the corresponding 10 fine-tuned BN classifiers, 3) a combination of all previous twenty BN classifiers. We also compared the results these three different ensembles of classifiers.

This work is structured as follows: in section II we review the related work on building ensembles of BN classifiers. In section III, we present our BN ensembles of classifiers. Section IV presents the experimental results and a comparison between the different ensembles. Section V is the conclusion.

II. Related Work

Diversity among individual classifiers is important in order for an ensemble to achieve better accuracy than the accuracy of any of constituent classifiers. Usually diversity of the base classifiers is achieved by training single learning algorithm on different data sets (bootstrap resampling) [11] [14] [15] [16], different parameters [14], or different features [1] [7] [8]. However, some works achieve diversity by training different learning algorithms on the same data. Ma and Shi [14] propose TAN learning algorithm called Random Tree-Augmented Naive Bayes (RTAN) that generates different TAN classifiers to be combined in an ensemble classifier. The algorithm builds TAN model by selecting the arcs whose conditional mutual information is larger than a certain threshold value. RTAN algorithm builds different TAN models by using different threshold values and different start edges. RTAN algorithm is trained on different training subsets, and then the different TAN classifiers to construct TAN ensemble classifier using a majority of votes. Their experimental results show that bagging Multi-TAN ensemble classifier has higher classification accuracy than the standard TAN classifier. Also, Shi et. al [11] used RTAN algorithm for boosting MultiTAN that shows higher classification accuracy than standard TAN classifier. Sun and Zhou [15] [16] used a boosting technique that is characterized by the way in which the hypothesis weights are selected, and by the instance weight update step. They used boosting to combine multiple TAN classifiers and compared it with Boosting-BAN classifiers. Their experimental results show that the Boosting-BAN has higher classification accuracy than Boosting-MultiTAN on noise-free data. Moreover, Sun and Zhou [17] built an ensemble combing Boosting-BAN and Boosting-MultiTAN using the sum voting methodology. The sum rule adds all confidence scores of sub-ensemble Prediction for each class and the class with the highest sum wins the election. They report that their proposed ensemble classifier is significantly more accurate than TAN, BAN, Boosting-BAN and Boosting-MultiTAN methods. Tsymbal et al. [8] developed an ensemble of NB classifiers that randomly samples the feature space. They found that the performance of their ensemble of classifiers performed better than a single naive Bayesian classifier. Lee and Cho [7] combined three different classifiers to build an ensemble. They created a General Bayesian network (GBN) to identify the variables inside the Markov blanket of GBN's class node, and then used those selected variables to create a GBN-assisted ensemble by combining GBN, decision tree, and/or SVM using voting and stacking combination strategies. They found that the ensemble systems generally improved the prediction accuracy. Sakkis, et al. [18] use stacked generalization approach to anti-spam filtering. They combined a memory-based classifier and a Naïve Bayes classifier in an ensemble classifier. Their experiments improved the performance of anti-spam filter and outperformed the two base classifiers. They report that the improvement of the ensemble of classifiers is due to the high diversity of the two base classifiers. Jing et al. [19] construct an ensemble Bayesian belief network (BBN) and exploit TAN learning algorithm to build a BBN structure. They combined parameter boosting and structure learning to improve the classification accuracy of BBN classifiers. Their algorithm goes through a fixed number of iterations and stops if the training error increases. At the beginning of each iteration a training set and its corresponding weights for the data points are given to the TAN algorithm to build a BBN Classifier. The TAN algorithm is used to build base classifier, it starts with an empty set and adds i edges with the highest mutual information to a naïve BBN. The training error of the resulting TAN classifier is then used to determine the weight of the test data points in subsequent iterations. According to their results, their boosted BBNs have comparable or reduced average testing error than NB and TAN. This work has an advantage over the previous works by using fine-tuned BN classifiers. Fine tuning process address the unreliable estimation of the attributes conditional probabilities due to the lack of data and improve BN classifiers accuracy by finding more accurate estimation of the probabilities terms.

III. Ensemble Bayesian Network Classifiers

In this work, we build an ensemble of BN classifiers. Each classifier in the ensemble is trained using the same data. Stacking [4] is employed to build three different ensembles. Stacking is employed to combine classifiers built by different learning algorithms. The main idea behind Stacking is to use the classifications of a set of base classifiers (level-0) estimated by using cross-validation, to learn a meta classifier (level-1) which gives the final prediction [20].

In this research, stacking splits the data set into two disjoint parts (using 10-fold Cross-Validation), then train all BN base learners on the first part. Then test the base learners on the second part. The predictions of all BN base classifiers are combined by using simple plurality voting to produce an ensemble of BN classifiers. Diversity is achieved by using three different types of BN (NB, TAN and BAN) classifiers. Moreover, we exploit the structure learning algorithms to build five different TAN classifiers and four different BAN classifiers. The search algorithms that were used to build different TAN classifiers are: TAN search, K2 search, Tabu Search, Hill climber Search and Repeated Hill Climber Search. The last four search techniques were used also to build four different BAN classifiers. We also, used their corresponding fine-tuned classifiers: fine-tune NB (FTNB) [12], fine-tune TAN (FTTAN) [13] and fine-tune BAN (FTBAN). Three ensemble classifiers were built; the first one combines 10 BN classifiers (NB, five models of TAN and four models of BAN). The second ensemble classifier combines the 10 corresponding fine-tuned classifiers, and the last one combines all 20 BN classifiers (fine-tuned and un-tuned).

A. *Distance-Based Diversity Measure (DDM)*

Since the diversity of the base classifiers has direct effect on the ensemble's classification accuracy, there is a need to be able to measure it. Kuncheva and Whitaker [21] compared several measures of diversity and concluded that all measures had approximately equally strong relationships and they were strongly correlated. Some of their experiments revealed the inadequacy of these measures to predict the accuracy of the ensemble. The low correlation between these measures on the one hand and the improvement in classification accuracy on the other hand, is discouraging. This work proposes a new distance-based diversity measure and uses it to analyze the relationship between the base classifiers diversity and the ensemble accuracy.

We have M classifiers and C classes, if we ignore accuracy and the ideal diverse ensemble would give equal votes for each class. For example, if we have 10 classifiers and 5 classes, the ideal (most diverse) vote vector for the five classes would be (2, 2, 2, 2, 2). In other words, the vote vector in which each class would get M/C votes. The least diverse ensemble is the one that has all its constituent classifiers voting for the same class, while all remaining classes have zero votes. In our example, the vote vector for the five classes would be something like (0, 0, 0, 0, 10). A good diversity measure would be based on the distance between the ideal vote vector and the actual vote vector for all instances. The small distance indicates more diverse classifiers and large distance indicates less diverse classifiers. We can compute the distance for an instance i giving its voting vector $\mathrm{V_i}$ as follows:

$$\mathrm{dist_i(V_i)} = \sum_{c=1}^{C} \left| \frac{M}{C} - V_i(c) \right| \quad (1)$$

This distance should be computed for N instances in the training set.

$$\mathrm{Dist} = \sum_{i=1}^{N} \mathrm{dist_i(V_i)} \quad (2)$$

The max distance, $\mathrm{MaxDist}$ that could be achieved a voting vector of a given instance is

$$\mathrm{MaxDist} = (C-1) * \frac{M}{C} + (\, M - \frac{M}{C}\,) \quad (3)$$

This is because $\mathrm{C}-1$ of the classes would get 0 votes and one class would get all votes. The distance for the $\mathrm{C}-1$ classes is $(C-1) * (\frac{M}{C} - 0)$ and the distance for the class that gets all votes is $(\, M - \frac{M}{C})$

Therefore, the maximum distances for all instances $\mathrm{MaxDist} * \mathrm{N}$.

The Distance-Based Diversity Measure (DDM) is defined as follows:

$$\mathrm{DDM} = 1 - \frac{\sum_{i=1}^{N} \mathrm{Dist_i(V_i)}}{\mathrm{MaxDist} * N} \quad (4)$$

Thus, diversity ranges from zero to one, where zero indicates the lowest diversity and one indicates the highest diversity.

IV. Experimental and Results

In all experiments, we used 40 data sets, obtained from the UCI repository [22]. The BAN models used in our experiments had a maximum of three parents for each attribute node. All ordinal attributes were discretized using Fayyad et al.'s [23] supervised discretization method, as implemented in Weka. The missing values in the data sets were simply replaced by the most common values. Ten-fold cross validation was used in all experiments. All experiments were implemented in the Weka workframe and used as much of the Weka classes as possible. We built three ensembles of classifiers that are based on different types of BN classifiers (NB, TAN and BAN). The ensemble of classifiers uses a simple majority (plurality) voting technique to classify instances.

We used classification formulas Eq. (5) for NB and Eq. (6) for TAN and BAN classifiers, as proposed by Friedman et al. [24].

$$\mathrm{c_{predicted}} = \arg\max_{c \in C} P(c) \prod_t P(a_t|c) \quad (5)$$

$$P(c|a_1 \ldots\ldots a_n) = P(c) \prod_{i=1}^{n} P(a_i|\mathrm{parent}(a_i) \wedge c) \quad (6)$$

We used Laplace estimator to estimate all probabilities values.

$$P(x = i) = (n_i + \mathrm{alpha})/(N + K * \mathrm{alpha}) \quad (7)$$

Where K is the number of different values of x; and Alpha is a small positive value. In our experiments, we used Weka simple estimator with Alpha = 0.5 to estimate the conditional probability of NB (which is the default value used by Weka for NB) and we choose Alpha = 0.2 for TAN and BAN (which gave us best results). We experimented with different values for Alpha and 0.2 gave us the best results.

The tables from 1 to 5 show the results of our three ensembles of BN classifiers. Also, it compares each ensemble BN classifier with its individual base classifiers. The last four rows of each table show the average values (classification accuracy over the 10 folds and ensemble diversity), the number of data sets with better results, and the number of data set with significantly better results at the 95% and 90% confidence levels. A paired t-test with confidence levels of 95% and 90% was used to determine whether the differences were statistically significant. The better results are highlighted in bold in the tables. The significant results, at 95% confidence level, are highlighted in bold and underlined, while the significant results at 90% confidence level are double underlined.

A. Stacking BN classifiers

The first experiment combined ten BN classifiers, an NB, five TAN classifiers and four BAN classifiers. Different structure learning algorithms were used to build different TAN and BAN classifiers. The five different TAN classifiers are distinguished by using the name of the search algorithm used to build them as a postfix. Thus, we have TAN-TAN search, TAN-K2, TAN-tabuSearch, TAN-HillClimber and TAN-RepeatedHillClimber. In the same way, we denoted the four different BAN classifiers (BAN-K2, BAN-tabuSearch, BAN-HillClimber and BAN-RepeatedHillClimber). The Ensemble classifier of the ten BN classifiers is called (EBN-10).

Table 1 shows the results of EBN-10 and the results of each of the constituent classifiers. It is obvious from the table that the average classification accuracy of EBN-10 is better than the average accuracy of any of the constituent classifiers. The average accuracy of the ensemble classifier is 71.69%, while the average accuracy of the constituent classifiers ranges from 64.92% to 70.88%. Moreover, the ensemble classifier outperforms all constituent classifiers in terms of the number data sets for which it achieves better and significantly better results, at the 95% confidence level. EBN-10 outperformed NB on 20 data sets; ten of them are significantly better results. Regarding the TAN models, EBN-10 outperformed TAN-TAN search, TAN-K2 and TAN-tabuSearch on 19 data sets, five, four and eight of them, respectively, are significant better results. Also, EBN-10 classifier outperforms TAN-HillClimber on 15 data sets; five of them are significant better results and outperforms TAN-RepeatedHillClimber on 21 data sets; seven of them are significant better results. The table also shows more obvious superior results of EBN-10 with BAN models. EBN-10 outperformed BAN-K2 and BAN-tabuSearch on 26 data sets, 19 and 16 of them, respectively, are significant better results. Moreover, EBN-10 classifier outperforms TAN-HillClimber on 24 data sets; 18 of them are significant better results. Also, EBN-10 classifier outperforms TAN-RepeatedHillClimber on 27 data sets, 19 of them are significant better results.

B. *Stacking Fine-Tuned BN Classifiers*

In second experiment, we constructed an ensemble of the same classifiers but after fine-tuning them. We used the fine-tune NB (FTNB) [12], and its adapted (FTTAN) [13] version to fine-tune TAN learning algorithm and fine-tune BAN (FTBAN). Fine-tuning each classifier improves the classification accuracy by finding more accurate estimation of probabilities terms. The enhanced accuracy of BN classifiers encouraged us to build an ensemble of these fine-tuned classifiers. The ensemble of the 10 fine-tuned BN classifiers is called (EFTBN-10).

Table 2 shows the results of EFTBN-10 and the results of the each of the constituent classifiers. The average classification accuracy of EFTBN-10 is better than all individual FTBN classifiers. The average accuracy of ensemble fine-tuned classifier is 72.48%, while the 10 FTBN classifiers average accuracy range between 66.08% and 72.01%. On other hand, the fine-tuned ensemble classifier outperform all individual FTBN classifiers in the number of better and significantly better number of data sets at the 95% confidence level. EFTBN-10 outperformed FTNB on 26 data sets; eight of them are significantly better results. Also, EFTBN-10 outperformed FTTAN-TAN search for 23 better data sets, six of them are significantly better results. Moreover, EFTBN-10 outperformed FTTAN-K2 and FTTAN-HillClimber search on 18 data sets, four and six of them are significantly better results, respectively. Also, EFTBN-10 outperformed FTTAN-TabuSearch and FTTAN-Repeated HillClimber on 22 better data sets, 10 and six of them are significantly better results, respectively.

The improvements of EFTBN-10 are even more obvious compared with the fine-tuned BAN classifiers (FTBAN). EFTBN-10 is better than FTBAN-K2 for 27 data sets, 17 of them are significantly better results. EFTBN-10 also achieved results for 27 data sets than FTBAN-TabuSearch, 12 of them are significantly better results. Moreover, EFTBN-10 outperformed FTBAN-HillClimber and FTBAN-RepeatedHillClimber on 27 and 30 better data sets, 19 and 16 of them are significantly better results, respectively. The superiority of EFTBN-10 is even more obvious at 90% confidence level (see the last row of Table 2).

C. *Stacking BN classifiers and their corresponding fine-tuned classifiers*

In the third experiment, we built an ensemble by combining the previous twenty BN classifiers (10 BN classifiers and their corresponding fine-tuned BN classifiers).We call this ensemble EBN-20.

Table 3 and Table 4 show the results of EBN-20 compared to the result of each of the constituent classifiers. The result of this ensemble is a compromise of the previous two classifiers. The tables show that the average classification accuracy of EBN-20 is 71.56% which is better than the average accuracy of any of its 20 constituent classifiers, except for FTTAN-TAN and FTTAN-K2. The result not a surprising because TAN search and K2 search algorithms have exhibited excellent performance in data mining [25] [26] and the fine tuning process makes them even better. The degradation of EBN-20 average accuracy is probably because EBN-20 combines fine-tuned and non-fine-tuned classifiers, which reduces diversity, as the constituent classifiers are not very different classifiers. In the terms of number of better and significantly better data sets at 90% confidence level, EBN-20 outperformed all the 20 individual classifiers. Also, EBN-20 outperformed all of the 20 classifiers with respect to the number of data sets it achieves better and significantly better data sets at 95% confidence level except for FTTAN-K2 classifier where it wins on four data sets and loses on five data sets (see Tables 3 and 4 for more details).

D. *Comparing the Three Ensembles*

Table 5 shows the results of comparing the three ensembles: ENB-10, EFTBN-10, and EBN-20. The table also shows the diversity value for each ensemble. As can be seen in table, EFTBN-10 outperforms EBN-10 with respect to the average classification accuracy, and the number of data sets for which it achieves better and significantly better results. EFTBN-10 achieves on average 72.47% classification accuracy, while EBN-10 achieves 71.69%. EFTBN-10 also achieves significantly better results for 6 data sets and worse results for only 1 data set. EFTBN-10 outperforms EBN-10 because its constituent classifiers, namely the fine-tuned classifiers, are more accurate than the constituent classifiers of EBN-10. In fact, the proposed diversity measure shows that both ensembles have the same average diversity of 0.44.

Comparing EFTBN-10 with EBN-20 shows that EFTBN-10 also outperforms EBN-20, which has an average classification accuracy of 71.56%. EFTBN-10 also achieves better results than EBN-20 for 13 datasets 3 of them are significantly better and 2 are significantly worse at 95% confidence level. At 90% confidence level, EFTBN-10 achieves better results for 5 data sets and worse results for 2 data sets. This result is a little bit surprising because EBN-20 contains much more classifiers. It contains the same classifiers of EFTBN-10 in addition to their un-tuned counterparts. This indicates EBN-20 must have less diversity than EFTBN-10, which is expected because the fine-tuned classifiers and their un-tuned counterparts are not very different classifiers. The proposed diversity measure actually supports this analysis. The diversity measure shows that EBN-20 has less diversity than

EFTBN-10. EBN-20 has an average diversity of 0.13 while EFTBN-10 has an average diversity of 0.44.

Although EBN-10 has more diversity than EBN-20, it achieves worse results. Its average classification accuracy is 71.56, while the average accuracy of EBN-20 is 71.69. Moreover, EBN-20 achieves better results for 16 data sets, 6 of them are significantly better at 90% confidence level. While ENB-10 achieves better results for 11 data sets only 2 of them are significantly better at 90% confidence level. This result occurred because EBN-20 contains the 10 fine-tuned version of the BN classifiers (in addition to their un-tuned counterparts), while EBN-10 contains only the less accurate un-tuned classifiers.

TABLE I. EBN-10 ENSEMBLE CLASSIFIER COMPARED TO THE 10 INDIVIDUAL BN CLASSIFIERS

Data sets	**C1**	**EBN-10**	**C2**	**EBN-10**	**C3**	**EBN-10**	**C4**	**EBN-10**	**C5**	**EBN-10**	**C6**	**EBN-10**	**C7**	**EBN-10**	**C8**	**EBN-10**	**C9**	**EBN-10**	**C10**	**EBN-10**
	NB		**TAN-TAN Search**		**TAN-K2**		**TAN-Tabu Search**		**TAN-Hill Climber**		**TAN-Repeated Hill Climber**		**BAN-K2**		**BAN-Tabu Search**		**BAN-Hill Climber**		**BAN-Repeated Hill Climber**	
abalone	25.65	**25.82**	24.62	**25.82**	**25.92**	25.82	25.58	**25.82**	**25.85**	25.82	25.65	**25.82**	25.00	**25.82**	**26.98**	25.82	**26.23**	25.82	**26.06**	25.82
auto-mpg	65.00	**72.92**	72.08	**72.92**	72.92	72.92	**73.75**	72.92	**73.75**	72.92	67.50	**72.92**	70.42	**72.92**	61.25	**72.92**	72.92	72.92	63.33	**72.92**
balance-scale	**77.60**	68.64	68.64	68.64	68.64	68.64	**68.80**	68.64	68.64	68.64	**69.28**	68.64	**71.36**	68.64	**70.24**	68.64	**71.52**	68.64	**70.24**	68.64
echocardiogram	**74.32**	72.97	72.97	72.97	72.97	72.97	**77.03**	72.97	72.97	72.97	72.97	72.97	72.97	72.97	72.97	72.97	72.97	72.97	72.97	72.97
breast-tissue-4class	59.43	59.43	**62.26**	59.43	59.43	59.43	59.43	59.43	59.43	59.43	**62.26**	59.43	59.43	59.43	50.94	**59.43**	53.77	**59.43**	48.11	**59.43**
car	73.21	**82.23**	81.71	**82.23**	**82.99**	82.23	**83.74**	82.23	**84.38**	82.23	82.18	**82.23**	66.96	**82.23**	66.90	**82.23**	67.36	**82.23**	66.32	**82.23**
cmc	41.14	**42.57**	**44.26**	42.57	42.36	**42.57**	**44.06**	42.57	**44.13**	42.57	**44.13**	42.57	40.19	**42.57**	37.95	**42.57**	39.92	**42.57**	39.65	**42.57**
column_2c_weka	**77.74**	75.48	**76.45**	75.48	73.87	**75.48**	**84.52**	75.48	75.48	75.48	**84.84**	75.48	73.87	**75.48**	75.16	**75.48**	**75.81**	75.48	65.16	**75.48**
column_3c_weka	60.32	60.32	**63.23**	60.32	**61.29**	60.32	35.16	**60.32**	46.13	**60.32**	**68.71**	60.32	60.32	60.32	53.87	**60.32**	26.77	**60.32**	60.00	**60.32**
dermatology	98.09	**98.36**	98.36	98.36	97.27	**98.36**	98.09	**98.36**	97.81	**98.36**	92.35	**98.36**	95.36	**98.36**	96.99	**98.36**	95.63	**98.36**	88.80	**98.36**
diabetes	78.26	78.13	77.60	**78.13**	77.34	**78.13**	**78.52**	78.13	78.13	78.13	77.60	**78.13**	77.60	**78.13**	**79.17**	78.13	**78.91**	78.13	77.08	**78.13**
disease	30.00	**50.00**	10.00	**50.00**	10.00	**50.00**	50.00	50.00	60.00	**50.00**	40.00	**50.00**	20.00	**50.00**	**60.00**	50.00	**60.00**	50.00	50.00	50.00
ecoli	**80.65**	79.46	79.17	**79.46**	**79.76**	79.46	79.17	**79.46**	79.46	79.46	79.46	79.46	**80.65**	79.46	**80.06**	79.46	79.46	79.46	**81.25**	79.46
fertility	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
glass	70.09	**72.90**	72.43	**72.90**	**74.77**	72.90	72.43	**72.90**	72.43	**72.90**	72.43	**72.90**	72.90	72.90	65.89	**72.90**	67.76	**72.90**	70.56	**72.90**
GL	61.21	**63.08**	61.21	**63.08**	62.15	**63.08**	**66.36**	63.08	**66.36**	63.08	**65.42**	63.08	62.62	**63.08**	52.34	**63.08**	45.79	**63.08**	50.47	**63.08**
graphic.tao.radial	76.80	**84.90**	**87.50**	84.90	84.90	84.90	76.54	**84.90**	84.90	84.90	84.90	84.90	84.90	84.90	76.54	**84.90**	84.90	84.90	84.90	84.90
hay-train	**60.86**	60.59	58.98	**60.59**	59.25	**60.59**	58.98	**60.59**	58.98	**60.59**	58.98	**60.59**	52.82	**60.59**	53.62	**60.59**	52.82	**60.59**	53.89	**60.59**
heart-h	**83.33**	81.97	**83.67**	81.97	**82.99**	81.97	**82.65**	81.97	**82.65**	81.97	**82.31**	81.97	80.61	**81.97**	80.27	**81.97**	75.85	**81.97**	76.19	**81.97**
iris.2D	92.67	92.67	**93.33**	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67	92.67
iris	**94.00**	93.33	93.33	93.33	92.67	**93.33**	93.33	93.33	41.33	**93.33**	41.33	**93.33**	92.67	**93.33**	93.33	93.33	41.33	**93.33**	50.67	**93.33**
landformidentification	98.33	**99.00**	98.33	**99.00**	99.00	99.00	98.33	**99.00**	98.33	**99.00**	99.00	99.00	99.00	99.00	99.00	99.00	98.33	**99.00**	98.00	99.00
led	**74.13**	73.93	**74.03**	73.93	73.78	**73.93**	73.92	**73.93**	73.92	**73.93**	73.92	**73.93**	**73.95**	73.93	73.77	**73.93**	73.90	**73.93**	73.75	**73.93**
lymph	84.46	**86.49**	**87.16**	86.49	**87.16**	86.49	**89.86**	86.49	**87.84**	86.49	83.78	**86.49**	70.27	**86.49**	85.14	**86.49**	70.27	**86.49**	67.57	**86.49**
machine	87.08	87.08	**88.52**	87.08	86.12	**87.08**	84.69	**87.08**	87.08	87.08	83.73	**87.08**	84.69	**87.08**	78.47	**87.08**	77.99	**87.08**	81.34	**87.08**
magic	76.54	**82.00**	**82.22**	82.00	81.39	**82.00**	**82.52**	82.00	**82.74**	82.00	**82.74**	82.00	41.21	**82.00**	73.48	**82.00**	72.47	**82.00**	76.55	**82.00**
nursery	**81.40**	78.06	74.05	**78.06**	75.80	**78.06**	74.26	**78.06**	75.35	**78.06**	75.79	**78.06**	72.56	**78.06**	73.56	**78.06**	73.28	**78.06**	66.15	**78.06**
pendigits	87.98	**97.45**	96.65	**97.45**	97.29	**97.45**	96.69	**97.45**	**97.54**	97.45	97.10	**97.45**	80.67	**97.45**	96.25	**97.45**	92.18	**97.45**	90.71	**97.45**
power_supply	16.24	16.24	15.63	**16.24**	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24	16.24
primary-tumor	45.72	**47.79**	45.13	**47.79**	46.61	**47.79**	46.90	**47.79**	46.02	**47.79**	46.61	**47.79**	28.61	**47.79**	42.77	**47.79**	27.43	**47.79**	26.25	**47.79**
sonar	**83.17**	84.13	79.81	**84.13**	83.17	**84.13**	82.69	**84.13**	83.65	**84.13**	79.81	**84.13**	78.85	**84.13**	84.13	84.13	**85.10**	84.13	82.21	**84.13**
SPECT-Heart	68.91	**69.66**	68.91	**69.66**	69.66	69.66	69.29	**69.66**	67.42	**69.66**	68.16	**69.66**	65.54	**69.66**	68.54	**69.66**	65.92	**69.66**	63.67	**69.66**
SyntheticDataFlow	57.54	**60.86**	**60.93**	60.86	**60.93**	60.86	60.47	**60.86**	**60.94**	60.86	**60.94**	60.86	16.58	**60.86**	25.55	**60.86**	15.71	**60.86**	15.71	**60.86**
tae	28.48	28.48	28.48	28.48	28.48	28.48	19.21	**28.48**	28.48	28.48	28.48	28.48	28.48	28.48	24.50	**28.48**	28.48	28.48	28.48	28.48
titanic	**71.01**	69.47	68.83	**69.47**	68.70	**69.47**	**69.70**	69.47	69.47	69.47	69.47	69.47	**73.69**	69.47	**73.24**	69.47	**76.51**	69.47	**71.38**	69.47
V1	87.36	**90.57**	**91.26**	90.57	**91.49**	90.57	90.57	90.57	89.43	**90.57**	87.82	**90.57**	88.74	**90.57**	90.11	**90.57**	88.74	**90.57**	89.20	**90.57**
waveform	81.60	**86.49**	82.43	**86.49**	85.71	**86.49**	85.89	**86.49**	86.00	**86.49**	85.97	**86.49**	79.97	**86.49**	83.66	**86.49**	78.63	**86.49**	81.43	86.49
wine_quality	48.18	**51.15**	50.77	**51.15**	50.54	**51.15**	50.25	**51.15**	50.25	**51.15**	50.74	**51.15**	46.08	**51.15**	45.23	**51.15**	43.48	**51.15**	45.41	**51.15**
yeast	**59.38**	58.81	**59.00**	58.81	58.71	**58.81**	58.81	58.81	**58.90**	58.81	58.81	58.81	56.02	**58.81**	56.02	**58.81**	56.02	**58.81**	56.40	**58.81**
zoo2_x	95.05	**96.04**	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	91.09	**96.04**	86.14	**96.04**	**98.02**	96.04	92.08	**96.04**	90.10	**96.04**
Average Accuracy	**70.02**	**71.69**	**70.45**	**71.69**	**70.48**	**71.69**	**70.88**	**71.69**	**70.23**	**71.69**	**69.83**	**71.69**	**65.71**	**71.69**	**68.07**	**71.69**	**65.08**	**71.69**	**64.92**	**71.69**
# Better	**12**	**20**	**14**	**19**	**9**	**19**	**11**	**19**	**11**	**15**	**9**	**21**	**4**	**26**	**7**	**26**	**7**	**24**	**4**	**27**
# Sig Better 95%	**1**	**10**	**1**	**5**	**2**	**4**	**2**	**8**	**2**	**5**	**2**	**7**	**1**	**19**	**1**	**16**	**1**	**18**	**0**	**19**
# Sig Better 90%	**3**	**12**	**4**	**9**	**3**	**8**	**4**	**10**	**4**	**9**	**3**	**9**	**3**	**19**	**4**	**18**	**2**	**21**	**0**	**20**

TABLE II. EFTBN-10 ENSEMBLE CLASSIFIER COMPARED TO THE 10 INDIVIDUAL FTBN CLASSIFIERS

Data sets	**C1**	**EFTBN-10**	**C2**	**EFTBN-10**	**C3**	**EFTBN-10**	**C4**	**EFTBN-10**	**C5**	**EFTBN-10**	**C6**	**EFTBN-10**	**C7**	**EFTBN-10**	**C8**	**EFTBN-10**	**C9**	**EFTBN-10**	**C10**	**EFTBN-10**
	FTNB		**FTTAN-TAN Search**		**FTTAN-K2**		**FTTAN-Tabu Search**		**FTTAN-HillClimber**		**FTTAN-Repeated HillClimer**		**FTBAN-K2**		**FTBAN-Tabu Search**		**FTBAN-HillClimber**		**FTBAN-Repeated HillClimber**	
abalone	25.10	**26.03**	24.73	**26.03**	**26.40**	25.82	25.62	**26.03**	**26.03**	26.03	25.65	**26.03**	25.03	**26.03**	**26.95**	26.03	**26.33**	26.03	**26.13**	26.03
auto-mpg	65.42	**72.08**	70.42	**72.08**	72.92	72.08	**72.92**	72.08	**72.50**	72.08	68.75	**72.08**	67.92	**72.08**	65.83	**72.08**	74.17	72.08	69.17	**72.08**
balance-scale	**77.60**	73.12	72.32	73.12	75.04	73.12	**74.08**	73.12	73.28	73.12	**73.92**	73.12	**73.44**	73.12	**74.72**	73.12	**69.60**	73.12	**72.32**	73.12
echocardiogram	**40.91**	75.68	75.68	75.68	75.68	75.68	**79.73**	75.68	75.68	75.68	71.62	75.68	75.68	75.68	75.68	75.68	75.68	75.68	71.62	75.68
breast-tissue-4class	60.38	59.43	**63.21**	59.43	55.66	59.43	59.43	59.43	59.43	59.43	**61.32**	59.43	58.49	59.43	51.89	**59.43**	55.66	**59.43**	47.17	**59.43**
car	76.22	**82.23**	82.35	**82.23**	**83.04**	82.23	**83.33**	82.23	**84.38**	82.23	82.00	**82.23**	66.90	**82.23**	65.34	**82.23**	67.48	**82.23**	65.86	**82.23**
cmc	48.95	**42.50**	**44.40**	42.50	42.50	**42.50**	**44.26**	42.50	**44.13**	42.50	**44.13**	42.50	40.26	**42.50**	38.09	**42.50**	39.92	**42.50**	39.99	**42.50**
column_2c_weka	**74.52**	80.97	**80.65**	80.97	80.00	**80.97**	**84.52**	80.97	76.45	80.97	**85.48**	80.97	79.03	**80.97**	82.26	**80.97**	**81.29**	80.97	69.35	**80.97**
column_3c_weka	60.32	60.00	**85.16**	60.00	**71.94**	60.00	34.52	**60.00**	46.13	**60.00**	**69.03**	60.00	79.03	60.00	53.55	**60.00**	26.77	**60.00**	67.10	**60.00**
dermatology	97.27	**98.36**	98.36	98.36	97.27	**98.36**	98.09	**98.36**	97.81	**98.36**	92.35	**98.36**	95.36	**98.36**	96.99	**98.36**	95.63	**98.36**	88.80	**98.36**
diabetes	78.13	78.13	**78.52**	78.13	77.34	**78.13**	**78.26**	78.13	77.73	**78.13**	77.34	**78.13**	77.47	**78.13**	**79.30**	78.13	**78.78**	78.13	76.82	**78.13**
disease	30.00	**30.00**	10.00	**30.00**	10.00	**30.00**	50.00	30.00	50.00	**30.00**	20.00	**30.00**	20.00	**30.00**	**60.00**	30.00	**60.00**	30.00	50.00	30.00
ecoli	**75.00**	80.36	80.36	**80.36**	**80.65**	80.36	79.76	**80.36**	80.36	80.36	80.36	80.36	**80.65**	80.36	**78.87**	80.36	80.36	80.36	**81.25**	80.36
fertility	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
glass	70.09	**72.43**	71.50	**72.43**	**72.90**	72.43	71.96	**72.43**	71.96	**72.43**	71.03	**72.43**	72.90	72.43	61.68	**72.43**	64.49	**72.43**	66.36	**72.43**
GL	64.02	**60.28**	56.07	**60.28**	57.94	**60.28**	**66.36**	60.28	**66.36**	60.28	**65.42**	60.28	63.55	**60.28**	51.40	**60.28**	46.73	**60.28**	52.34	**60.28**
graphic.tao.radial	84.38	**84.64**	**76.54**	84.64	84.64	84.64	78.97	**84.64**	84.64	84.64	84.64	84.64	84.64	84.64	78.97	**84.64**	84.64	84.64	84.64	84.64
hay-train	**61.66**	60.59	58.98	**60.59**	59.79	**60.59**	59.79	**60.59**	59.79	**60.59**	59.79	**60.59**	51.74	**60.59**	52.55	**60.59**	52.01	**60.59**	53.35	**60.59**
heart-h	**82.99**	83.33	**82.65**	83.33	**82.31**	83.33	**82.99**	83.33	**82.65**	83.33	**83.33**	83.33	79.93	**83.33**	79.59	**83.33**	73.47	**83.33**	77.55	**83.33**
iris.2D	97.33	97.33	**93.33**	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33	97.33
iris	**95.33**	93.33	94.66	93.33	94.00	**93.33**	94.00	93.33	42.00	**93.33**	42.00	**93.33**	93.33	**93.33**	94.00	93.33	42.00	**93.33**	51.33	**93.33**
landformidentification	98.66	**99.00**	98.33	**99.00**	99.00	99.00	98.33	**99.00**	98.33	**99.00**	99.00	99.00	99.00	99.00	99.00	99.00	98.33	**99.00**	98.00	99.00
led	**74.03**	74.15	**73.97**	74.15	73.78	**74.15**	73.90	**74.15**	74.02	**74.15**	74.00	**74.15**	**73.97**	74.15	73.85	**74.15**	73.95	**74.15**	73.53	**74.15**
lymph	85.81	**86.49**	**87.16**	86.49	**87.16**	86.49	**89.19**	86.49	**87.84**	86.49	83.78	**86.49**	70.27	**86.49**	85.14	**86.49**	70.27	**86.49**	67.57	**86.49**
machine	86.60	86.12	**88.52**	86.12	86.12	**86.12**	84.69	**86.12**	86.60	86.12	83.25	**86.12**	84.69	**86.12**	79.43	**86.12**	77.99	**86.12**	82.30	**86.12**
magic	53.74	**82.06**	**82.23**	82.06	81.37	**82.06**	**82.55**	82.06	**82.75**	82.06	**82.75**	82.06	41.36	**82.06**	73.47	**82.06**	72.47	**82.06**	76.54	**82.06**
nursery	**84.74**	77.99	73.81	**77.99**	75.91	**77.99**	74.05	**77.99**	75.30	**77.99**	75.69	**77.99**	72.61	**77.99**	73.69	**77.99**	73.68	**77.99**	66.12	**77.99**
pendigits	93.73	**97.45**	96.65	**97.45**	97.29	**97.45**	96.69	**97.45**	**97.54**	97.45	97.10	**97.45**	80.67	**97.45**	96.25	**97.45**	92.18	**97.45**	90.71	**97.45**
power_supply	14.73	16.19	15.70	**16.19**	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19	16.19
primary-tumor	44.25	**46.61**	44.54	**46.61**	46.31	**46.61**	46.02	**46.61**	46.31	**46.61**	46.61	**46.61**	28.32	**46.61**	43.07	**46.61**	27.73	**46.61**	25.66	**46.61**
sonar	**83.17**	84.62	81.25	**84.62**	83.65	**84.62**	83.17	**84.62**	84.62	**84.62**	79.33	**84.62**	79.33	**84.62**	84.13	84.62	**85.10**	84.62	82.21	**84.62**
SPECT-Heart	70.04	**68.54**	67.79	**68.54**	68.54	68.54	68.54	**68.54**	67.04	**68.54**	68.54	**68.54**	64.42	**68.54**	67.79	**68.54**	65.92	**68.54**	62.55	**68.54**
SyntheticDataFlow	54.45	**60.86**	**60.93**	60.86	**60.93**	60.86	60.48	**60.86**	**60.94**	60.86	**60.94**	60.86	16.58	**60.86**	25.56	**60.86**	15.72	**60.86**	15.72	**60.86**
tae	26.49	62.25	62.25	62.25	62.25	62.25	34.44	**62.25**	62.25	62.25	62.25	62.25	62.25	62.25	56.29	**62.25**	62.25	62.25	62.25	62.25
titanic	70.06	**74.56**	71.42	**74.56**	74.33	**74.56**	71.97	**74.56**	44.30	**74.56**	44.30	**74.56**	73.97	**74.56**	53.48	**74.56**	73.97	**74.56**	67.83	**74.56**
V1	87.82	**90.34**	**90.80**	90.34	**91.26**	90.34	90.11	90.34	89.66	**90.34**	87.82	**90.34**	88.51	**90.34**	90.11	**90.34**	87.82	**90.34**	89.20	**90.34**
waveform	85.43	**86.49**	82.43	**86.49**	85.71	**86.49**	85.89	**86.49**	86.00	**86.49**	85.97	**86.49**	79.97	**86.49**	83.66	**86.49**	78.63	**86.49**	81.43	86.49
wine_quality	48.76	**51.33**	50.74	**51.33**	50.28	**51.33**	50.45	**51.33**	50.45	**51.33**	50.80	**51.33**	46.14	**51.33**	45.14	**51.33**	43.66	**51.33**	45.35	**51.33**
yeast	**58.81**	59.00	**58.61**	59.00	58.81	**59.00**	58.81	59.00	**58.90**	59.00	58.90	59.00	56.59	**59.00**	53.90	**59.00**	55.05	**59.00**	56.40	**59.00**
zoo2_x	91.09	**96.04**	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	91.09	**96.04**	86.14	**96.04**	**98.02**	96.04	92.08	**96.04**	90.10	**96.04**
Average Accuracy	**69.15**	**72.47**	**71.78**	**72.47**	**72.01**	**72.47**	**71.64**	**72.47**	**70.54**	**72.47**	**69.80**	**72.47**	**67.29**	**72.47**	**68.78**	**72.47**	**66.08**	**72.47**	**66.15**	**72.47**
# Better	**12**	**25**	**12**	**22**	**13**	**16**	**13**	**21**	**12**	**17**	**9**	**21**	**7**	**25**	**7**	**27**	**7**	**26**	**5**	**29**
# Sig Better 95%	**3**	**8**	**3**	**6**	**1**	**4**	**3**	**9**	**2**	**5**	**2**	**6**	**1**	**17**	**2**	**12**	**2**	**19**	**2**	**16**
# Sig Better 90%	**3**	**9**	**5**	**10**	**2**	**6**	**4**	**11**	**4**	**9**	**4**	**11**	**1**	**18**	**4**	**19**	**2**	**21**	**2**	**21**

TABLE III. EBN-20 ENSEMBLE CLASSIFIER COMPARED TO THE 20 INDIVIDUAL CLASSIFIERS

Data sets	C1		C2		C3		C4		C5		C6		C7		C8		C9		C10	
	NB	EBN-20	FTNB	EBN-20	TAN-TAN Search	EBN-20	FTTAN-TAN Search	EBN-20	TAN-K2	EBN-20	FTTAN-K2	EBN-20	TAN-Tabu Search	EBN-20	FTTAN-Tabu Search	EBN-20	TAN-HillClimber	EBN-20	FTTAN-HillClimber	EBN-20
abalone	25.65	**26.09**	25.10	**26.09**	24.62	**26.09**	24.73	**26.09**	25.92	**26.09**	**26.40**	26.09	25.58	**26.09**	25.62	**26.09**	25.85	**26.09**	26.03	**26.09**
auto-mpg	65.00	**72.50**	65.42	**72.50**	72.08	**72.50**	70.42	**72.50**	**72.92**	72.50	**72.92**	72.50	**73.75**	72.50	**72.92**	72.50	**73.75**	72.50	72.50	72.50
balance-scale	**77.60**	69.12	**77.60**	69.12	68.64	**69.12**	**72.32**	69.12	68.64	**69.12**	**75.04**	69.12	68.80	**69.12**	**74.08**	69.12	68.64	**69.12**	**73.28**	69.12
echocardiogram	**74.32**	72.97	40.91	**72.97**	72.97	72.97	**75.68**	72.97	72.97	72.97	**75.68**	72.97	**77.03**	72.97	**79.73**	72.97	72.97	72.97	**75.68**	72.97
breast-tissue-4class	59.43	59.43	**60.38**	59.43	**62.26**	59.43	**63.21**	59.43	59.43	59.43	55.66	**59.43**	59.43	59.43	59.43	59.43	59.43	59.43	59.43	59.43
car	73.21	**82.12**	76.22	**82.12**	81.71	**82.12**	**82.35**	82.12	**82.99**	82.12	**83.04**	82.12	**83.74**	82.12	**83.33**	82.12	**84.38**	82.12	**84.38**	82.12
cmc	41.14	**42.57**	**48.95**	42.57	**44.26**	42.57	**44.40**	42.57	42.36	**42.57**	42.50	**42.57**	**44.06**	42.57	**44.26**	42.57	**44.13**	42.57	**44.13**	42.57
column_2c_weka	77.74	**80.00**	74.52	**80.00**	76.45	**80.00**	**80.65**	80.00	73.87	**80.00**	80.00	80.00	**84.52**	80.00	**84.52**	80.00	75.48	**80.00**	76.45	**80.00**
column_3c_weka	**60.32**	60.00	**60.32**	60.00	**63.23**	60.00	**85.16**	60.00	**61.29**	60.00	**71.94**	60.00	35.16	**60.00**	34.52	**60.00**	46.13	**60.00**	46.13	**60.00**
dermatology	97.27	**98.36**	97.27	**98.36**	98.36	98.36	98.36	98.36	97.27	**98.36**	97.27	**98.36**	98.09	**98.36**	98.09	**98.36**	97.81	**98.36**	97.81	**98.36**
diabetes	**78.26**	78.13	78.13	78.13	77.60	**78.13**	**78.52**	78.13	77.34	**78.13**	77.34	**78.13**	**78.52**	78.13	**78.26**	78.13	78.13	78.13	77.73	**78.13**
disease	30.00	30.00	30.00	30.00	10.00	**30.00**	10.00	**30.00**	10.00	**30.00**	10.00	**30.00**	**50.00**	30.00	**50.00**	30.00	**60.00**	30.00	**50.00**	30.00
ecoli	**80.65**	80.36	75.00	**80.36**	79.17	**80.36**	80.36	80.36	79.76	**80.36**	**80.65**	80.36	79.17	**80.36**	79.76	**80.36**	79.46	**80.36**	80.36	80.36
fertility	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
glass	70.09	**74.30**	70.09	**74.30**	72.43	**74.30**	71.50	**74.30**	**74.77**	74.30	72.90	**74.30**	72.43	**74.30**	71.96	**74.30**	72.43	**74.30**	71.96	**74.30**
GL	61.21	**62.15**	**64.02**	62.15	61.21	**62.15**	56.07	**62.15**	62.15	62.15	57.94	**62.15**	**66.36**	62.15	66.36	62.15	66.36	**62.15**	66.36	62.15
graphic.tao.radial	76.80	**86.44**	84.38	**86.44**	**87.50**	86.44	76.54	**86.44**	84.90	**86.44**	84.64	**86.44**	76.54	**86.44**	78.97	**86.44**	84.90	**86.44**	84.64	**86.44**
hay-train	**60.86**	60.05	**61.66**	60.05	58.98	**60.05**	58.98	**60.05**	59.25	**60.05**	59.79	**60.05**	58.98	**60.05**	59.79	**60.05**	58.98	**60.05**	59.79	**60.05**
heart-h	**83.33**	82.65	**82.99**	82.65	**83.67**	82.65	82.65	82.65	**82.99**	82.65	82.31	**82.65**	82.65	82.65	**82.99**	82.65	82.65	82.65	82.65	82.65
iris.2D	92.67	**93.33**	**97.33**	93.33	93.33	93.33	93.33	93.33	92.67	**93.33**	**97.33**	93.33	92.67	**93.33**	**97.33**	93.33	92.67	**93.33**	**97.33**	93.33
iris	**94.00**	93.33	**95.33**	93.33	93.33	93.33	**94.66**	93.33	92.67	**93.33**	**94.00**	93.33	93.33	93.33	**94.00**	93.33	41.33	**93.33**	42.00	**93.33**
landformidentification	98.33	**99.00**	98.66	**99.00**	98.33	**99.00**	98.33	**99.00**	99.00	99.00	99.00	99.00	98.33	**99.00**	98.33	**99.00**	98.33	**99.00**	98.33	**99.00**
led	**74.13**	73.98	**74.03**	73.98	**74.03**	73.98	73.97	**73.98**	73.78	**73.98**	73.78	73.98	73.92	**73.98**	73.90	**73.98**	73.92	**73.98**	**74.02**	73.98
lymph	84.46	**86.49**	85.81	**86.49**	**87.16**	86.49	**87.16**	86.49	**87.16**	86.49	**87.16**	86.49	**89.86**	86.49	**89.19**	86.49	**87.84**	86.49	**87.84**	86.49
machine	**87.08**	86.60	86.60	86.60	**88.52**	86.60	**88.52**	86.60	86.12	86.60	86.12	86.60	84.69	**86.60**	84.69	**86.60**	**87.08**	86.60	86.60	86.60
magic	76.54	**82.05**	53.74	**82.05**	**82.22**	82.05	**82.23**	82.05	81.39	**82.05**	81.37	**82.05**	**82.52**	82.05	**82.55**	82.05	**82.74**	82.05	**82.75**	82.05
nursery	**81.40**	78.09	**84.74**	78.09	74.05	**78.09**	73.81	**78.09**	75.80	**78.09**	75.91	**78.09**	74.26	**78.09**	74.05	**78.09**	75.35	**78.09**	75.30	**78.09**
pendigits	87.98	**97.45**	93.73	**97.45**	96.65	**97.45**	96.65	**97.45**	97.29	**97.45**	97.29	**97.45**	96.69	**97.45**	96.69	**97.45**	**97.54**	97.45	**97.54**	97.45
power_supply	**16.24**	16.14	14.73	**16.14**	15.63	**16.14**	15.70	**16.14**	**16.24**	16.14	**16.19**	16.14	**16.24**	16.14	**16.19**	16.14	**16.24**	16.14	**16.19**	16.14
primary-tumor	45.72	**47.49**	44.25	**47.49**	45.13	**47.49**	44.54	**47.49**	46.61	**47.49**	46.31	**47.49**	46.90	**47.49**	46.02	**47.49**	46.02	**47.49**	46.31	**47.49**
sonar	83.17	**84.62**	83.17	**84.62**	79.81	**84.62**	81.25	**84.62**	83.17	**84.62**	83.65	**84.62**	82.69	**84.62**	83.17	**84.62**	83.65	**84.62**	84.62	84.62
SPECT-Heart	**68.91**	68.54	**70.04**	68.54	68.91	**68.54**	67.79	**68.54**	**69.66**	68.54	68.54	68.54	**69.29**	68.54	68.54	68.54	67.42	**68.54**	67.04	**68.54**
SyntheticDataFlow	57.54	**60.86**	54.45	**60.86**	**60.93**	60.86	**60.93**	60.86	**60.93**	60.86	**60.93**	60.86	60.47	**60.86**	60.48	**60.86**	**60.94**	60.86	**60.94**	60.86
tae	28.48	**31.79**	26.49	**31.79**	28.48	**31.79**	**62.25**	31.79	28.48	**31.79**	**62.25**	31.79	19.21	**31.79**	**34.44**	31.79	28.48	**31.79**	**62.25**	31.79
titanic	71.01	**74.56**	70.06	**74.56**	68.83	**74.56**	71.42	**74.56**	68.70	**74.56**	74.33	**74.56**	69.70	**74.56**	71.97	**74.56**	69.47	**74.56**	44.30	**74.56**
V1	87.36	**90.11**	87.82	**90.11**	**91.26**	90.11	**90.80**	90.11	**91.49**	90.11	**91.26**	90.11	**90.57**	90.11	90.11	90.11	89.43	**90.11**	89.66	90.11
waveform	81.60	**86.49**	85.43	**86.49**	82.43	**86.49**	82.43	**86.49**	85.71	**86.49**	85.71	**86.49**	85.89	**86.49**	85.89	**86.49**	86.00	**86.49**	86.00	**86.49**
wine_quality	48.18	**51.24**	48.76	**51.24**	50.77	**51.24**	50.74	**51.24**	50.54	**51.24**	50.28	**51.24**	50.25	**51.24**	50.45	**51.24**	50.25	**51.24**	50.45	**51.24**
yeast	**59.38**	59.00	58.81	**59.00**	59.00	59.00	58.61	**59.00**	58.71	**59.00**	58.81	**59.00**	58.81	**59.00**	58.81	**59.00**	58.90	**59.00**	58.90	**59.00**
zoo2_x	95.05	**96.04**	91.09	**96.04**	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04	96.04
Average Accuracy	**70.00**	**71.56**	**69.15**	**71.56**	**70.45**	**71.56**	**71.78**	**71.56**	**70.48**	**71.56**	**72.01**	**71.56**	**70.88**	**71.56**	**71.64**	**71.56**	**70.23**	**71.56**	**70.54**	**71.56**
# Better	**14**	**23**	**12**	**24**	**11**	**22**	**15**	**19**	**10**	**23**	**14**	**19**	**13**	**22**	**15**	**19**	**10**	**24**	**13**	**17**
# Sig Better 95%	**1**	**8**	**4**	**10**	**4**	**5**	**4**	**6**	**2**	**6**	**5**	**4**	**2**	**10**	**4**	**10**	**4**	**6**	**6**	**7**
# Sig Better 90%	**3**	**11**	**5**	**13**	**4**	**12**	**6**	**11**	**2**	**10**	**6**	**7**	**3**	**14**	**5**	**11**	**4**	**13**	**8**	**10**

TABLE IV. CONTINUED - FTBN-20 ENSEMBLE CLASSIFIER COMPARED TO THE 20 INDIVIDUAL CLASSIFIERS

Data sets	**C11**	**EBN-20**	**C12**	**EBN-20**	**C13**	**EBN-20**	**C14**	**EBN-20**	**C15**	**EBN-20**	**C16**	**EBN-20**	**C17**	**EBN-20**	**C18**	**EBN-20**	**C19**	**EBN-20**	**C20**	**EBN-20**
	TAN-RepeatedHillClimber		**FTTAN-RepeatedHillClimber**		**BAN-K2**		**FTBAN-K2**		**BAN-TabuSearch**		**FTBAN-TabuSearch**		**BAN-HillClimber**		**FTBAN-HillClimber**		**BAN-RepeatedHillClimber**		**FTBAN-RepeatedHillClimber**	
abalone	25.65	**26.09**	25.65	**26.09**	25.00	**26.09**	25.03	**26.09**	**26.98**	26.09	**26.95**	26.09	**26.23**	26.09	**26.33**	26.09	26.06	**26.09**	**26.13**	26.09
auto-mpg	67.50	**72.50**	68.75	**72.50**	70.42	**72.50**	67.92	**72.50**	61.25	**72.50**	65.83	**72.50**	**72.92**	72.50	**74.17**	72.50	63.33	**72.50**	69.17	**72.50**
balance-scale	**69.28**	69.12	**73.92**	69.12	**71.36**	69.12	**73.44**	69.12	**70.24**	69.12	**74.72**	69.12	**71.52**	69.12	**69.60**	69.12	**70.24**	69.12	**72.32**	69.12
echocardiogram	72.97	72.97	71.62	**72.97**	72.97	72.97	**75.68**	72.97	72.97	72.97	**75.68**	72.97	72.97	72.97	**75.68**	72.97	72.97	72.97	71.62	**72.97**
breast-tissue-4class	**62.26**	59.43	**61.32**	59.43	59.43	59.43	58.49	**59.43**	50.94	**59.43**	51.89	**59.43**	53.77	**59.43**	55.66	**59.43**	48.11	**59.43**	47.17	**59.43**
car	**82.18**	82.12	82.00	**82.12**	66.96	**82.12**	66.90	**82.12**	66.90	**82.12**	65.34	**82.12**	67.36	**82.12**	67.48	**82.12**	66.32	**82.12**	65.86	**82.12**
cmc	**44.13**	42.57	**44.13**	42.57	40.19	**42.57**	40.26	**42.57**	37.95	**42.57**	38.09	**42.57**	39.92	**42.57**	39.92	**42.57**	39.65	**42.57**	39.99	**42.57**
column_2c_weka	**84.84**	80.00	**85.48**	80.00	73.87	**80.00**	79.03	**80.00**	75.16	**80.00**	**82.26**	80.00	75.81	**80.00**	**81.29**	80.00	65.16	**80.00**	69.35	**80.00**
column_3c_weka	**68.71**	60.00	**69.03**	60.00	**60.32**	60.00	**79.03**	60.00	53.87	**60.00**	53.55	**60.00**	26.77	**60.00**	26.77	**60.00**	60.00	60.00	**67.10**	60.00
dermatology	92.35	**98.36**	92.35	**98.36**	95.36	**98.36**	95.36	**98.36**	96.99	**98.36**	96.99	**98.36**	95.63	**98.36**	95.63	**98.36**	88.80	**98.36**	88.80	**98.36**
diabetes	77.60	**78.13**	77.34	**78.13**	77.60	**78.13**	77.47	**78.13**	**79.17**	78.13	**79.30**	78.13	**78.91**	78.13	**78.78**	78.13	77.08	**78.13**	76.82	**78.13**
disease	**40.00**	30.00	20.00	**30.00**	20.00	**30.00**	20.00	**30.00**	**60.00**	30.00	**60.00**	30.00	**60.00**	30.00	**60.00**	30.00	**50.00**	30.00	**50.00**	30.00
ecoli	79.46	**80.36**	80.36	80.36	**80.65**	80.36	**80.65**	80.36	80.06	**80.36**	78.87	**80.36**	79.46	**80.36**	80.36	80.36	**81.25**	80.36	**81.25**	80.36
fertility	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00	88.00
glass	72.43	**74.30**	71.03	**74.30**	72.90	**74.30**	72.90	**74.30**	65.89	**74.30**	61.68	**74.30**	67.76	**74.30**	64.49	**74.30**	70.56	**74.30**	66.36	**74.30**
GL	**65.42**	62.15	**65.42**	62.15	**62.62**	62.15	**63.55**	62.15	52.34	**62.15**	51.40	**62.15**	45.79	**62.15**	46.73	**62.15**	50.47	**62.15**	52.34	**62.15**
graphic.tao.radial	84.90	**86.44**	84.64	**86.44**	84.90	**86.44**	84.64	**86.44**	76.54	**86.44**	78.97	**86.44**	84.90	**86.44**	84.64	**86.44**	84.90	**86.44**	84.64	**86.44**
hay-train	58.98	**60.05**	59.79	**60.05**	52.82	**60.05**	51.74	**60.05**	53.62	**60.05**	52.55	**60.05**	52.82	**60.05**	52.01	**60.05**	53.89	**60.05**	53.35	**60.05**
heart-h	82.31	**82.65**	**83.33**	82.65	80.61	**82.65**	79.93	**82.65**	80.27	**82.65**	79.59	**82.65**	75.85	**82.65**	73.47	**82.65**	76.19	**82.65**	77.55	**82.65**
iris.2D	92.67	**93.33**	**97.33**	93.33	92.67	**93.33**	**97.33**	93.33	92.67	**93.33**	**97.33**	93.33	92.67	**93.33**	**97.33**	93.33	92.67	**93.33**	**97.33**	93.33
iris	41.33	**93.33**	42.00	**93.33**	92.67	**93.33**	93.33	93.33	93.33	93.33	**94.00**	93.33	41.33	**93.33**	42.00	**93.33**	50.67	**93.33**	51.33	**93.33**
landformidentification	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	98.33	**99.00**	98.33	**99.00**	98.00	**99.00**	98.00	**99.00**
led	73.92	**73.98**	**74.00**	73.98	73.95	**73.98**	73.97	**73.98**	73.77	**73.98**	73.85	**73.98**	73.90	**73.98**	73.95	**73.98**	73.75	**73.98**	73.53	**73.98**
lymph	83.78	**86.49**	83.78	**86.49**	70.27	**86.49**	70.27	**86.49**	85.14	**86.49**	85.14	**86.49**	70.27	**86.49**	70.27	**86.49**	67.57	**86.49**	67.57	**86.49**
machine	83.73	**86.60**	83.25	**86.60**	84.69	**86.60**	84.69	**86.60**	78.47	**86.60**	79.43	**86.60**	77.99	**86.60**	77.99	**86.60**	81.34	**86.60**	82.30	**86.60**
magic	**82.74**	82.05	**82.75**	82.05	41.21	**82.05**	41.36	**82.05**	73.48	**82.05**	73.47	**82.05**	72.47	**82.05**	72.47	**82.05**	76.55	**82.05**	76.54	**82.05**
nursery	75.79	**78.09**	75.69	**78.09**	72.56	**78.09**	72.61	**78.09**	73.56	**78.09**	73.69	**78.09**	73.28	**78.09**	73.68	**78.09**	66.15	**78.09**	66.12	**78.09**
pendigits	97.10	**97.45**	97.10	**97.45**	**80.67**	97.45	**80.67**	97.45	96.25	**97.45**	96.25	**97.45**	92.18	**97.45**	92.18	**97.45**	90.71	**97.45**	90.71	**97.45**
power_supply	**16.24**	16.14	**16.19**	16.14	**16.24**	16.14	**16.19**	16.14	**16.24**	16.14	**16.19**	16.14	**16.24**	16.14	**16.19**	16.14	**16.24**	16.14	**16.19**	16.14
primary-tumor	46.61	**47.49**	46.61	**47.49**	28.61	**47.49**	28.32	**47.49**	42.77	**47.49**	43.07	**47.49**	27.43	**47.49**	27.73	**47.49**	26.25	**47.49**	25.66	**47.49**
sonar	79.81	**84.62**	79.33	**84.62**	78.85	**84.62**	79.33	**84.62**	84.13	**84.62**	84.13	**84.62**	**85.10**	84.62	**85.10**	84.62	82.21	**84.62**	82.21	**84.62**
SPECT-Heart	68.16	**68.54**	68.54	68.54	65.54	**68.54**	64.42	**68.54**	68.54	**68.54**	67.79	**68.54**	65.92	**68.54**	65.92	**68.54**	63.67	**68.54**	62.55	**68.54**
SyntheticDataFlow	**60.94**	60.86	**60.94**	60.86	16.58	**60.86**	16.58	**60.86**	25.55	**60.86**	25.56	**60.86**	15.71	**60.86**	15.72	**60.86**	15.71	**60.86**	15.72	**60.86**
tae	28.48	**31.79**	**62.25**	31.79	28.48	**31.79**	**62.25**	31.79	24.50	**31.79**	**56.29**	31.79	28.48	**31.79**	**62.25**	31.79	28.48	**31.79**	**62.25**	31.79
titanic	69.47	**74.56**	44.30	**74.56**	73.69	**74.56**	73.97	74.56	73.24	**74.56**	53.48	**74.56**	**76.51**	74.56	73.97	**74.56**	71.38	**74.56**	67.83	**74.56**
V1	87.82	**90.11**	87.82	**90.11**	88.74	**90.11**	88.51	**90.11**	90.11	90.11	90.11	90.11	88.74	**90.11**	87.82	**90.11**	89.20	**90.11**	89.20	**90.11**
waveform	85.97	**86.49**	85.97	**86.49**	**79.97**	86.49	**79.97**	86.49	83.66	**86.49**	83.66	**86.49**	78.63	**86.49**	78.63	**86.49**	81.43	**86.49**	81.43	**86.49**
wine_quality	50.74	**51.24**	50.80	**51.24**	46.08	**51.24**	46.14	**51.24**	45.23	**51.24**	45.14	**51.24**	43.48	**51.24**	43.66	**51.24**	45.41	**51.24**	45.35	**51.24**
yeast	58.81	**59.00**	58.90	**59.00**	56.02	**59.00**	56.59	**59.00**	58.81	**59.00**	58.81	**59.00**	56.02	**59.00**	55.05	**59.00**	56.40	**59.00**	56.40	**59.00**
zoo2_x	91.09	**96.04**	91.09	**96.04**	86.14	**96.04**	86.14	**96.04**	**98.02**	96.04	**98.02**	96.04	92.08	**96.04**	92.08	**96.04**	90.10	**96.04**	90.10	**96.04**
Average Accuracy	**69.83**	**71.56**	**69.80**	**71.56**	**65.71**	**71.56**	**67.29**	**71.56**	**68.14**	**71.56**	**68.90**	**71.56**	**65.08**	**71.56**	**66.08**	**71.56**	**64.92**	**71.56**	**66.15**	**71.56**
# Better	**11**	**26**	**13**	**23**	**7**	**29**	**10**	**26**	**6**	**29**	**11**	**26**	**8**	**29**	**11**	**27**	**4**	**33**	**8**	**31**
# Sig Better 95%	**2**	**6**	**5**	**7**	**2**	**13**	**6**	**16**	**1**	**15**	**4**	**14**	**1**	**18**	**3**	**18**	**0**	**20**	**3**	**18**
# Sig Better 90%	**4**	**13**	**6**	**11**	**2**	**17**	**7**	**16**	**2**	**19**	**8**	**17**	**1**	**21**	**4**	**19**	**0**	**22**	**5**	**22**

TABLE V. COMPARISONS BETWEEN THE THREE ENSEMBLE CLASSIFIERS (EBN-20, EBN-10 AND EFTBN-10)

Data sets	**EBN-20 vs EBN-10**				**EBN-20 vs EFTBN-10**				**EBN-10 vs EFTBN-10**	
	EBN-20	**Diversity**	**EBN- 10**	**Diversity**	**EBN-20**	**Diversity**	**EFTBN- 10**	**Diversity**	**EBN- 10**	**EFTBN-10**
abalone	**26.09**	0.00	25.82	0.50	**26.09**	0.00	26.03	0.50	25.82	**26.03**
auto-mpg	72.50	0.21	**72.92**	0.48	**72.50**	0.21	72.08	0.50	**72.92**	72.08
balance-scale	**69.12**	0.12	68.64	0.42	69.12	0.12	**73.12**	0.44	68.64	**73.12**
echocardiogram	72.97	0.25	72.97	0.50	72.97	0.25	**75.68**	0.50	72.97	**75.68**
breast-tissue-4class	59.43	0.12	59.43	0.39	59.43	0.12	59.43	0.41	59.43	59.43
car	82.12	0.20	**82.23**	0.45	82.12	0.20	**82.23**	0.45	82.23	82.23
cmc	42.57	0.30	42.57	0.54	**42.57**	0.30	42.50	0.54	**42.57**	42.50
column_2c_weka	**80.00**	0.17	75.48	0.50	80.00	0.17	**80.97**	0.50	75.48	**80.97**
column_3c_weka	60.00	0.33	**60.32**	0.57	60.00	0.33	60.00	0.54	**60.32**	60.00
dermatology	98.36	0.03	98.36	0.33	98.36	0.03	98.36	0.33	98.36	98.36
diabetes	78.13	0.07	78.13	0.50	78.13	0.07	78.13	0.50	78.13	78.13
disease	30.00	0.55	**50.00**	0.50	30.00	0.55	30.00	0.50	**50.00**	30.00
ecoli	**80.36**	0.01	79.46	0.32	80.36	0.01	80.36	0.32	79.46	**80.36**
fertility	88.00	0.00	88.00	0.50	88.00	0.00	88.00	0.50	88.00	88.00
glass	**74.30**	0.07	72.90	0.37	**74.30**	0.07	72.43	0.39	**72.90**	72.43
GL	62.15	0.09	**63.08**	0.39	**62.15**	0.09	60.28	0.41	**63.08**	60.28
graphic.tao.radial	**86.44**	0.14	84.90	0.50	**86.44**	0.14	84.64	0.50	**84.90**	84.64
hay-train	60.05	0.19	**60.59**	0.45	60.05	0.19	**60.59**	0.45	60.59	60.59
heart-h	**82.65**	0.05	81.97	0.34	82.65	0.05	**83.33**	0.35	81.97	**83.33**
iris.2D	**93.33**	0.02	92.67	0.39	93.33	0.02	**97.33**	0.39	92.67	**97.33**
iris	93.33	0.27	93.33	0.55	93.33	0.27	93.33	0.56	93.33	93.33
landformidentification	99.00	0.00	99.00	0.31	99.00	0.00	99.00	0.31	99.00	99.00
led	**73.98**	0.01	73.93	0.29	73.98	0.01	**74.15**	0.29	73.93	**74.15**
lymph	86.49	0.13	86.49	0.41	86.49	0.13	86.49	0.41	86.49	86.49
machine	86.60	0.04	**87.08**	0.34	**86.60**	0.04	86.12	0.34	**87.08**	86.12
magic	**82.05**	0.28	82.00	0.50	82.05	0.28	**82.06**	0.50	82.00	**82.06**
nursery	**78.09**	0.14	78.06	0.41	78.09	0.14	77.99	0.41	78.06	77.99
pendigits	97.45	0.04	97.45	0.30	97.45	0.04	97.45	0.30	97.45	97.45
power_supply	16.14	0.00	**16.24**	0.50	16.14	0.00	**16.19**	0.50	**16.24**	16.19
primary-tumor	47.49	0.00	**47.79**	0.50	**47.49**	0.00	46.61	0.50	**47.79**	46.61
sonar	**84.62**	0.14	84.13	0.50	84.62	0.14	84.62	0.50	84.13	**84.62**
SPECT-Heart	68.54	0.25	**69.66**	0.50	68.54	0.25	68.54	0.50	**69.66**	68.54
SyntheticDataFlow	60.86	0.28	60.86	0.53	60.86	0.28	60.86	0.53	60.86	60.86
tae	**31.79**	0.38	28.48	0.41	31.79	0.38	**62.25**	0.44	28.48	**62.25**
titanic	**74.56**	0.17	69.47	0.50	74.56	0.17	74.56	0.50	69.47	**74.56**
V1	90.11	0.09	**90.57**	0.50	90.11	0.09	**90.34**	0.50	**90.57**	90.34
waveform	86.49	0.12	86.49	0.45	86.49	0.12	86.49	0.45	86.49	86.49
wine_quality	**51.24**	0.06	51.15	0.41	51.24	0.06	**51.33**	0.41	51.15	**51.33**
yeast	**59.00**	0.03	58.81	0.29	59.00	0.03	59.00	0.30	58.81	**59.00**
zoo2_x	96.04	0.04	96.04	0.36	96.04	0.04	96.04	0.36	96.04	96.04
Average	**71.56**	**0.13**	**71.69**	**0.44**	**71.56**	**0.13**	**72.47**	**0.44**	**71.69**	**72.47**
# Better	**16**		**11**		**8**		**13**		**12**	**13**
# Sig Better 95%	**0**		**0**		**2**		**3**		**1**	**6**
# Sig Better 90%	**6**		**2**		**2**		**8**		**4**	**9**

V. CONCLUSION

This work shows that an ensemble of fine tune BN classifiers is an effective way to increase the classification accuracy of BN classifiers. It also empirically concludes that the ensemble of the fine-tuned classifiers outperforms an ensemble of un-tuned classifiers. Although the two ensembles have the same average diversity, the ensemble of the fine-tuned classifiers combines more accurate classifiers. However, constructing a larger ensemble that combines the fine-tuned and un-tuned classifiers does not improve the classification accuracy because the combined classifiers are not very different. The work also proposes a distance-based diversity measure and uses it in analyzing the results. The ensemble of classifiers combines different types of BN classifiers (NB, TAN, and BAN). Different learning algorithms that use different search methods were used to build TAN and BAN classifiers. The variation of the BN classifiers increases the diversity of the ensemble while using fine-tuned classifiers increase accuracy of the constituent classifiers. The work compares between three different ensembles of BN classifiers. The first ensemble, EBN-10, combines ten un-tuned classifiers; the second, EFTBN-10, combines ten fine-tuned BN classifiers while the third ensemble combines all the previous 20 BN classifiers (EBN-20). The experimental results using 40 data sets and a simple majority voting method shows that the ensembles outperform all the individual constituent classifiers. It also states that the EFTBN-10 is the superior one because it has more accurate constituents and is more diverse.

VI. Future Work

As a future work, we intend to develop ensembles of BN classifiers by using different other BN classifiers. Also it is interesting to develop a new fine tuning algorithm to improve the accuracy of the ensemble base classifiers. Moreover, different ensemble method and voting techniques can be used.

References

[1] Akhlaqur Rahman and Sumaira Tasnim, "Ensemble Classifiers and Their Applications: A Review," International Journal of Computer Trends and Technology (IJCTT), vol. 10, no. 1, pp. 31-35, Apr 2014.

[2] Lior Rokach, "Ensemble Methods for Classifiers," in Data Mining and Knowledge Discovery Handbook.: Springer, 2010, pp. 957-980.

[3] Lior Rokach, "Ensemble-based classifiers," Artif Intell Rev , pp. 33:1–39, 2010.

[4] David H. WOLPERT, "Stacked Generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[5] Rich Caruana, Alexandru Niculescu-Mizil , Geoff Crew, and Alex Ksikes , "Ensemble Selection from Libraries of Models," in Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 18-26.

[6] Robi Polikar, "Ensemble Based Systems in Decision Making," IEEE Circuits and Systems Magazine, vol. 6, no. 3, pp. 21-45, 2006.

[7] Kun Chang Lee and Heeryon Cho, "Performance of Ensemble Classifier for Location Prediction Task: Emphasis on Markov Blanket Perspective," International Journal of u- and e- Service, vol. 3, no. 3, September 2010.

[8] Alexey Tsymbal, Seppo Puuronen, and David Patterson, "Feature Selection for Ensembles of Simple Bayesian classifiers," Information Fusion, pp. 87-100, 2003.

[9] Hal Daumé III, a course in machine learning., 2012.

[10] Ming Ting Kia and Zijian Zheng, "A Study of AdaBoost with Naive Bayesian Classifiers: Weakness and Improvement," Computational Intelligence 19(2): (2003), vol. 19, no. 2, pp. 186-200, 2003.

[11] Hongbo Shi, Zhihai Wang, and Houkuan Huang, "Improving Classification Performance by Combining Multiple TAN Classifiers," in Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing., 2003, vol. 2639, pp. 631-634.

[12] Khalil El Hindi, "Fine tuning the naïve bayesian learning algorithm," AI Communications, vol. 27, no. 2, pp. 133-141, 2014.

[13] Amel Alhussan and Khalil Elhindi, "Fine Tuning the Tree Augmented Naïve Bayes (FTTAN) Learning Algorithm," in SAI Intelligent Systems Conference (IntelliSys), London, 2015, pp. 72 - 79.

[14] SHANG-CAI MA and HONG-BO sm, "Tree-Augmented Naive Bayes Ensembles," in Proceedings of the Third Intemational Conference on Machine Leaming and Cybernetics, Shanghai, 2004, pp. 26-29.

[15] Xiaowei Sun and Hongbo Zhou, "Experiments with Two New Boosting Algorithms," Intelligent Information Management, vol. 2, pp. 386-390, June 2010.

[16] Xiaowei Sun and Hongbo Zhou, "An Empirical Comparison of Two Boosting Algorithms on Real Data Sets based on Analysis of Scientific Materials," Advances in Intelligent and Soft, vol. 105, pp. 324-327, 2011.

[17] Xiao Wei Sun and Hong Bo Zhou , "An Empirical Evaluation of Boosting-BAN and Boosting-MultiTAN," Applied Mechanics and Materials, vol. 513-517, February 2014.

[18] Georgios Sakkis et al., "Stacking classifiers for anti-spam filtering of e-mail," in Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing (EMNLP 2001), 2001, pp. 44–50.

[19] Yushi Jing, Vladimir Pavlovic, and James M. Rehg, "Boosted Bayesian network classifiers," Machine Learning, vol. 73, no. 2, pp. 155-184, November 2008.

[20] Lior Rokach, "Ensemble Methods in Supervised Learning," in Data Mining and Knowledge Discovery Handbook., 2010, pp. 959-979.

[21] LUDMILA I. KUNCHEVA and CHRISTOPHER J. WHITAKER, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," Machine Learning, vol. 51, no. 2, pp. 181-207, May 2003.

[22] C. a. M. C. Blake, UCI Repository of Machine Learning Databases [online], 1998, Available from: http://www.ics.uci.edu/~mlearn/MLRepository.html.

[23] U. Fayyad and K. Irani, "Multi-interval discritization of continuous values attributes for classification learning," in in Proceedings of 13th International Joint Conference on Artificial Intelligence, 1993.

[24] Nir Friedman, Dan Geiger, and Moises Goldszmidt, "Bayesian Network Classifiers," vol. 29, no. 2-3, pp. 131-163, Nov/Dec 1997.

[25] JESUS CERQUIDES and RAMON LO PEZ DE MANTARAS, "TAN Classifiers Based on Decomposable Distributions," Machine Learning, vol. 59, no. 3, pp. 323-354, June 2005.

[26] Mradul Dhakar and Akhilesh Tiwari, "The Conceptual and Architectural Design of an Intelligent Intrusion Detection System," in Improving Information Security Practices through Computational Intelligence.: IGI global disseminator of knowledge, 2016, p. 26 pages.

Cloud-Based Processing for Data Science Visualization

Ahmad Ashari
Department of Computer Science
and Electronics
Faculty of Mathematic and Natural
Sciences, Universitas Gadjah Mada
Yogyakarta, Indonesia

A Min Tjoa
Institute of Software Technology
Vienna University of Technology
Vienna, Austria

Mardhani Riasetiawan
Department of Computer Science
and Electronics
Faculty of Mathematic and Natural
Sciences,
Universitas Gadjah Mada
Yogyakarta, Indonesia

***Abstract*—Data scientists need to process and visualize data science for scientific and decision purposes. The data have different size, type, real-time or batch forms, and validity. Data science visualization has a challenge in processing, management, and technique. The research works to investigate, design, and develop the cloud-based processing for data science visualization. The research uses Google Drive as file storage, Google App Engine as the processing tool, and Google Fusion for the visualization. Financial and banking data from Indonesia are used in the research to provide geolocation data, transaction flows, and bank networks information. Cloud-based processing consists of a data mapping process, data tagging, data manipulation, and data visualization. The research focus is on the data source manipulation, data preparation, storage management, data processing, and visualization. This research contributes to delivering cloud-based approach to handle data science visualization of financial-banking data networks in Indonesia.**

Keywords—component; cloud-based processing; display; data science; big data

I. Introduction

The increase in data size, type of data, data stream or batch, and the data structure is one of the issues in big data processing [1]. Computer processing has a different method and approach based on the data characteristics. It will become complicated for the data scientist to deliver the processing plan. There is also a need to understand the business process, information architecture, information system design, data structures, and delivery system designs [2]. In the term data science, we need to define the business process that should be used to deliver the information. The data science needs the word of knowledge to define business process [3]. The data that come from different sources is managed together in the store and arranged in structured or unstructured formats.

The information architecture specifies the detail of data and information [4]. The structure is used to define the data feature in the first process and the results. The information system design needs to know the information structure to describe the process and related information [5]. The interaction between information architecture and information system design requires establishing the process. The data architecture manages the data science collection by identifying the data details, in this case metadata and content [6]. The data processing method can deliver in several ways, such as integration, offline by using tools, online by using web application, and hybrid by using them in combination [7]. The processing technology approach uses real-time, batch, and stream. The method and technical approach are combined based on the purposes.

The research investigates cloud-based processing in data process for data science visualization. The research designs the cloud-based processing steps for managing the data. The technology approach used in this study are cloud-based applications such as Google Drive, Google App Engines, and Google Fusion. The study uses the financial-banking data in Indonesia provided by Open Data Indonesia [8]. The research goal is to deliver the data science visualization of intercity-network bank in Indonesia. This research has a contribution to the methods of cloud-based processing for data visualization as a best practice to deliver the data knowledge on particular issues. The paper has the following sections: Section II presents the current approach and method for data visualization. Section III delivers the step-by-step method on cloud-based processing. Section IV shows the result and discussion. The last section shows the conclusion and future direction.

II. Data Science Process and Visualization

The primary issues in data processing and display are big data and data science research, such as machine learning, data mining, semantic web, social networks, and information fusion [9]. The research is based on an investigation and discovers a new technique in data processing, data representation, pattern mining, data storage, and visualization. The combination of the algorithm and the process approach is the primary concern to the resulting information. The big data and little (small) data management can combine to support many purposes. The use of little (small) data as a sample and generated to answer a question has been used for many reasons. The little (small) data can be used for defining the sample of the big data. It will improve the quality of data and the process itself. The big data will enable in spreading data and enhance the quality of the sample and results [10].

Data management for long-term use and access, especially for big data, is an important issue in managing the data value and usage. Data processing has the capability to address the problem of long-term access and use, not only in the present but also in the future [11]. Data processing for big data can be done by using distributed data mechanism at the storage and and work management levels. The technique of distributed data storage can increase the efficiency when provided through an Internet-enabled environment [12]. The mechanism supports the system architecture for cloud-based processing. Data science needs an enormous volume of resources. In several cases, the processing needs to share with other resources to enhance capacity. The shared resources become the big data services that need protection. An authentication scheme is implemented to protect user privacy on the research conducted by Jeong and Shin [13]. Big data processing focuses on end-to-end processing of data science integration, model, and evidence [14]. The approach delivers by process mining and bridges the gap between data science and process science. The process mining use big data technologies, service based and cloud services.

The big data system architecture consists of several components, that is, data visualization, processing (include real-time, structured database, interactive analytics, and batch processing), data structure, and infrastructure. The data visualization in big data science delivers the intelligence visualization [15]. The intelligence visualization displays information and knowledge. The real-time process, analytics, and batch processing need to address speed, reliabilities, and data spread especially in processing purposes [16]. The data is classified into structured and unstructured data [17]. The infrastructure needs to address the high-performance infrastructure to support the processing needs [18]. Figure 1 presents the interaction between the components.

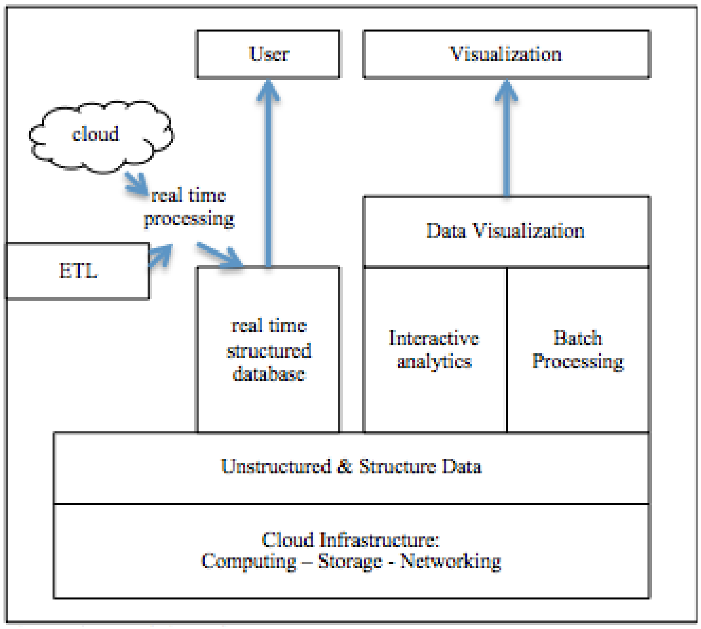


Fig. 1. Big Data Science System

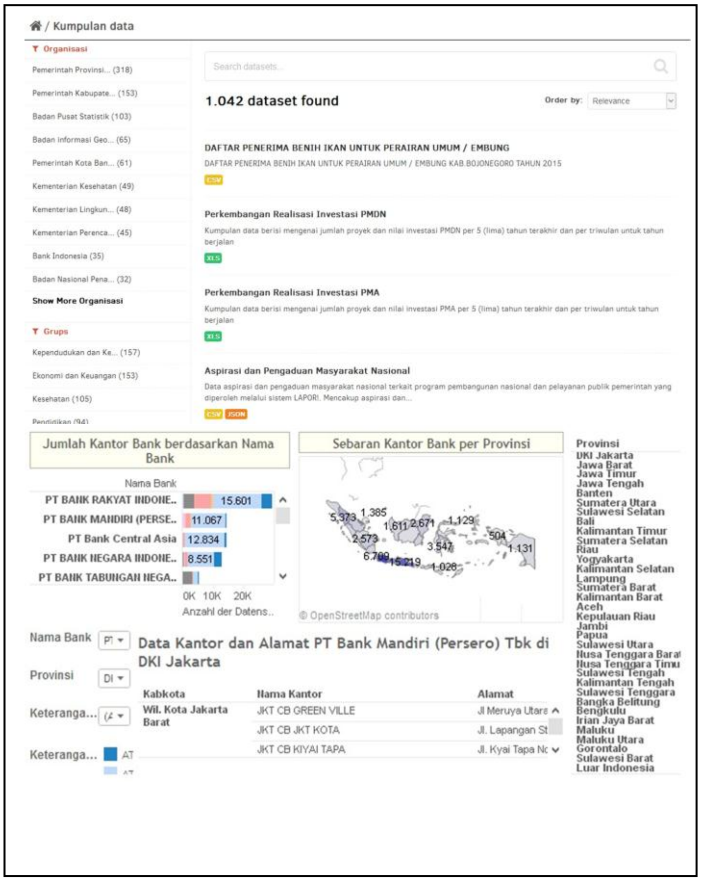
Emergency management is used in the case study and helps in overcoming the trending issue in emergency management. The visualization has been used to describe the difference between the type of record and history based on the provenance [19, 20]. The research is an organizational framework to specify the origin and design knowledge on it. Reactive Vega has presented a system architecture for graphic visualization and interaction [21]. The research constructs the data flow graph, scene graph, and interaction with streaming data. The display has been built with the help of time scale, relational, and hierarchical data.

III. Cloud -Based Processing

This section talks about the research design and works. The research was divided into several steps such as data preparation, storage management, data processing and manipulation, data integration, and data visualization.

A. Data Sources

The research uses the data from Indonesia Open Data portal. Open Data Portal (data.go.id) is a data portal built by the Indonesian Government to establish the open data movement and free data service. The open data portal itself has 1042 datasets, 31 institutions, and 18 groups of data. The research uses economic and financial data, provided by the Bank of Indonesia. There are 153 datasets consisting of economic and financial information from a broad range of regions in Indonesia. Figure 2 shows the set of the collections.



1.042 dataset found

DAFTAR PENERIMA BENIH IKAN UNTUK PERAIRAN UMUM / EMBUNG

DAFTAR PENERIMA BENIH IKAN UNTUK PERAIRAN UMUM / EMBUNG KAB.BOJONEGORO TAHUN 2015

Perkembangan Realisasi Investasi PMDN

Kumpulan data berisi mengenai jumlah proyek dan nilai investasi PMDN per 5 (lima) tahun terakhir dan per triwulan untuk tahun berjalan

Perkembangan Realisasi Investasi PMA

Kumpulan data berisi mengenai jumlah proyek dan nilai investasi PMA per 5 (lima) tahun terakhir dan per triwulan untuk tahun berjalan

Aspirasi dan Pengaduan Masyarakat Nasional

Data aspirasi dan pengaduan masyarakat nasional terkait program pembangunan nasional dan pelayanan publik pemerintah yang diperoleh melalui sistem LAPOR!. Mencakup aspirasi dan...

Data Kantor dan Alamat PT Bank Mandiri (Persero) Tbk di DKI Jakarta

Kabkota	Nama Kantor	Alamat
Wil. Kota Jakarta Barat	JKT CB GREEN VILLE	Jl Meruya Utara
	JKT CB JKT KOTA	Jl. Lapangan St
	JKT CB KIYAI TAPA	Jl. Kyai Tapa Nc

Fig. 2. Open Data Portal

The research uses the dataset from the portal that was involved in the process, that is.:

- Bank Location

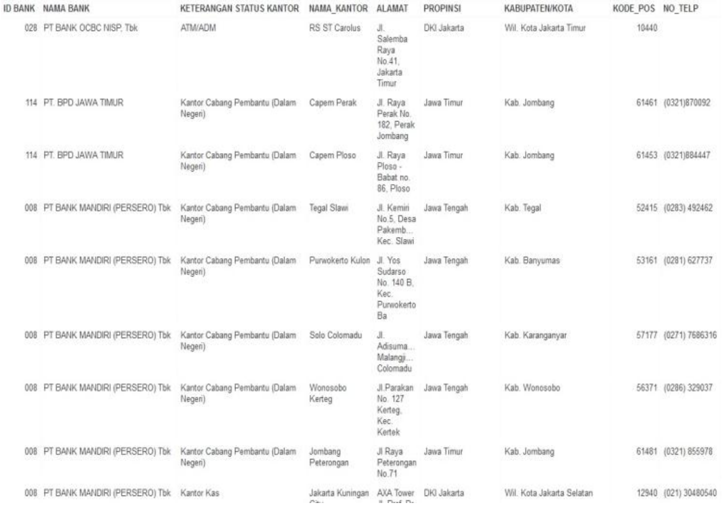
- Indonesia Bank Operation
- Transaction Volume
- Regional economic indicators

B. Data Preparation

The data preparation uses the data bank locator, operation, transaction volume, and economic indicators. The data preparation has several steps; there is data normalization, data cleansing, and data tagging. Data normalization standardizes the data. The normalization identifies the region name, the bank office, the name of the bank, and the region classification.

Data cleansing is done to minimize the data error in geotagging and relation. The data cleansing process consists of taking a data sample of at least 30 items of data. The data is transformed into the visualization prototype. The process is to figure out whether there are data items that cannot be processed based on the current data.

Data tagging has two options. Geotagging is used to give location information to the data object such as location, bank office, and transaction data. The second option or geolocation uses Google Map API facilities to attach to it. The result of this process is presented in Figure 3.



ID BANK	NAMA BANK	KETERANGAN STATUS KANTOR	NAMA_KANTOR	ALAMAT	PROPINSI	KABUPATEN/KOTA	KODE_POS	NO_TELP
028	PT BANK OCBC NISP, Tbk	ATM/ADM	RS ST Carolus	Jl. Salemba Raya No.41, Jakarta Timur	DKI Jakarta	Wil. Kota Jakarta Timur	10440	
114	PT. BPD JAWA TIMUR	Kantor Cabang Pembantu (Dalam Negeri)	Capem Perak	Jl. Raya Perak No. 182, Perak Jombang	Jawa Timur	Kab. Jombang	61461	(0321)870092
114	PT. BPD JAWA TIMUR	Kantor Cabang Pembantu (Dalam Negeri)	Capem Ploso	Jl. Raya Ploso - Babat no. 86, Ploso	Jawa Timur	Kab. Jombang	61453	(0321)884447
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Tegal Slawi	Jl. Kemiri No.5, Desa Pakemb... Kec. Slawi	Jawa Tengah	Kab. Tegal	52415	(0283) 492462
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Purwokerto Kulon	Jl. Yos Sudarso No. 140 B, Kec. Purwokerto Ba	Jawa Tengah	Kab. Banyumas	53161	(0281) 627737
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Solo Colomadu	Jl. Adisuma... Malangji... Colomadu	Jawa Tengah	Kab. Karanganyar	57177	(0271) 7686316
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Wonosobo Kerteg	Jl.Parakan No. 127 Kerteg, Kec. Kertek	Jawa Tengah	Kab. Wonosobo	56371	(0286) 329037
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Cabang Pembantu (Dalam Negeri)	Jombang Peterongan	Jl Raya Peterongan No.71	Jawa Timur	Kab. Jombang	61481	(0321) 855978
008	PT BANK MANDIRI (PERSERO) Tbk	Kantor Kas	Jakarta Kuningan	AXA Tower	DKI Jakarta	Wil. Kota Jakarta Selatan	12940	(021) 30480540

Fig. 3. Data Preparation

C. Storage Management

The cloud-based processing is stored the data and the process in the Internet facilities. The research uses Google Drive to place the data, Google App Engine to access the data stored in the intermediate storage and database engine, and Google Fusion to process the data and visualize it as presented in Figure 4.

D. Data Processing and Manipulation

The data processing and manipulation have several steps: card process, mapping, chart, and summary. The data proceed first into the card. In this process, the data are collected into the record. The data become an individual item that will be used to continue the relation and data network. The card process result is presented in Figure 5.

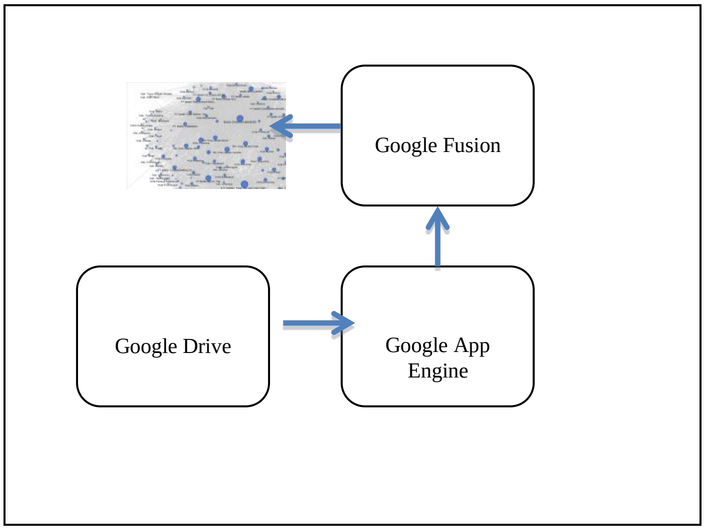
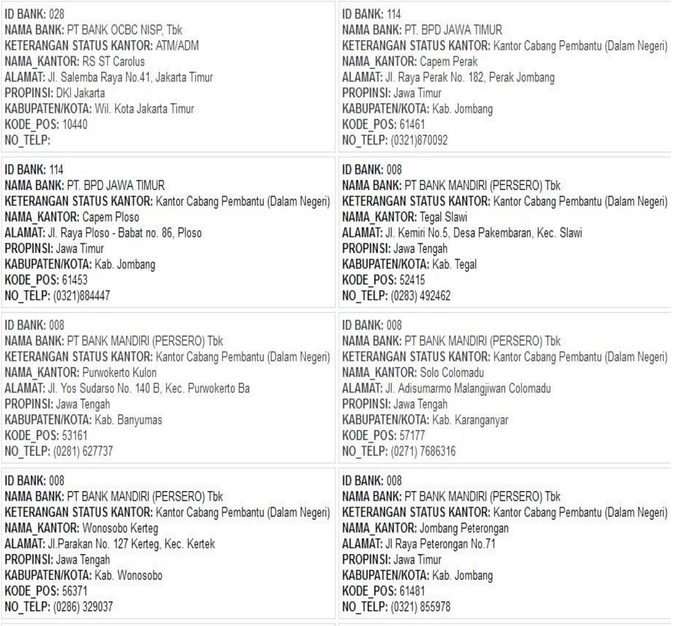


Fig. 4. Storage Management



ID BANK: 028
NAMA BANK: PT BANK OCBC NISP, Tbk
KETERANGAN STATUS KANTOR: ATM/ADM
NAMA_KANTOR: RS ST Carolus
ALAMAT: Jl. Salemba Raya No.41, Jakarta Timur
PROPINSI: DKI Jakarta
KABUPATEN/KOTA: Wil. Kota Jakarta Timur
KODE_POS: 10440
NO_TELP:

ID BANK: 114
NAMA BANK: PT. BPD JAWA TIMUR
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Capem Perak
ALAMAT: Jl. Raya Perak No. 182, Perak Jombang
PROPINSI: Jawa Timur
KABUPATEN/KOTA: Kab. Jombang
KODE_POS: 61461
NO_TELP: (0321)870092

ID BANK: 114
NAMA BANK: PT. BPD JAWA TIMUR
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Capem Ploso
ALAMAT: Jl. Raya Ploso - Babat no. 86, Ploso
PROPINSI: Jawa Timur
KABUPATEN/KOTA: Kab. Jombang
KODE_POS: 61453
NO_TELP: (0321)884447

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Tegal Slawi
ALAMAT: Jl. Kemiri No.5, Desa Pakembaran, Kec. Slawi
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Tegal
KODE_POS: 52415
NO_TELP: (0283) 492462

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Purwokerto Kulon
ALAMAT: Jl. Yos Sudarso No. 140 B, Kec. Purwokerto Ba
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Banyumas
KODE_POS: 53161
NO_TELP: (0281) 627737

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Solo Colomadu
ALAMAT: Jl. Adisumarmo Malangjiwan Colomadu
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Karanganyar
KODE_POS: 57177
NO_TELP: (0271) 7686316

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Wonosobo Kerteg
ALAMAT: Jl.Parakan No. 127 Kerteg, Kec. Kertek
PROPINSI: Jawa Tengah
KABUPATEN/KOTA: Kab. Wonosobo
KODE_POS: 56371
NO_TELP: (0286) 329037

ID BANK: 008
NAMA BANK: PT BANK MANDIRI (PERSERO) Tbk
KETERANGAN STATUS KANTOR: Kantor Cabang Pembantu (Dalam Negeri)
NAMA_KANTOR: Jombang Peterongan
ALAMAT: Jl Raya Peterongan No.71
PROPINSI: Jawa Timur
KABUPATEN/KOTA: Kab. Jombang
KODE_POS: 61481
NO_TELP: (0321) 855978

Fig. 5. Card Process

The data is also used in the mapping process. The data uses the location parameter by rendering and process to have geolocation based on Google Map. The mapping process resulted in a card that had the information location. It also processes the transaction data. The next process is a chart and summary. The process is used to create a relation between datasets to map the network process. The summarizing will give weight to every data and the bank location to visualize in the representation burden.

E. Visualization

The visualization process works to display all the information that resulted from previous steps appropriately.. The visualization process itself has a particular format. The process identified the bank institution and location (city) as primary nodes, and transaction and other data as weight indicator for primary nodes. It needs not only for visualization and give the value of nodes. Table I shows the visualization process. The visualization process is rendered from the dataset and displayed in the HTML format.

TABLE I. VISUALIZATION PARAMETER

Primary	***Weight***
Location Bank Name	Transaction Volume Indicator

IV. RESULT

The research has resulted in a working visualization prototype for displaying the bank, location, and transaction weight based on the cloud-based processing. The display shows the network maps chart as shown in Figure 6.

The visualization result demonstrates the bank, location, and the transaction. The nodes have a different size based on the transaction weight on it. The visualization can be dynamic and comes out with the other data.

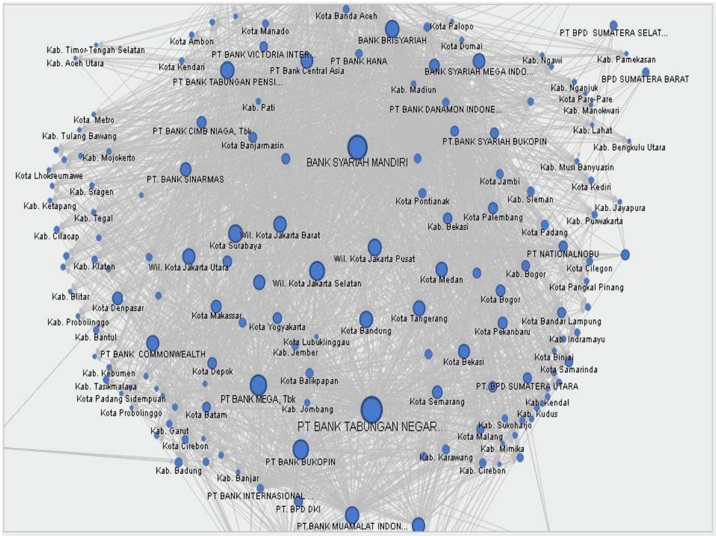


Fig. 6. Visualization Result

The visualization can display the network between the banks that operate in several cities, as shown in Figure 7. The relation between cities and banks is illustrated in Figure 8.

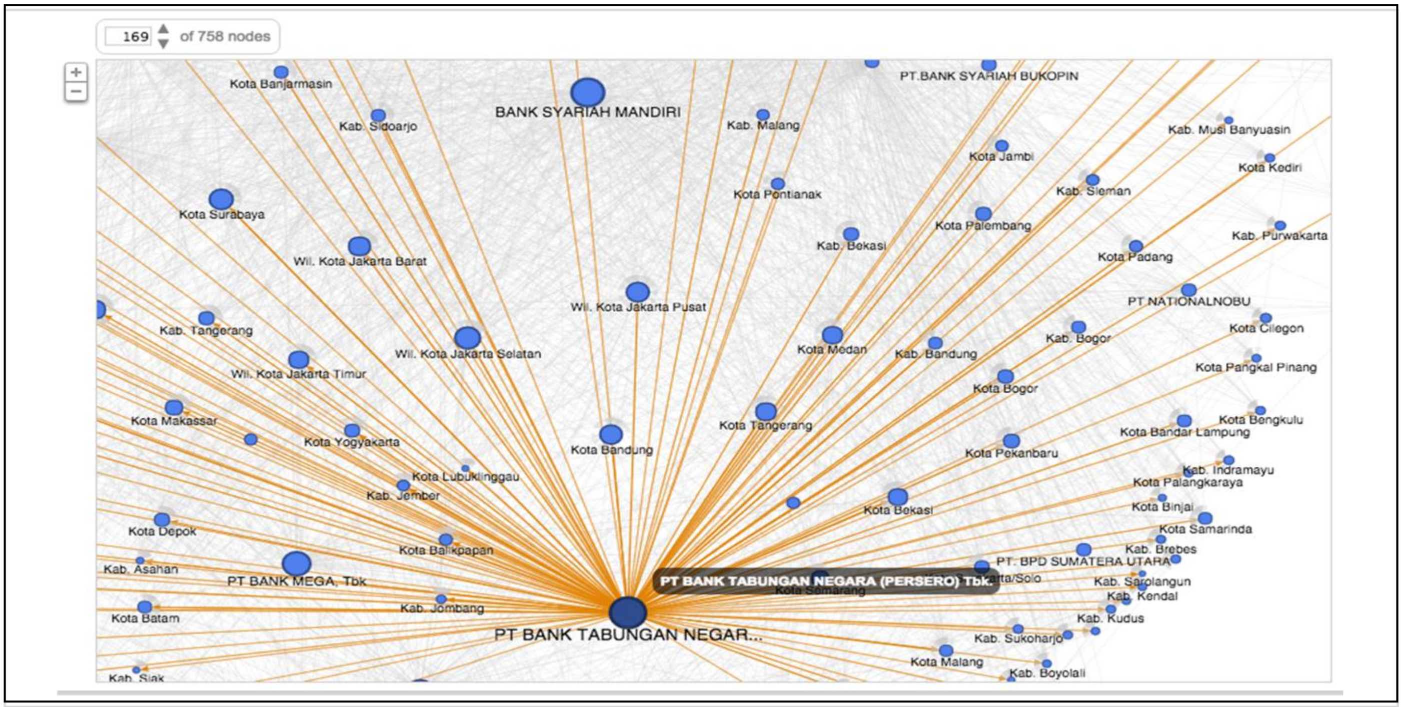


Fig. 7. Dynamic Visualization based on The Bank

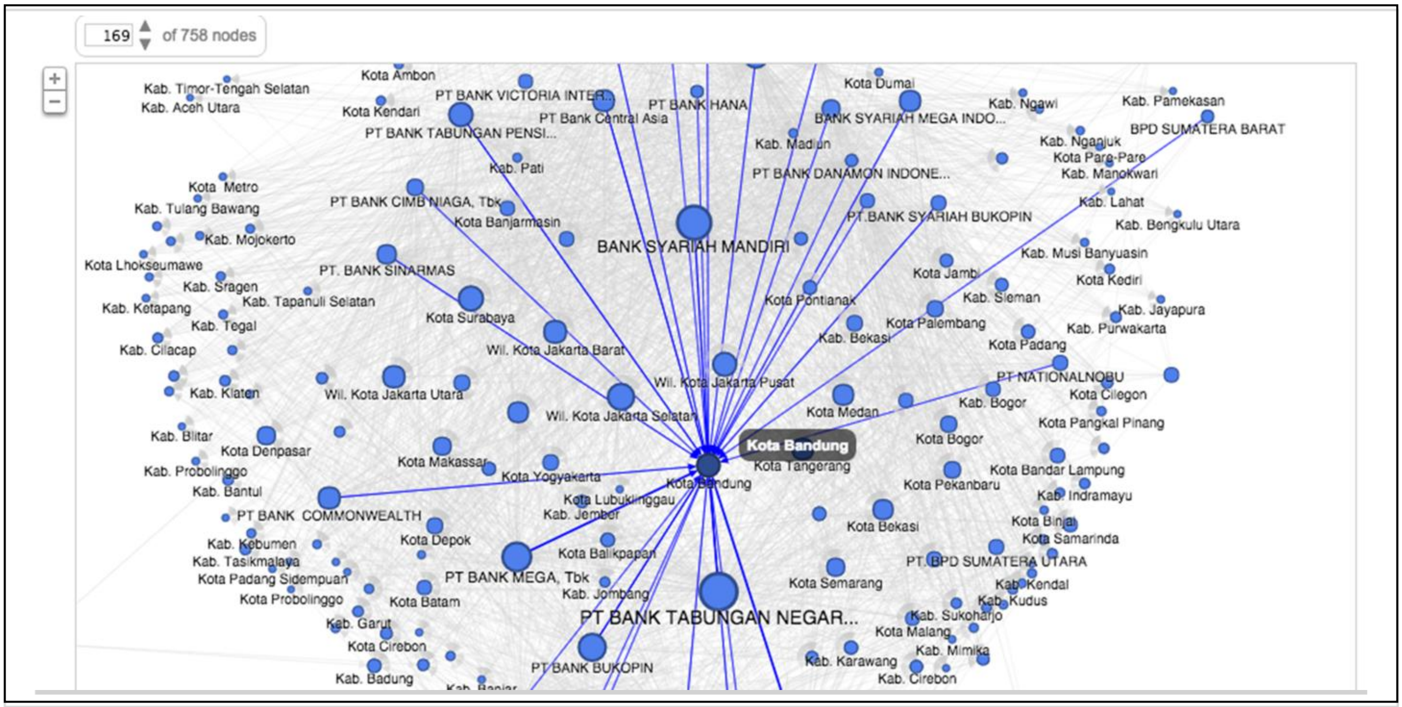


Fig. 8. Dynamic Visualization based on the city

V. Conclusion and Discussion

The research has shown the best practice of using cloud-based approach to process the data science, which is big data, with several steps. The conclusion of this research is that the cloud-based approach utilized for data science purposes, in this case, uses Google application. The research works with an open data sample, and the visualization is presented in http://makeiswork.com/2015/12/23/show-case/ as a working prototype. The research work has a contribution to the process of data science into the visualization that uses cloud-based approach. The process consists of data preparation, storage management, data processing and manipulation, and display. In this, every process needs a unique approach to ensure the quality of data. The data process is unique and depends on the data characteristic itself. The process involving more datasets will need more processing. The work on visualization depends on the process.

The research on data process and display has a challenge in the multiple datasets involved. The process even uses a framework tool but still needs to have a well-designed approach and methods. The cloud-based approach addresses the process on the Internet. The approach needs to address the multiple sources handled in the cloud-based process. The work on the approach will be the key for many organizations in business decision-making, business analysis, and intelligence, or scientific analysis.

Acknowledgment

The research was supported by ASEA UNINET, Vienna University, and Vienna University of Technology. The research uses sample data from Open Data Indonesia (data.go.id), and places the visualization in www.makeiswork.com which in turn is supported by DALA-intelligent technology initiatives.

References

[1] A.A. Al-Habsi, D.K. Kang, M.J. kim, “Enhancing Dataset Processing in Hadoop Yarn Performance for Big Data Applications,” Lecture Notes in Electrical Engineering, vol. 354, pp. 9-15, 2016.

[2] G. Quicmayr, “The Boeing Information Services 2010 Study,” Lecturing Materials, Institut fur Distributed and Multimedia Systems, Universitat Wien, 2015.

[3] F. Rahimi, C. Moller, L. Hvam, “Business Process Management and IT Management: The missing Integration, “ International Journal of Information Management, vol. 36, pp. 142-154, 2016.

[4] C. Maican, R. Kixandroiu, “A System Architecture Based on Open Source Enterprise Content Management System for Supporting Educational Institutions,” International Journal of Information Management, vol. 36, pp. 2017-214, 2016.

[5] R. Dijkman, I. Vanderfeesten, H.A Reijers, “Business Process Architecture: Overview, Comparison, and Framework,” Enterprise Information System, vol. 10, pp. 129-158, 2016.

[6] A.A. Neznanov, A.A. Parinov, “Distributed Architecture of Data Analysis System Based on Formal Concept Analysis Approach,” Studies in Computational Intelligence, vol. 616, pp. 265-271.

[7] C. Hu, X. Cheng, Z. Liu, “A Virtual Dataspaces Model for Large-scale Materials Scientific Data Access,” Future Generation Computer Systems, vol. 54, pp. 456-468, 2016.

[8] Open Government Indonesia, available at www.data.go.id

[9] G. Bello-Orgaz, J.J Jung, D. Camacho, “Social Big Data: Recent Achievements and New Challenges,” Information Fusion, vol. 28, pp. 45-59, 2016.

[10] R. Kitchin, T.P. Lauriault, “Small Data in the era of big data,” Geojournal, vol. 80, pp. 463-475, 2015.

[11] M. Riasetiawan, AK. Mahmood, “Managing and Preserving Large Data Volume in Data Grid Environment,” 2010 International Conference on Information Retrieval and Knowledge Management, Shah Alam Malaysia, pp. 91-96, 2010.

[12] S.T. Park, Y.R. Kim, S.P. Jeong, C.I. Hong, T.G. Kang, “A Case Study on Effective Technique of Distributed Data Storage for Big Data

Processing in The Wireless Internet Environment," Wireless Pers Communication, vol. 86, pp. 239-253, 2016.

[13] Y.S. Jeong, S.S. Shin, "An Efficient Authentication Scheme to Protect User Privacy in Seamless Big Data Services," Wireless Pers Communication, vol. 86, pp. 7-19, 2016.

[14] W.v.d. Aalst, E. Damiani, "Processes Meet Big Data: Connecting Data Science with Process Science," IEEE Transaction on Service Computing, vol. 8, pp. 810-819, 2015.

[15] E.A Mohammed, C. Naugler, B.H. Far, "Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics," Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools, pp. 577-602, 2015.

[16] Cloudera, available at www.cloudera.com.

[17] B. Funk, P. Niemeyer, J.M. Gomez, "Information Technology in Environmental Engineering: Selected Contributions to the Sixth International Conference on Information Technologies in Environmental Engineering (ITEE 2013)," 2013.

[18] F. Teng. Management Des Donnees Et Ordonnancement Des Taches Sur Architectures Distributes. Thesis, Ecole Centrale Paris, 2011.

[19] F. Dusse, P:S. Junior, A.T Alves, R. Novais, V. Vieira, M. Mendoca, "Information Visualization for Emergency Management: A Systematic Mapping Study," Expert Systems with Application, vol. 45, pp.424-437, March 2016.

[20] E.D Ragan, A. Endert, J. Sanyal, J. Chen, "Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework for Provenance Types and Purposes," IEEE Transaction on Visualization and Computer Graphic, vol. 22, pp. 31-40, January 2016.

[21] A. Satyanarayan, R. Russel, J. Hoffswell, J. Heer, "Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization," IEEE Transaction on Visualization and Computer Graphic, vol. 22, pp. 659-668, January 2016.

Comparative Study of Robust Control Strategies for a Dfig-Based Wind Turbine

Mohamed BENKAHLA, Rachid TALEB, Zinelaabidine BOUDJEMA
Electrical Engineering Department, Hassiba Benbouali University
Laboratoire Génie Electrique et Energies Renouvelables (LGEER)
Chlef, Algeria

***Abstract*—Conventional vector control configurations which use a proportional-integral (PI) regulator for the powers DFIGs driven have some drawbacks such as parameter tuning difficulties, mediocre dynamic performances and reduced robustness. So, based on analysis of the DFIG model, fed by a direct AC-AC converter, two nonlinear algorithms: sliding mode and adaptive fuzzy logic is used to control independently active and reactive powers provided by the stator side of the DFIG to the grid. Their respective performances are compared with the conventional PI controller regarding reference tracking, response to sudden speed variation and robustness against machine parameters variations.**

Keywords—Wind turbine (WT); doubly fed induction generator (DFIG); sliding mode controller (SMC); PI controller; adaptive fuzzy logic controller (AFLC)

I. Introduction

Wind energy is the most promising renewable source of Electrical power generation for the future. Many countries promote the wind power technology through various national Programs and market incentives. Wind energy technology has evolved rapidly over the past three decades with increasing rotor diameters and the use of sophisticated power electronics to allow operation at variable speed [1].

Doubly fed induction generator (DFIG) is one of the most popular variable speed wind turbines in use nowadays. It is usually supplied by a voltage source inverter. However, currently the three phase matrix converters have received considerable attention because they may become a good alternative to voltage-source inverter Pulse-Width-Modulation (PWM) topology. The matrix converter provides bi-directional power flow, nearly sinusoidal input/output waveforms, and a controllable input power factor [2]. Furthermore, the matrix converter allows a compact design due to the lack of dc-link capacitors for energy storage. Consequently, in this work, a three-phase matrix converter is used to drive the DFIG.

Many research works have been presented with different control schemes of DFIG. These control diagrams are usually based on vector control notion with conventional PI controllers as proposed in [3, 4]. The similar conventional controllers are also used to realize control techniques of DFIG when grid faults appear like unbalanced voltages [5, 6] and voltage dips [7]. It has also been shown in [8, 9] that glimmer problems could be resolved with suitable control strategies. Many of these works prove that stator reactive power control can be an adapted solution to these diverse problems.

In recent years, the sliding mode control (SMC) Methodology has been widely used for robust control of nonlinear systems. Sliding mode control, based on the theory of variable structure systems (VSS), has attracted a lot of research on control systems for the last two decades. It achieves robust control by adding a discontinuous control signal across the sliding surface, satisfying the sliding condition. Nevertheless, this type of control has an essential disadvantage, which is the chattering phenomenon caused by the discontinuous control action. In order to overcome these difficulties, several modifications to the original sliding control law have been proposed, the most popular being the boundary layer approach [10].

Fuzzy logic is a technology based on engineering experience and observations. In fuzzy logic, an exact mathematical model is not necessary because linguistic variables are used to define system behavior rapidly.

This work is organized as follows. We briefly review the modelling of the device studied in Section II. In Section III we present the field oriented control of the DFIG. Section IV provides the control of stator active and reactive powers of the DFIG by using three different controllers: PI, SMC and AFLC. In Section V, The three controllers are compared regarding reference power tracking, sensitivity to perturbations and robustness against machine parameters variations. Finally, a summary of the results is presented in the Conclusion.

II. System Modeling

A. Wind turbine model

For a horizontal axis wind turbine, the mechanical power captured from the wind is given by [11]:

$$P_t = \frac{1}{2} C_P(\lambda,\beta)\pi R^2 \rho v^3 \tag{1}$$

where, R is the radius of the turbine (m), ρ is the air density (kg/m^3), v is the wind speed (m/s), and C_P is the power coefficient which is a function of both tip speed ratio λ, and blade pitch angle β (deg). In this work, the C_P equation is approximated using a nonlinear function according to [12].

$$C_P = (0.5 - 0.167)(\beta - 2)\sin\left[\frac{\pi(\lambda + 0.1)}{18.5 - 0.3(\beta - 2)}\right] - 0.0018(\lambda - 3)(\beta - 2) \tag{2}$$

The tip speed ratio is given by:

$$\lambda = \frac{\Omega_t R}{v} \tag{3}$$

where Ω_t is the angular velocity of Wind Turbine.

B. The matrix converter model

The matrix converter performs the power conversion directly from AC to AC without any intermediate dc link. It is very simple in structure and has powerful controllability.

The converter consists of a matrix of bi-directional switches linking two independent three-phase systems. Each output line is linked to each input line via a bi-directional switch. Fig. 1 shows the basic diagram of a matrix converter.

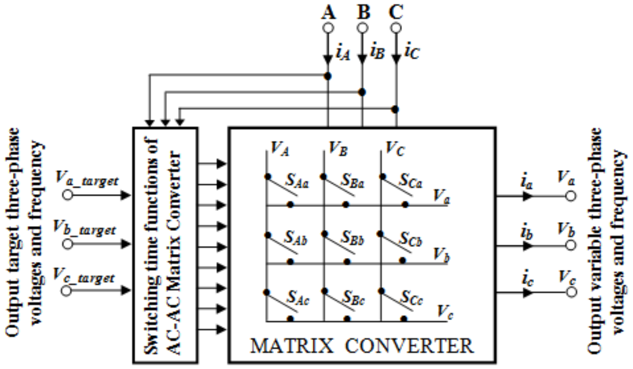


Fig. 1. Schematic representation of the matrix converter

The switching function of a switch S_{mn} in Fig. 1 is given by:

$$S_{mn} = \begin{cases} 1 & S_{mn} \ \text{closed} \\ 0 & S_{mn} \ \text{open} \end{cases} \quad m \in \{A,B,C\}, n \in \{a,b,c\} \tag{4}$$

The mathematical expression that represents the operation of the matrix converter in figure 1 can be written as:

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} S_{Aa} & S_{Ab} & S_{Ac} \\ S_{Ba} & S_{Bb} & S_{Bc} \\ S_{Ca} & S_{Cb} & S_{Cc} \end{bmatrix} \cdot \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} i_A \\ i_B \\ i_C \end{bmatrix} = \begin{bmatrix} S_{Aa} & S_{Ba} & S_{Ca} \\ S_{Ab} & S_{Bb} & S_{Cb} \\ S_{Ac} & S_{Bc} & S_{Cc} \end{bmatrix}^T \cdot \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \tag{6}$$

To determine the behavior of the matrix converter at output frequencies well below the switching frequency, a modulation duty cycle can be defined for each switch.

The input/output relationships of voltages and currents are related to the states of the nine switches and can be expressed as follows:

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} k_{Aa} & k_{Ab} & k_{Ac} \\ k_{Ba} & k_{Bb} & k_{Bc} \\ k_{Ca} & k_{Cb} & k_{Cc} \end{bmatrix} \cdot \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix} \tag{7}$$

$$\begin{bmatrix} i_A \\ i_B \\ i_C \end{bmatrix} = \begin{bmatrix} k_{Aa} & k_{Ba} & k_{Ca} \\ k_{Ab} & k_{Bb} & k_{Cb} \\ k_{Ac} & k_{Bc} & k_{Cc} \end{bmatrix}^T \cdot \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \tag{8}$$

with: $0 \le k_{mn} \le 1,$ $m = A, B, C, n = a, b, c$ (10)

The variables k_{mn} are the duty cycles of the nine switches S_{mn} and can be represented by the duty-cycle matrix k. In order to prevent a short circuit on the input side and ensure uninterrupted load current flow, these duty cycles must satisfy the three following constraint conditions:

$$k_{Aa} + k_{Ab} + k_{Ac} = 1 \tag{10}$$

$$k_{Ba} + k_{Bb} + k_{Bc} = 1 \tag{11}$$

$$k_{Ca} + k_{Cb} + k_{Cc} = 1 \tag{12}$$

The high-frequency synthesis technique introduced by Venturini (1980) and Alesina and Venturini (1988), allows a control of the S_{mn} switches so that the low frequency parts of the synthesized output voltages (V_a, V_b and V_c) and the input currents (i_A, i_B and i_C) are purely sinusoidal with the prescribed values of the output frequency, the input frequency, the displacement factor and the input amplitude. The output voltage is given by :

$$\begin{bmatrix} V_a \\ V_b \\ V_c \end{bmatrix} = \begin{bmatrix} 1+2\delta\cos\alpha & 1+2\delta\cos(\alpha-\frac{2\pi}{3}) & 1+2\delta\cos(\alpha-\frac{4\pi}{3}) \\ 1+2\delta\cos(\alpha-\frac{4\pi}{3}) & 1+2\delta\cos\alpha & 1+2\delta\cos(\alpha-\frac{2\pi}{3}) \\ 1+2\delta\cos(\alpha-\frac{2\pi}{3}) & 1+2\delta\cos(\alpha-\frac{4\pi}{3}) & 1+2\delta\cos\alpha \end{bmatrix} \cdot \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix} \tag{13}$$

where $\begin{cases} \alpha = \omega_m + \theta \\ \omega_m = \omega_{output} - \omega_{input} \end{cases}$

The running matrix converter with Venturini algorithm generates at the output a three-phases sinusoidal voltages system having in that order pulsation ω_m, a phase angle θ and amplitude $\delta.V_s$ ($0 < \delta < 0.866$ with modulation of the neural) [13].

C. Doubly fed induction generator modeling

For a doubly fed induction machine, the Concordia and Park transformation's application to the traditional *a*, *b*, *c* model allows to write a dynamic model in a *d-q* reference frame as follows:

$$\begin{cases} V_{ds} = R_s I_{ds} + \frac{d}{dt}\psi_{ds} - \omega_s \psi_{qs} \\ V_{qs} = R_s I_{qs} + \frac{d}{dt}\psi_{qs} + \omega_s \psi_{ds} \\ V_{dr} = R_r I_{dr} + \frac{d}{dt}\psi_{dr} - \omega_r \psi_{qr} \\ V_{qr} = R_r I_{qr} + \frac{d}{dt}\psi_{qr} + \omega_r \psi_{dr} \end{cases}, \quad \begin{cases} \psi_{ds} = L_s I_{ds} + M I_{dr} \\ \psi_{qs} = L_s I_{qs} + M I_{qr} \\ \psi_{dr} = L_r I_{dr} + M I_{ds} \\ \psi_{qr} = L_r I_{qr} + M I_{qs} \end{cases} \tag{14}$$

The stator and rotor angular velocities are linked by the following relation : $\omega_s = \omega + \omega_r$. This electrical model is completed by the mechanical equation:

$$C_{em} = C_r + J\frac{d\Omega}{dt} + f\Omega \tag{15}$$

where the electromagnetic torque C_{em} can be written as a function of stator flux and rotor currents:

$$C_{em} = p\frac{M}{L_s}(\psi_{qs}I_{dr} - \psi_{ds}I_{qr}) \tag{16}$$

III. Field Oriented Control of the DFIG

In order to easily control the production of electricity by the wind turbine, we will carry out an independent control of active and reactive powers by orientation of the stator flux.

Since the stator is directly connected to the grid and the stator flux can be considered constant, and if the voltage dropped in the stator resistance has been neglected, the voltage equations, flux equations, currents equations and stator active and reactive powers equations can be simplified in study state as: If the stator flux is linked to the d-axis of the frame we have:

$$\psi_{ds} = \psi_s \text{ and } \psi_{qs} = 0 \tag{17}$$

And the electromagnetic torque can then be expressed as follows:

$$C_{em} = -p\frac{M}{L_s}I_{qr}\psi_{ds} \tag{18}$$

by substituting Eq. 18 in Eq. 15, the following rotor flux equations are obtained:

$$\begin{cases}\psi_s = L_s I_{ds} + MI_{dr}\\ 0 = L_s I_{qs} + MI_{qr}\end{cases} \tag{19}$$

In addition, the stator voltage equations are reduced to:

$$\begin{cases}V_{ds} = 0\\ V_{qs} = \omega_s\psi_s\end{cases} \tag{20}$$

By supposing that the electrical supply network is stable, having for simple voltage V_s, that led to a stator flux ψ_s constant. This consideration associated with Eq. 19 shows that the electromagnetic torque only depends on the *q*-axis rotor current component. Using Eq. 20, a relation between the stator and rotor currents can be established:

$$\begin{cases}I_{ds} = -\frac{M}{L_s}I_{dr} + \frac{\psi_s}{L_s}\\ I_{qs} = -\frac{M}{L_s}I_{qr}\end{cases} \tag{21}$$

The stator active and reactive powers are written:

$$\begin{cases}P_s = V_{ds}I_{ds} + V_{qs}I_{qs}\\ Q_s = V_{qs}I_{ds} - V_{ds}I_{qs}\end{cases} \tag{22}$$

By using Eqs. 14, 15, 12 and 22, the statoric active and reactive power, the rotoric fluxes and voltages can be written versus rotoric currents as:

$$\begin{cases}P_s = -\frac{\omega_s\psi_s M}{L_s}I_{qr}\\ Q_s = -\frac{\omega_s\psi_s M}{L_s}I_{dr} + \frac{\omega_s\psi_s{}^2}{L_s}\end{cases} \tag{23}$$

$$\begin{cases}\psi_{dr} = (L_r\text{-}\frac{M^2}{L_s})I_{dr} + \frac{M\psi_s}{L_s}\\ \psi_{qr} = (L_r\text{-}\frac{M^2}{L_s})I_{qr}\end{cases} \tag{24}$$

$$\begin{cases}V_{dr} = R_r I_{dr} + (L_r\text{ - }\frac{M^2}{L_s})\frac{dI_{dr}}{dt} - g\omega_s(L_r\text{ - }\frac{M^2}{L_s})I_{qr}\\ V_{qr} = R_r I_{qr} + (L_r\text{ - }\frac{M^2}{L_s})\frac{dI_{qr}}{dt} + g\omega_s(L_r\text{ - }\frac{M^2}{L_s})I_{dr} + g\omega_s\frac{M\psi_s}{L_s}\end{cases} \tag{25}$$

In steady state, the second derivative terms of the two equations in 26 are nil. We can thus write:

$$\begin{cases}V_{dr} = R_r I_{dr} - g\omega_s(L_r\text{-}\frac{M^2}{L_s})I_{qr}\\ V_{qr} = R_r I_{qr} + g\omega_s(L_r\text{-}\frac{M^2}{L_s})I_{dr} + g\omega_s\frac{M\psi_s}{L_s}\end{cases} \tag{26}$$

The third term, which constitutes cross-coupling terms, can be neglected because of their small influence. These terms can be compensated by an adequate synthesis of the regulators in the control loops.

IV. Controllers Synthesis

In this section, we have chosen to compare the performances of the DFIG with three different controllers: PI, SMC and AFLC.

Based on relations (23) and (25), the control system can be designed as shown in Fig. 2. The blocks R_1, R_2, R_3 and R_4 represent respectively the stator powers and the rotor currents regulators.

A. PI controller synthesis

This controller is simple to elaborate. Figure 2 shows the block diagram of the system implemented with this controller. The terms k_p and k_i represent respectively the proportional and integral gains. The quotient *B/A* represents the transfer function to be controlled, where *A* and *B* are presently defined as follows:

$$A = L_s R_r + s.L_s\left(L_r - \frac{M^2}{L_s}\right) \text{ and } B = \omega_s\psi_s M \tag{27}$$

The regulator terms are calculated with a pole-compensation method. The time response of the controlled system will be fixed at 10 *ms*. This value is sufficient for our application and a lower value might involve transients with important overshoots. Fig. 3 shows the system with PI controller, the calculated terms are:

$$k_p = \frac{1}{1\times10^{-3}}\frac{L_s(L_r - \frac{M^2}{L_s})}{M\omega_s\psi_s}\ ,\ k_i = \frac{1}{1\times10^{-3}}\frac{L_s R_r}{M\omega_s\psi_s} \tag{28}$$

It is important to specify that the pole compensation is not the only method to calculate a *PI* regulator but it is simple to elaborate with a first-order transfer-function and it is sufficient in our case.

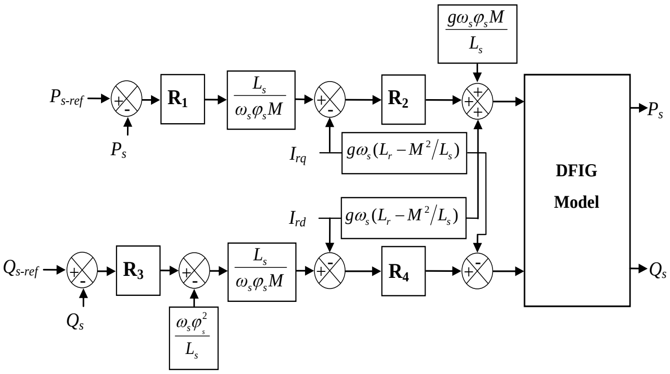

$\frac{g\omega_s\varphi_s M}{L_s}$
$P_{s\text{-}ref}$
P_s
R1
$\frac{L_s}{\omega_s\varphi_s M}$
R2
I_{rq}
$g\omega_s(L_r - M^2/L_s)$
I_{rd}
$g\omega_s(L_r - M^2/L_s)$
DFIG Model
P_s
$Q_{s\text{-}ref}$
Q_s
R3
$\frac{L_s}{\omega_s\varphi_s M}$
R4
$\frac{\omega_s\varphi_s^2}{L_s}$
Q_s

Fig. 2. Power control of the DFIG

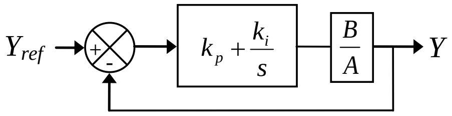


Fig. 3. System with PI controller

B. Sliding mode controller (SMC)

The sliding mode technique is developed from variable structure control to solve the disadvantages of other designs of nonlinear control systems. The sliding mode is a technique to adjust feedback by previously defining a surface. The system which is controlled will be forced to that surface, then the behavior of the system slides to the desired equilibrium point.

The main feature of this control is that it only needs to drive the error to a "switching surface". In this study, the errors between the measured and references stator powers have been chosen as sliding mode surfaces, so the following expression can be written [14].

$$\begin{cases} S_C(P) = P_{S-ref} - P_{S-mes} \\ S_C(Q) = Q_{S-ref} - Q_{S-mes} \end{cases} \tag{29}$$

the first derivative of Eq. (29), gives:

$$\begin{cases} \dot{S}_C(P) = \dot{P}_{S-ref} - \dot{P}_{S-mes} \\ \dot{S}_C(Q) = \dot{Q}_{S-ref} - \dot{Q}_{S-mes} \end{cases} \tag{30}$$

we replace (23) in (30), we obtain:

$$\begin{cases} \dot{S}_C(P) = \dot{P}_{S-ref} + \frac{\omega_S\varphi_S M}{L_S}\dot{I}_{qr} \\ \dot{S}_C(Q) = \dot{Q}_{S-ref} + \frac{\omega_S\varphi_S M}{L_S}\dot{I}_{dr} - \frac{\omega_S\varphi_S^2}{L_S} \end{cases} \tag{31}$$

We pull derivatives of currents I_{dr} and I_{qr} of the Eq. (25) replaced in the Eq. (31), and during the sliding mode and in steady state, $S_c(P) = S_c(Q) = 0$ and $\dot{S}_c(P) = \dot{S}_c(Q) = 0$, we find equivalent command I_{dr-eq} and I_{qr-eq} :

$$\begin{cases} I_{dr-eq} = Q_{S-ref}\frac{L_S\sigma L_r}{V_S M R_r} + \left(\frac{1}{R_r}v_{dr} - \frac{L_S\sigma\omega_r}{V_S M R_r}I_{qr}\right) \\ I_{qr-eq} = P_{S-ref}\frac{L_S\sigma L_r}{V_S M R_r} + \left(\frac{1}{R_r}v_{qr} - \frac{L_r\sigma\omega_r}{R_r}I_{dr} - gM\frac{V_S}{R_r L_S}\right) \end{cases} \tag{32}$$

$$\sigma = 1 - \frac{M^2}{L_S L_r} \tag{33}$$

To obtain good executions, dynamics and commutations around surfaces, the control vector is written as follows:

$$\begin{cases} I_{dr-n} = K_{dr}\cdot sign(S_C(Q_S)) \\ I_{qr-n} = K_{qr}\cdot sign(S_C(P_S)) \end{cases} \tag{34}$$

The sliding mode will exist only if the following condition is verified:

$$S\cdot\dot{S} < 0 \tag{35}$$

C. Adaptive fuzzy logic controller (AFLC)

Gain scheduling means a technique where PI controller parameters are tuned during control of the system in a predefined way [15], it enlarges the operation area of linear controller (PI) to perform well also with a nonlinear system. The diagram of this technique is illustrated in Fig. 4.

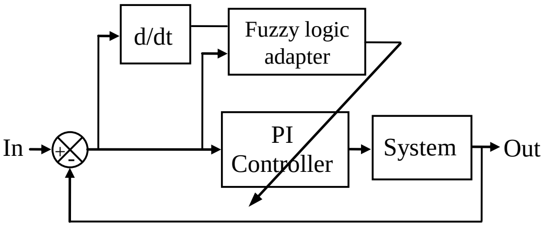


Fig. 4. Principle of adaptation of PI by fuzzy logic

The fuzzy inference mechanism adjusts the PI parameters and generates new parameters during process control, so that the fuzzy logic adapts the PI parameters to operating conditions based on the error and its first time difference. The parameters of the PI controller used in the direct chain K_p and K_i are normalized into the range between zero and one by using the following linear transformations [16]:

$$\begin{cases} K'_p = (K_p - K_{p-min})/(K_{p-max} - K_{p-min}) \\ K'_i = (K_i - K_{i-min})/(K_{i-max} - K_{i-min}) \end{cases} \tag{36}$$

The inputs of the fuzzy adapter are: The error *e* and the derivative of error d*e*, the outputs are: the normalized value of the proportional action K'_p and the normalized value of the integral action K'_i.

The problem of selecting the suitable fuzzy controller rules remain relying on expert knowledge and try and error tuning methods. The fuzzy subsets of the input variables are defined as follows: *NB* Negative Big, *NM* Negative Middle, *NS* Negative Small, *EZ* Equal Zero, *PS* Positive Small, *PM* Positive Middle, *PB* Positive Big. The fuzzy subsets of the output variables are defined as follows: *B* Big, *S* Small.

The membership functions for *e* inputs and d*e* are defined in the range [-1, 1] (Fig. 5). And memberships functions for the outputs are defined in the interval [0, 1] (Fig. 6).

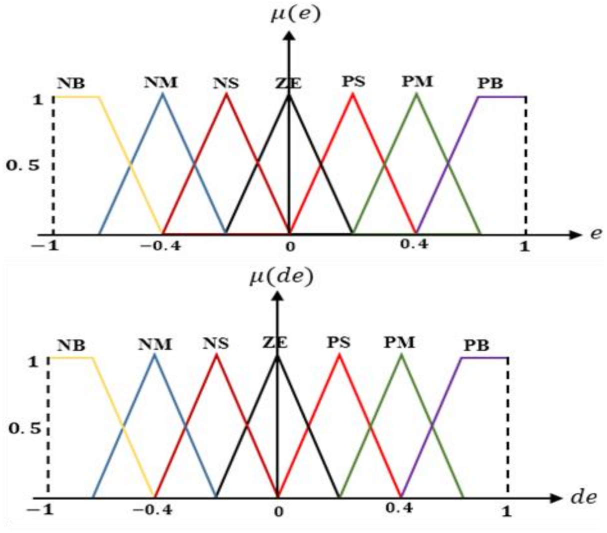


Fig. 5. Function of membership for *e* and *de*

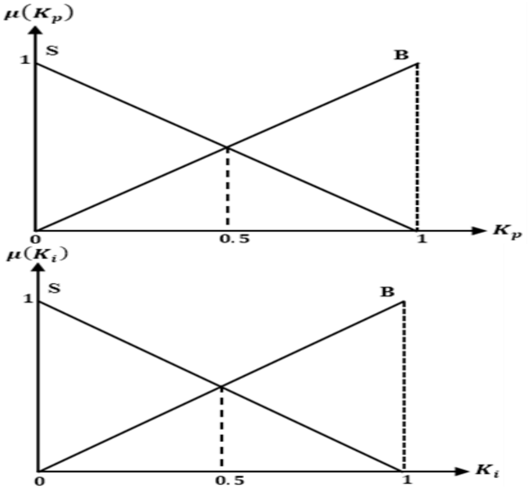


Fig. 6. Function of membership for *Kp* and *Ki*

The fuzzy rules may be extracted from operator's expertise or based on the step response of the process [15]. The tuning rules for K'_p and K'_i are given in Tables 1 and 2 respectively.

TABLE I. BASIS OF RULES FOR THE OUTPUT K_P

e / de	*NB*	*NM*	*NS*	*ZE*	*PS*	*PM*	*PB*
NB	B	B	B	B	B	B	B
NM	B	B	B	B	B	B	S
NS	S	S	B	B	B	S	S
ZE	S	S	S	B	S	S	S
PS	S	S	B	B	B	S	S
PM	S	B	B	B	B	B	S
PB	B	B	B	B	B	B	B

TABLE II. BASIS OF RULES FOR THE OUTPUT K_I

e / de	*NB*	*NM*	*NS*	*ZE*	*PS*	*PM*	*PB*
NB	B	B	B	B	B	B	B
NM	B	S	S	S	S	S	S
NS	B	B	S	S	S	S	S
ZE	B	B	B	S	B	B	B
PS	B	B	S	S	S	B	B
PM	B	B	B	S	B	B	B
PB	B	B	B	B	B	B	B

Once the values K_p and K_i obtained the new parameters of the PI controller are calculated by the equations:

$$\begin{cases} K_p = (K_{p-max} - K_{p-min}) / K'_p + K_{p-min} \\ K_i = (K_{i-max} - K_{i-min}) / K'_i + K_{i-min} \end{cases} \tag{37}$$

V. SIMULATION RESULT

In this section, simulations are realized with a 7.5 KW generator coupled to a 380V/50Hz grid. Parameters of the machine are given in appendix A. With an aim to evaluate the performances of the three controllers: PI, SMC and AFLC, three categories of tests have been realized: pursuit test, sensitivity to the speed variation and robustness against machine parameter variations.

A. Reference tracking

The objective of this test is the study of the three controllers' behavior in reference tracking, while the machine's speed is considered constant and equal to its nominal value. The simulation results are presented in Fig. 7. As it's shown by this Figure, for the three controllers, the stator active and reactive powers tracks almost perfectly their references but with an important response time for the PI controller compared to the other controllers. On the other hand it can be noticed that the SMC ensures a perfect decoupling between the two axis, nevertheless a clear coupling effect is appear on the curves with PI and AFLC. Therefore it can be considered that the SMC have a very good performance for this test. In addition, through this same figure we notice that the two components of rotor current have forms that reflect Eq. 23.

B. Sensitivity to the speed variation

The aim of this test is to analyze the influence of a speed variation of the DFIG on active and reactive powers for the three controllers. For this objective and at time = 0.04s, the speed was varied from 150 rad/s to 170 rad/s (Fig. 8).

The simulation results are shown in Fig. 9. This figure express that the speed variation produce a slight effect on the powers curves of the system with PI and AFLC controllers, while the effect is almost negligible for the system with SMC one. It can be noticed that this last has a nearly perfect speed disturbance rejection, indeed; only very small power variations can be observed (fewer than 2%). This result is attractive for wind energy applications to ensure stability and quality of the generated power when the speed is varying.

C. Robustness tests

To test the robustness of the controllers used, parameters of the machine have been modified: the values of the rotor and the stator resistances R_r and R_s are multiplied by 2, while the values of inductances L_r, L_s and M are divided by 2. The machine is running at its nominal speed. The results presented in Fig. 10 show that parameter variations of the DFIG present a clear effect on the power curves (their errors curves) and that the effect appears more significant for PI and AFLC controllers than that with SMC one. Thus it can be concluded that this last is the most robust controller used in this work.

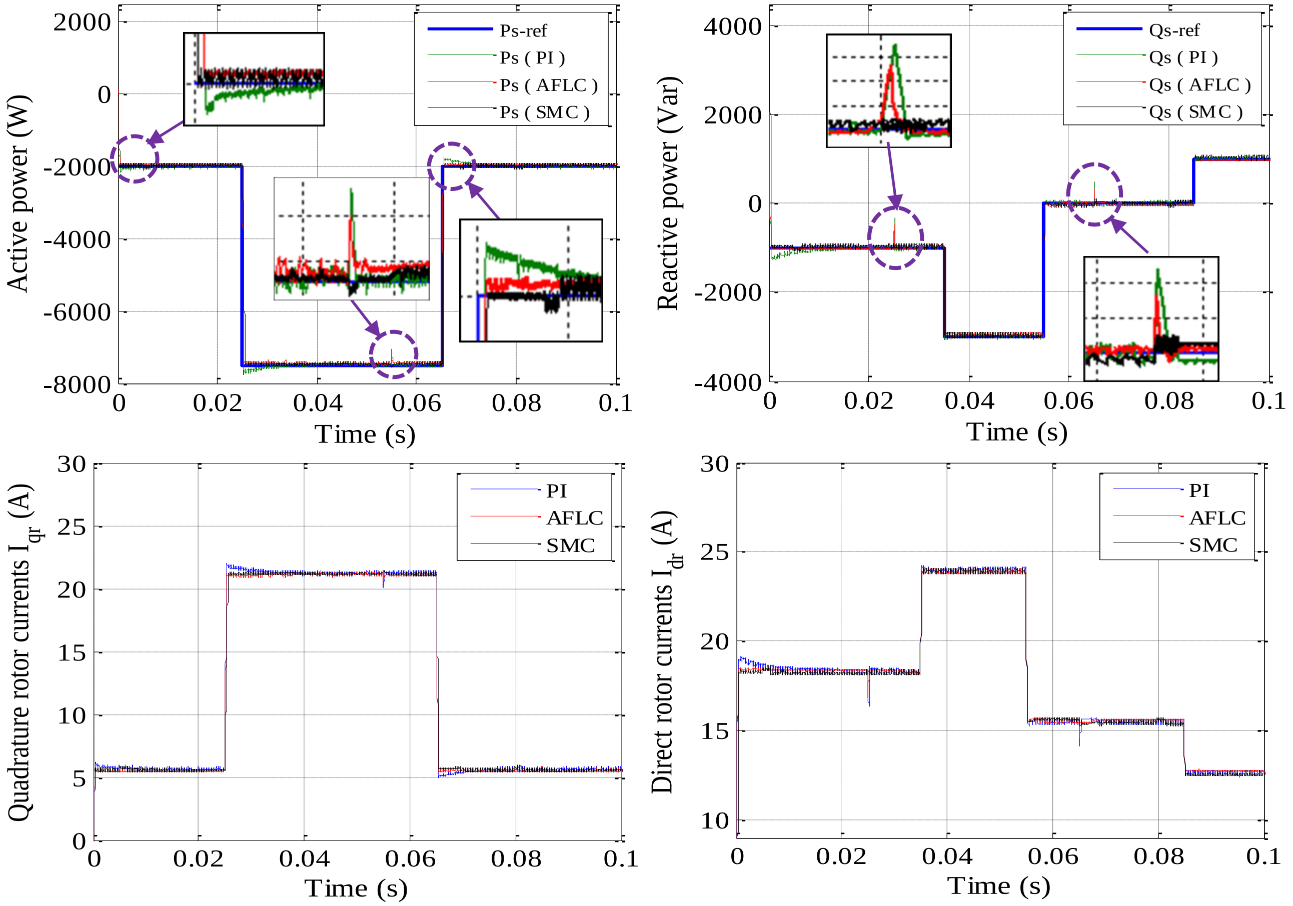


Fig. 7. Reference tracking test

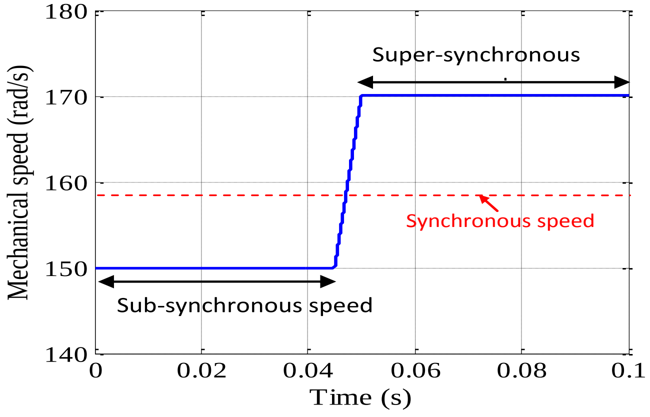


Fig. 8. Mechanical speed profile

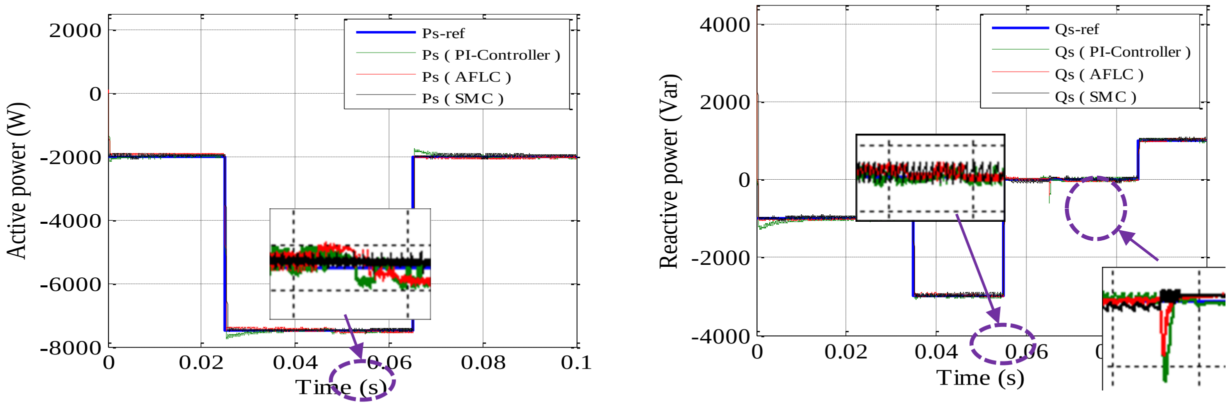


Fig. 9. Sensitivity to the speed variation

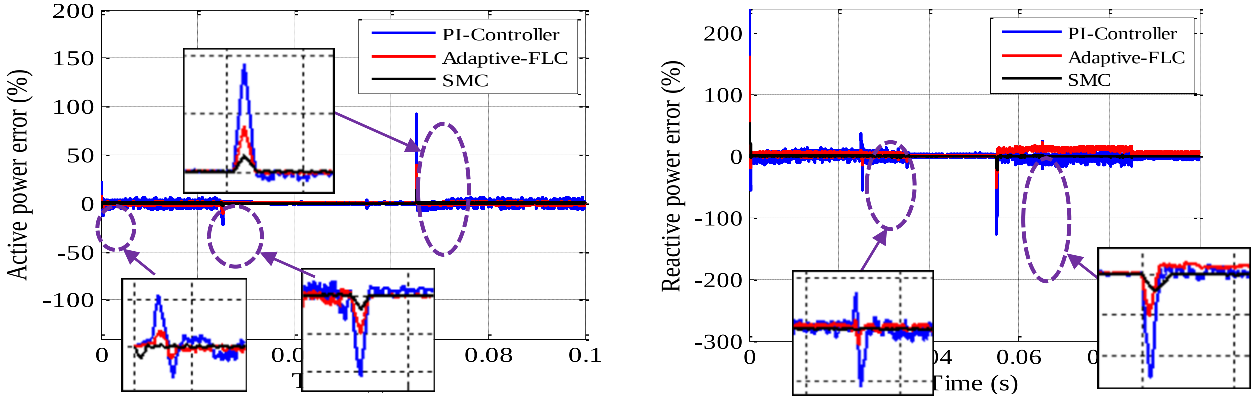


Fig. 10. Effect machine’s parameters variation on the robust control of the DFIG

VI. Conclusions

The modeling, the control and the simulation of an electrical power conversion system based on a DFIG connected directly to the grid by the stator and fed by a matrix converter On the rotor side has been presented in this paper. The objective was the implementation of a robust decoupled control the system of active and reactive powers generated by the side stator of the DFIG, to ensure of the high performance and a better execution of the DFIG, and to make the system insensible with the external disturbances and the parametric variations. In the first step, we started with a study of modeling on the matrix converter controlled by the Venturini modulation technique, because this later present a reduced harmonic rate and the possibility of operation of the converter at the input Unit power factor. The second step, we adopted a vector control strategy to control active and reactive power Exchanged between the stator of the DFIG and the grid. In third step, three different controllers are synthesized and compared. Regarding power reference tracking with the DFIG in ideal conditions, the SMC ensures a perfect decoupling between the two axes comparatively to the other controllers where the coupling effect between them is clear.

When the machine’s speed is modified, a slight effect is appeared on the powers curves of the system with PI and AFLC controllers, while the effect is almost negligible for a system with SMC one. A robustness test has also been investigated where the machine’s parameters have been modified. These changes induce some disturbances on the power responses but with an effect almost doubled with the PI And AFLC controllers than on that with SMC one. Basing on all these results we conclude that robust control method as SMC can be a very attractive solution for devices using DFIG such as wind energy conversion systems.

Appendix

TABLE III. Machine Parameters

Parameters	Value	IS-Unit
Nominal power	7.5	KW
Turbine radius	35.5	m
Gearbox gain	90	
Stator voltage	398	V
Stator frequency	50	Hz
Number of pairs poles	2	
Nominal speed	150	rad/s
Rotor resistance	0.62	Ω
Stator inductance	0.084	H
Rotor inductance	0.081	H
Mutual inductance	0.078	H
Inertia	0.01	$Kg.m^2$

TABLE IV. LIST OF SYMBOLS

Symbol	Significance
V_{ds}, V_{qs}, V_{dr}, V_{qr}	Two-phase stator and rotor voltages,
φ_{ds}, φ_{qs}, φ_{dr}, φ_{qr}	Two-phase stator and rotor fluxes,
I_{ds}, I_{qs}, I_{dr}, I_{qr}	Two-phase stator and rotor currents,
R_s, R_r	Per phase stator and rotor resistances,
L_s, L_r	Per phase stator and rotor inductances,
M	Mutual inductance,
p	Number of pole pairs,
s	Laplace operator,
ω_s, ω_r	Stator and rotor currents frequencies (rad/s),
ω	Mechanical rotor frequency (rad/s),
P_s, Q_s	Active and reactive stator power,
J	Inertia,
f	Coefficient of viscous frictions,
C_r	Load torque,
C_{em}	Electromagnetic torque.

REFERENCES

[1] O. Anaya-Lara, N. Jenkins, J. Ekanayake, P. Cartwright, and M. Hughes,: *Wind energy generation*, In: Wiley, 2009.

[2] L. Empringham, J. W. Kolar, J. Rodriguez, P. W. Wheeler, and J. C. Clare,: *Technological issues and industrial application of matrix converters: A review*, In: *IEEE Transactions on Industrial Electronics*, vol. 60, no. 10, pp. 4260–4271, Oct. 2013.

[3] F. Poitiers, T. Bouaouiche, and M. Machmoum,: *Advanced control of a doubly fed induction generator wind energy conversion*, In: Electric Power Systems Research, 79 (2009) 1085–1096.

[4] J. P. da Costa, H. Pinheiro, T. Degner and G. Arnold,: *robust controller for DFIGs of grid-connected wind turbines*, IEEE Transactions on Industrial Electronics, Vol. 58, no. 9, September (2011) 4023-4038.

[5] M. A. Poller,: *Doubly-fed induction machine models for stability assessment of wind farms*, In: Power Tech Conference Proceedings, 2003, IEEE, Bologna, vol. 3, 23–26 June 2003.

[6] T. Brekken, N. Mohan,: *A novel doubly-fed induction wind generator control scheme for reactive power control and torque pulsation compensation under unbalanced grid voltage conditions*, In: IEEE 34th Annual Power Electronics Specialist Conference, 2003, PESC'03, vol. 2, pp. 760-764 15-19 June 2003.

[7] T. K. A. Brekken, N. Mohan,: *Control of a doubly fed induction wind generator under unbalanced grid voltage conditions*, In: IEEE Transaction on Energy Conversion, 129–135, 22 (March (1)) 2007.

[8] J. Lopez, P. Sanchis, X. Roboam, and L. Marroyo,: *Dynamic behavior of the doubly fed induction generator during three-phase voltage dips*, In: IEEE Transaction on Energy Conversion, 709-717, 22 (September (3)) 2007.

[9] T. Sun, Z. Chen, and F. Blaabjerg,: *Flicker study on variable speed wind turbines with doubly fed induction generators*, IEEE Transactions on Energy Conversion, 896–905, 20 (December (4)) 2005.

[10] M. A. A. Morsy, M. Said, A. Moteleb, and H. Dorrah,: *Design and Implementation of Fuzzy Sliding Mode Controller for Switched Reluctance Motor*, Proceedings of the International Multi-Conference of Engineers and Computer Scientists, Vol. 2, IMECS, Hong Kong, 19-21 March 2008.

[11] O. Barambones and J. M. Gonzalez de Durana,: *An Adaptive Sliding Mode Control Law for the Power Maximization of the Wind Turbine System*, In: 2011 International Conference on Power Engineering, Energy and Electrical Drives (POWERENG), pp. 1- 6, Spain (Malaga), 11-13 May 2011.

[12] E. S. Abdin, W. Xu,: *Control design and Dynamic Performance Analysis of a Wind Turbine Induction Generator Unit*, IEEE Trans. On Energy conversion, vol.15, No1, March 2000.

[13] M. Venturini,: *A new sine wave in sine wave out conversion technique which eliminates reactive elements*, In: Proc Powercon 7, San Diego, CA, pp. E3-1, E3-15, 27-24 March 1980.

[14] Z. Yan, C. Jin, and V. I. Utkin,: *Sensorless sliding mode control of induction motors*, In: IEEE Trans. Ind. electronic. 47 (2000).

[15] D. Kairous and B. Belmadani, *Robust Fuzzy-Second Order Sliding Mode based Direct Power Control for Voltage Source Converter*, International Journal of Advanced Computer Science and Applications (IJACSA), pp. 167-175, vol. 6, no. 8, 2015.

[16] Ou Sheng, Liu Haishan, Liu Guoying, Zeng Guohui, Zhan Xing, Wang Qingzhen, Liu Haishan, *A Fuzzy PI Speed Controller based on Feedback Compensation Strategy for PMSM*, International Journal of Advanced Computer Science and Applications (IJACSA), pp. 49-54, vol. 6, no. 5, 2015.

Android Malware Detection & Protection: A Survey

Saba Arshad
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Abid Khan
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Munam Ali Shah
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Mansoor Ahmed
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

***Abstract*—Android has become the most popular smartphone operating system. This rapidly increasing adoption of Android has resulted in significant increase in the number of malwares when compared with previous years. There exist lots of antimalware programs which are designed to effectively protect the users' sensitive data in mobile systems from such attacks. In this paper, our contribution is twofold. Firstly, we have analyzed the Android malwares and their penetration techniques used for attacking the systems and antivirus programs that act against malwares to protect Android systems. We categorize many of the most recent antimalware techniques on the basis of their detection methods. We aim to provide an easy and concise view of the malware detection and protection mechanisms and deduce their benefits and limitations. Secondly, we have forecast Android market trends for the year up to 2018 and provide a unique hybrid security solution and take into account both the static and dynamic analysis an android application.**

***Keywords*—Android; Permissions; Signature**

I. Introduction

Since 2008, the rate of smartphone adoption has increased tremendously. Smartphones provide different connectivity options such as Wi-Fi, GSM, GPS, CDMA and Bluetooth etc. which make them a ubiquitous device. Google says, 1.3 million Android devices are being activated each day [1]. Android operating system left its competitors far behind by capturing more than 78% of total market share in 2013 [2]. Gartner report 2013 of smartphone sales shows that there is 42.3% increase in sales of smartphones in comparison with 2012. According to International data corporation IDC, Android OS dominates with 82.8% of total market shares in 2Q 2015 [3]. Figure 1 shows the market shares of Android operating system on yearly basis. It could be observed that Android has become the most widely used operating system over the years.

Android platform offers sophisticated functionalities at very low cost and has become the most popular operating system for handheld devices. Apart from the Android popularity, it has become the main target for attackers and malware developers. The official Android market hosts millions of applications that are being downloaded by the users in a large number everyday [4]. Android offers an open market model where no any application is verified by any security expert and this makes Android an easy target for developers to embed malicious content into their applications. The users' sensitive data can be easily compromised and can be transferred to other servers. Furthermore, the existence of third party application stores contribute in spreading malwares for Android because Google Play also hosts the applications of third-party developers. Android official market uses Bouncer for protection of marketplace against malwares [5]. However, Bouncer does not analyze the vulnerabilities of the uploaded apps. Malware developers take advantage of vulnerabilities among apps by repackaging the popular apps of Google Play and distributing them on other third-party app-stores. This degrades the reputation of the app-store and of the reputation of the developer. Malwares includes computer viruses, Trojan horses, adware, backdoors, spywares and other malicious programs which are designed to disrupt or damage the operating system and to steal personal, financial, or business information. Malware developers use code obfuscation methods, dynamic execution, stealth techniques, encryption and repackaging to bypass the existing antimalware techniques provided by Android platform.

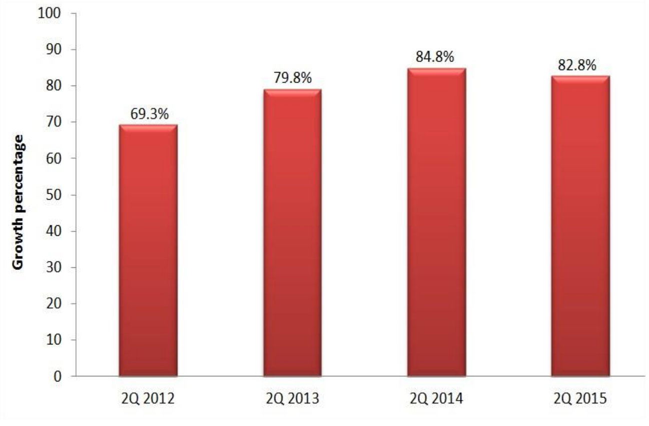


Fig. 1. Android Market Shares

In order to prevent such malwares, it is important to have accurate and deep understanding of them so that security measures to protect users' data could be taken accordingly. There are large numbers of attack scenarios where an attacker can compromise a user's data by taking advantage of the vulnerabilities of Android operating system. For example, a

Trojan app downloads some HD wallpapers with user's permission but this permission may allow this app to access the user's contacts or other personal information and it leaks user's confidential data to some other server from the device secretly. In such a case, the wallpapers app will have Internet permissions for download purpose. The user might not give much attention towards other requested access permissions and might grant READ_CONTATCS permission accidentally. As a result, the app may modify the device settings, corrupt the user's data and can transfer private data to some unknown remote servers. This results in user's business data loss and other personal information. The attackers can use the stolen data for kidnapping, blackmailing or business loss purposes. In an another attack scenario, attackers distribute the malicious apps as a repackaged version of some popular apps which may offer location-based services so in that scenario malicious app kill the victim device by draining its battery with the excessive use of GPS and radio etc. Some of the malicious programs get the user's device IMEI numbers and send it to remote server. These IMEI numbers have significant worth in black markets where IMEI numbers of stolen devices can be altered with user's IMEI [6].

There are hundreds of malware techniques identified which attack the Android platforms in several ways such as sending messages without the victim's knowledge and deleting them by itself, sending user's private information to some other server and many more. So there is a great need to protect user's data from these malwares.

This ever increasing malware threats have forced the Android antimalware industry to develop the solutions for mitigating malicious app threat on Android smartphones and other Android devices. Two main approaches are used for this purpose: *Static approach* and *Dynamic approach*. Antivirus programs use any of these approaches to protect the mobile systems from the malware attacks. They detect the malicious apps and notify the user about such apps and take measures to remove these malwares**.** With the increasing number of threat level, the antivirus detection rate has also increased. As a result of threat & malware, and protection mechanism offered by Android antimalware programs, the overall risk situation of Android users is difficult to assess [7].

In this paper, we have analyzed different malwares, their behaviors and techniques used by different malware types to attack Android devices. Furthermore, the paper provides detailed review on different antimalware techniques, their advantages and limitations. On the basis of this review, a hybrid solution for Android security has been proposed. The rest of the paper is organized as follow. Section II classifies the existing malwares on the basis of their behavior. Section III consists of malware penetration techniques employed by the attackers. In Section IV, a detailed analysis on the malware detection and removal methods for the protection of Android devices has been performed. Section V consists of performance evaluation of antimalware mechanisms. The future trends for Android market shares and malware growth and limitations for existing antimalware approaches are provided in Section VI. A solution has also been proposed in this section which is aimed at providing better security mechanism. The paper is concluded in Section VII.

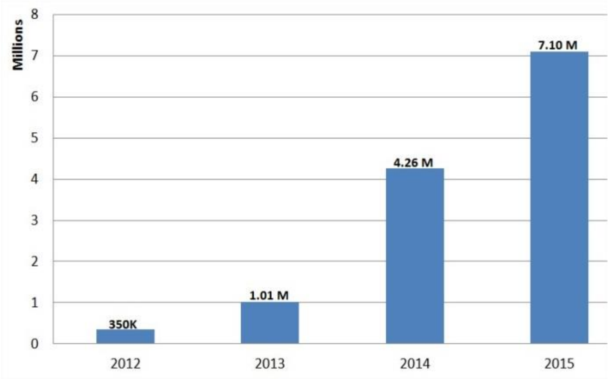


Fig. 2. Android Malware Growth

II. Android Malware Analysis

Wide range of malwares has been detected and the number of malwares are increasing every year. According to TrendMicro, malwares have increased to 7.10 million in first half (1H) of 2015 [8][9]. Figure 2 shows the increased number of Android malwares over the years. The behavior of different malware families is provided in subsequent sections.

A. Trojans

Trojans appear to a user as a Benign app [5]. In fact, they actually steal the user's confidential information without the user's knowledge. Such apps can easily get access to the browsing history, messages, contacts and device IMEI numbers etc. of victim's device and steal this information without the consent of user. *FakeNetflix* [10] is an example of such malwares that provide user interface identical to original Netflix app and collect the user's login credentials. SMS Trojans exploit the premium services to incur financial loss to the victim. *Fakeplayer* is a well-known SMS Trojan that sends messages to premium rate numbers without user awareness [11]. *Zsone* [12] and *Android.foney* are also the examples of such SMS Trojan apps. Malwares also capture the user's banking information such as account number and password. *Zitmo* and *Spitmo* Trojans are designed to steal the user's mTANs (Mobile Transaction Authentication Number) which then complete the transactions silently [13].

B. Backdoors

Backdoors employ the root exploits to grant root privileges to the malwares and facilitate them to hide from antiviruses. *Exploid*, *Rageagainstthecage (RATC)* and *Zimperlich* are the top three root exploits which gain full- control of device [14]. *DroidKungFu* [15] uses root exploits, *Exploid and Rageagainstthecage,* in an encrypted form. When *DroidKungFu* executes, it first decrypts and launches the root exploits. If the root exploit succeed to gain control over device and root privilege, the malware become able to perform any operation on the device even the installation of applications keeping the user unaware of this act [16].

C. Worms

Such malwares create copies of it and distribute them over the network. For example, Bluetooth worms spread malware through the Bluetooth network by sending copies of it to the paired devices. *Android.Obad.OS* is the example of Bluetooth worm [17].

D. *Spyware*

Nickspy [11] and *GPSSpy* [18] are the examples of spyware apps which appear as benign app, but it actually monitors the user's confidential information such as messages, contacts, bank mTANs, location etc. for some undesirable consequences. Personal spywares can install the malicious payload without the victim's knowledge. It sends the user's information such as text messages, contacts etc. to the attacker who installed that software on victim's device [6].

E. *Botnets*

Botnet is a network of compromised Android devices. Botmaster, a remote server, controls the botnet through the C&C network. Geinimi [11] is one of the Android botnets.

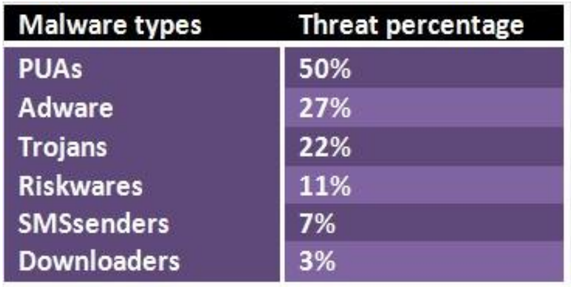
F. *Ransomwares*

Ransomware prevent the user from accessing their data on device by locking the device, until ransom amount is paid. FakeDefender.B [19] is a malware that masquerades itself as avast!, an antivirus. It locks the victim's device and force the user to pay ransom amount to unlock the device.

G. *Riskwares*

Riskwares are the legitimate software exploited by the malicious authors to reduce the performance of device or harm the data e.g., delete, copy or modify etc. [20]. Table 1 below shows the top malware types detected in 2015 by TrendMicro [21].

TABLE I. TOP ANDROID MALWARE TYPES IN 2015



Malware types	Threat percentage
PUAs	50%
Adware	27%
Trojans	22%
Riskwares	11%
SMSsenders	7%
Downloaders	3%

The statistical data obtained from [21] has been computed and plotted in Figure 3 which presents the top Android malware families recorded by TrendMicro in second quarter (2Q) of 2015. According to the report, 24% of the total malwares were guided variants, which do not have any GUIs and silently run at the background without the user's knowledge.

III. MALWARE PENETRATION TECHNIQUES

A. *Repackaging*

Malware authors repackage the popular applications of Android official market, Google Play, and distribute them on other less monitored third party app-store. Repackaging includes the disassembling of the popular benign apps, both free and paid; append the malicious content and reassembling of app .This process of repackaging is done by reverse-engineering tools. During repackaging, malicious authors change the signature of repackaged app and so the app seems new to the antimalware. TrendMicro report have shown that 77% of the top 50 free apps available in Google Play are repackaged [22].

B. *Drive By Download*

It refers to an unintentional download of malware in the background. Drive by download attacks occur when a user visit a website that contains malicious content and injects malware into the victim's device without the user's knowledge. Malware developers use *Android/NotCompatible* [23] which is one of the drive-by download app.

C. *Dynamic Payloads*

Malwares also penetrate into Android devices through dynamic payload technique. They encrypt the malicious content and embed it within APK resources. After installation, the app decrypts the encrypted malicious payload and executes the malicious code. Some malwares, instead of embedding payload as resource, download the malicious content from remote servers dynamically and are not detected by static analysis approach [24].

D. *Stealth Malware Techniques*

On Android device malware scanners cannot perform deep analysis because of the availability of limited resources such as battery. Malware developers exploit these hardware vulnerabilities and obfuscate the malicious code to easily bypass the antimalware. Different stealth techniques such as key permutation, dynamic loading, native code execution, code encryption and java reflection are used to attack the victim's device.

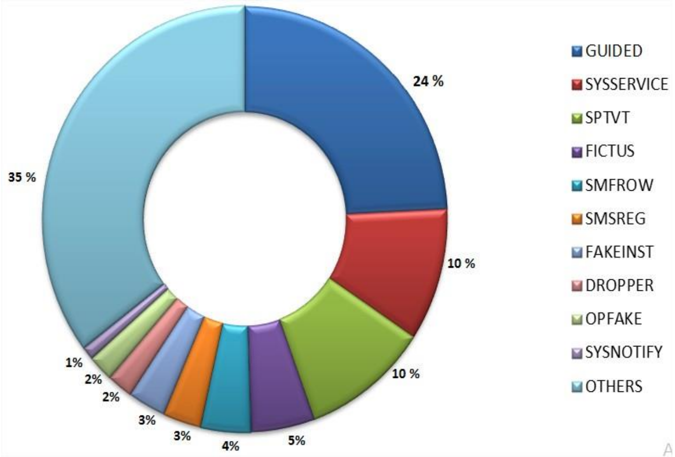


Fig. 3. Malware families seen in 2015

IV. ANDROID MALWARE DETECTION

There are mainly two approaches to analyze the Android malwares: Static and Dynamic Approach. We have further categorized the antimalware using static and dynamic approaches. Figure 4 shows the taxonomy of existing antimalware techniques based on our study.

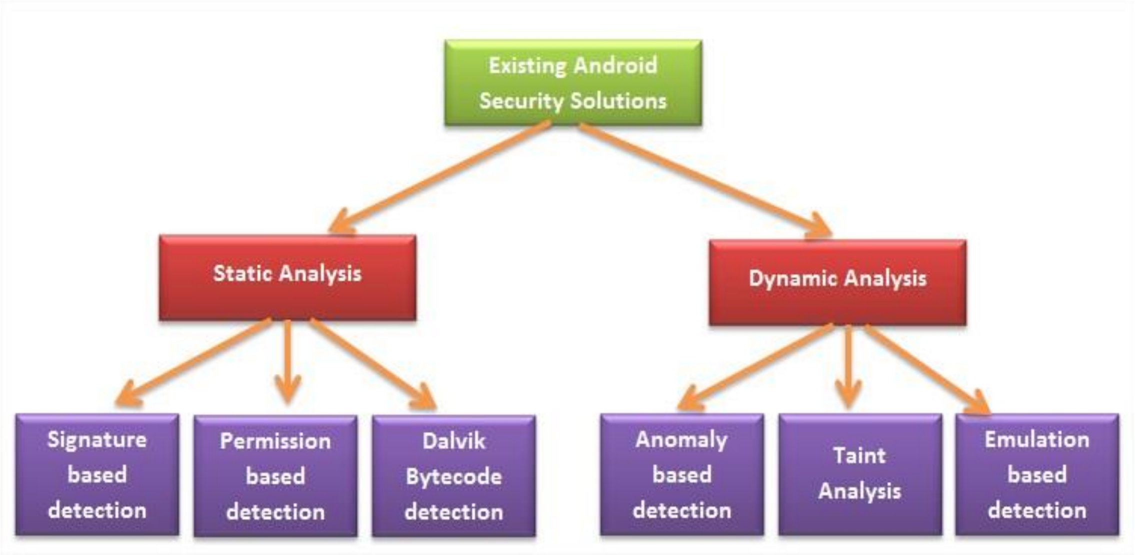

Existing Android Security Solutions
Static Analysis
Dynamic Analysis
Signature based detection
Permission based detection
Dalvik Bytecode detection
Anomaly based detection
Taint Analysis
Emulation based detection

Fig. 4. Taxonomy of Existing Android Antimalwares

A. Static Approach

Static approach is a way to check functionalities and maliciousness of an application by disassembling and analyzing its source code, without executing the application. It is useful for finding malicious behaviors that may not operate until the particular condition occurs.

1) Signature Based Approach

Signature based malware detection methods are commonly used by commercial antimalware products. This method extracts the semantic patterns and creates a unique signature [25]. A program is classified as a malware if its signature matches with existing malware families' signatures. The major drawback of signature based detection is that it can be easily circumvented by code obfuscation because it can only identify the existing malwares and fails against the unseen variants of malwares. It needs immediate update of malware variants as they are detected.

Faruki *et al*. [26] proposed *AndroSimilar*, a robust statistical signature method to detect the unknown variants of existing malwares that are usually generated by using repackaging and code obfuscation techniques. It generates the variable length signature for the application under test and compares it with the signatures in AndroSimilar malware database and identify the app as malware and benign on the basis of similarity percentage. Authors tested the AndroSimilar against 1260 apps among which 6779 apps were Google Play apps and 545 apps were from third party app store. They also used code obfuscation techniques such as method renaming, string encryption, control flow obfuscation and junk method insertion techniques to change the signature of the code and tested the effectiveness of AndroSimilar against 426 samples. The solution detected more than 60% samples correctly. AndroSimilar compares the signatures of the applications in order to distinct between the malwares and benign apps but it has limited signature database as compared to the other antivirus solutions. So any unseen malwares will remain undetected. Also the similarity percentage creates the false positives as it may classify the clean apps as malicious on the basis of percentage.

DroidAnalytics [27] is a signature based analytic system which extract and analyze the apps at op-code level. It not only generates the signature but also associate the malware with existing malwares after identifying the malicious content. It generates 3 level signatures. First it generates signature at method level by API call tracing then combining all the signatures of methods in a class it generates the class level signatures and at third level it generates the application signature by combining the signatures of the classes in the application. Authors have used DroidAnalytics to detect 2,494 malware samples from 102 malware families and 342 repackaged malwares from other six malware families. The limitations of this method includes, it classifies the apps as malware on the basis of classes mostly used by malware families but during experiment they found some signatures that are used by both the legitimate apps and malwares. Also the similarity score used for detection of repackaged malwares do not provide 100% solution or it may also provide false positive, classify the legitimate app as malware.

- *Limitation of Signature Based Detection:* Although signature based detection is very efficient for known malwares but it cannot detect the unknown malware types. Also because of limited signature database most of the malwares remain undetected.

2) Permission Based Analysis:

In Android system, permissions requested by the app plays a vital role in governing the access rights. By default, apps have no permission to access the user' data and effect the system security. During installation, user must allow the app to access all the resources requested by the app. Developers must mention the permissions requested for the resources in the AndroidManifest.xml file. But all declared permissions are not necessarily the required permissions for that specific application.

Ref. [28] has shown that most of the time developers have declared the permissions that are not actually required by the application which makes it difficult to detect the malicious behavior of application. Antimalware analyzes the Android Manifest.xml file where all the permissions for the resources required by the app are mentioned. *Stowaway* [28] exposes the permission over privilege problem in Android where an app requests more permissions than it actually uses. *Stowaway* performs static analysis to determine the API calls invoked by the application and then it maps the permissions required by the API calls. They found that one third applications are over privileged among 940 Android application samples. It cannot resolve the API calls invoked by applications with the use of java reflections.

In [29], authors have proposed a light weight malware detection mechanism which only analyze the manifest file and extract the information such as permissions, intent filters (action, category and priority), process name and number of redefined permissions to detect the malicious behavior of an application. After extracting such information, they compare it with the keyword list provide in the proposed method and then calculate the malignancy score. They used *Weka* [30] which is a data mining tool for calculation of threshold value. At last they compare the malignancy score with threshold value and classify the app as malware if malignancy score exceeds threshold value. They have used 365 samples to test the efficiency of proposed solution and the solution provides 90% accurate detection. It is cost saving mechanism as it only includes the analysis of manifest file and can be implemented in other detection architectures easily to detect malwares efficiently. Also it can detect even those malwares that remain undetected by signature based detection method. This proposed solution is limited to manifest file information. Also it cannot detect the adware samples.

C. Y. Haung *et al*. [31] proposed a method for better detection of permission based malware detection which includes the analysis of both requested and required permissions as most of the time malware authors declare more permissions in the manifest file than they actually require for the application. Also it analyses the easy to retrieve features and then labels the application as benign or malware. Three different labeling types are used for this purpose which includes site based labeling; scanner based labeling and mixed labeling. In site based labeling it labels the app as benign if it is downloaded from Google official app market and if it is downloaded from some malicious source then the app is labeled as malicious. In the second labeling scheme, if the antivirus scanner declares the app as benign the app is label as benign and same for the malware case. In the mixed labeling the app is labeled on the basis of both site based and scanner based labels. After labeling all the samples are divided into three datasets and requested permissions of these datasets are analyzed by the machine learning algorithms such as *Naive Bayes, AdaBoost, Support Vector Machine and Decision Tree* [32]. On the basis of results generated by these classifiers we can evaluate the performance of permission based detection method. in [31] authors have performed experiment on data set of 124,769 benign and 480 malicious apps. They analyzed the performance of permission based detection of malware and showed that more than 81% of malicious apps samples can be detected by the permission based detection method. Proposed method provides the quick filter for malware detection but the performance values generated by the classifiers are not perfect and we cannot completely rely on those results.

Sanz Borja *et al*. [33] presented *PUMA* for detection of malicious apps by analyzing the requested permissions for application. They used permission tags such as <uses-permission> and <uses-features> present in AndroidManifest.xml file to analyze the malicious behavior of apps and applied different classifier algorithms on dataset of 357 benign apps and 249 malicious apps. The solution provides high detection rate but results generated have high false positives rate also it is not adequate for efficient detection of malware it still requires information related to other features and dynamic analysis.

Shin *et al*. [34] used a state machine based approach and formally analyze the permission based Android security model. They also verified that the specified system satisfy the security property.

Tang, Wei *et al*. [35] proposed a Security Distance Model for mitigation of Android malware. Security Distance Model is based on the concept that not a single permission is enough for an application to threaten the security of Android devices. For example an application requesting permission READ_PHONE_STATE can access the phone number and IMEI but it cannot move data out of the device. There must be a combination of permissions to affect the security model of device such as INTERNET permission allows to concept the device with the network and will be needed to move data to some remote server. The SD measure the dangerous level of application on the basis of permissions requested by the app. Authors classify the combinations of permissions into four groups and assigned threat points (TP) to each group such as TP-0, 1, 5 and 25 to Safe SD, Normal SD, Dangerous SD and Severe SD. Before the installation of new application it calculates the threat point from the combination of permissions requested by the application. That helps the user to get aware of more dangerous permissions while installation of app. It can easily detect the unknown malwares with very high threat points. They found 500 threat points for the Geinimi malware which is a very clear variation from benign apps. A limitation of this solution includes that applications with threat points between 50 and 100 are not easy to identify as benign and malware. They could be the benign apps with such permission combinations or malwares.

Enck *et al*. [36] developed *KIRIN*, a tool that provides light weight certification at installation time. It defines the security rules and simply compares the requested permissions of app with its security rules and certifies the app as malware if it fails to pass all the security rules. The installation of app is aborted if the app is attributed as malware. Authors have tested 311 applications downloaded from official Android market and found that 5 applications failed to pass the specified rules. Proposed solution is light weight as it only analyzes the Menifest.xml file. The limitation of *KIRIN* includes that it may also declare some legitimate applications as malware because the information provided for application certification is not adequate for detection of malware.

DroidMat [37] is a tool that extracts the information from manifest file such as permissions, message passing through intents and API call tracing to analyze the behavior of application. It applies K-means clustering that increases the malware detection capability and classify the applications as benign or malware by using KNN algorithm [38]. It is more efficient than Androgaurd [39] as it takes lesser time to identify the 1,738 apps as malware or benign. Also it is cost saving as it doesn't require dynamic simulation and manual efforts. But as a static based detection method it cannot detect the malwares which dynamically load the malicious content such as *DroidKngFu* and *BaseBridge*.

- *Limitation of Permission Based Detection:* Permission based detection is a quick filter for the application scanning and identifying that whether the application is benign or malware but it only analyses the manifest file it do not analyze other files which contain the malicious code. Also there is very small difference in permissions used by the malicious and benign apps. Permission based methods require second pass to provide efficient malware detection.

3) Dalvik Bytecode Analysis:

In Android, Dalvik is a register-based VM. Android apps are developed in java language, compiled in java bytecode and then translated to dalvik byte code. Bytecode analysis helps us to analyze the app behavior. Control and data flow analysis detect the dangerous functionalities performed by malicious apps.

Jinyung Kim *et al.* [40] developed *SCANDAL*, a static analyzer that analyze the dalvik byte code of applications and detects the privacy leakage in applications. It determines the data flow from information source to any remote server. Dalvik bytecode contains branch, method invocation and jump instructions which alters the order of execution of code and obfuscates the code. During execution, the possible paths that an application can take can be identified by the Bytecode analysis. In [40] Authors have examined 90 applications from Android official market and 8 malicious applications from third party market place. They found privacy leakage in 11 Google market applications and 8 third party market applications. There is a need of performance optimization techniques to implement as SCANDAL consumes more time and memory for analysis of application. Also it does not support the applications which use reflections for data leakage. In the SCANDAL authors have implemented reflection semantics manually to detect the privacy leakage in malicious apps taken from black market.

Karlsen *et al*. [41] presented the first formalization of Dalvik Bytecode along with java reflective features. They examined 1700 popular Android Apps to determine what Dalvik Bytecode instructions and features are mostly used by the Android Apps. Such formalization helps to perform control and data flow analysis in order to detect the malicious apps or to identify the sensitive API calls invoked during execution. It supports the dynamic dispatch and reflective features. But it requires extension in analysis of concurrency and reflection handling.

Zhou *et al*. [42] implemented *DroidMOSS* that extract the Dalvik Byte code sequence and developer information of application by using baksmali tool [43] and generate finger prints for each app by using fuzzy hashing techniques to create the fixed sized 80 byte signature to detect the repackaged applications. On the basis of similarity score it identifies the repackaged apps. Authors have applied *DroidMOSS* to test 200 samples from six different third party market places and detected that 5% to 13% apps were repackaged. The proposed solution cannot detect the repackaged apps if the original app is not present in database. Also because of limited database most of the malwares remains undetected. Google play store may also contain malwares. The limitation of this solution also includes that they have assumed all the Google Play apps as legitimate apps and then matched the signature of the apps taken from other app store to detect the repacked apps.

DroidAPIMiner [44], build upon Androgaurd [39], identifies the malware by tracking the sensitive API calls , dangerous parameters invoked and package level information within the bytecode. To classify the application as benign or malware it implements KNN algorithm [38] and detected up to 99 % accuracy and 2.2% false positive rate.

Fuchs *et al*. [45] presented *SCandroid* which analyze the Android application statically as they are installed and performs data flow analysis to checks whether the data flow through the applications is consistent or not. On the basis of data flows it declares the application as safe to be run with requested permissions. Authors use it as a security certification tool for Android apps.

Many researchers worked on conversion of Dalvik bytecode to Java bytecode and then performed static analysis on java code to detect the malicious behavior of the app. *ded* [46] and *Dare* [47] are the tools used for conversion of dalvik bytecode into java bytecode. These tools are also useful when developers don't distribute the java source code, in such case one must analyze the source code to detect the malware through static analysis. *Dexpler* tool [48] converts the Dalvik bytecode into Jimple code which is used by static analysis framework named Soot [49]. It makes the Soot to read the Dalvik Bytecode directly and perform the static analysis without converting Dalvik bytecode into java bytecode. Well known static analysis framework used by researchers is WALA which perform static analysis on java bytecode to detect privacy leakage within malicious apps [50].

Chin *et al*. [51] presented a tool named *ComDroid* that detect the communication based vulnerabilities among Android apps. They have analyzed 20 samples and detected 34 exploitable vulnerabilities among 12 applications. It uses *Dedexer* tool [52] to disassemble the dex files in the app. It performs the static analysis on Dalvik files, analyzes the permissions listed in the manifest.xml file of the app, performs intraprocedural analysis and examines the Intents of the apps to detect the communication vulnerabilities

- *Limitations of Dalvik Bytecode Detection*: In this method analysis is performed at instruction level and consumes more power and storage space. As the android devices are resource poor so they limits this detection approach.

B. *Dynamic Approach*

Dynamic analysis examines the application during execution. It may miss some of the code sections that are not executed but it can easily identify the malicious behaviors that are not detected by static analysis methods. Although static analysis methods are faster to malware detection but they fail against the code obfuscation and encryption malwares.

In [53] , Egele provided a detailed overview of different dynamic analysis methods used for discrimination between malware and benign apps. Dynamic analysis approach is effective against polymorphic and metamorphic code obfuscation techniques employed by the malwares [54] but it requires more resources.

1) Anomaly Based Detection

Iker *et al*. [55] proposed *CrowDroid* to detect the behavior of applications dynamically. Details of system calls invoked by the app are collected by the *Strace* tool [56] and then crowdsourcing app, which is installed on the device, creates a log file and sends it to remote server. Log file may include the following information: Device information, apps installed on device and system calls. 2-mean clustering algorithm is applied at server side to classify the application as malware or benign. Results are stored at server database. The solution provides deep analysis and thus require large amount of resources. The solution requires client app to be installed on the user' device and may classify the legitimate app as malware if it invoke more system calls.

Shabtai *et al*. [57] proposed *Andromly* , a behavior based Android malware detection system. In order to classify the application as benign or malware it continuously monitor the different features and patterns that indicate the device state such as battery level, CPU consumption etc. while it is running and then apply the machine learning algorithms to discriminate between malicious and Benign apps. the solution can detect continuous attacks and can notify the user about these attacks.

AntiMalDroid [58], a malware detection framework using SVM algorithm is proposed by Zhao, can identify the malicious apps and their variants during execution. First it monitors the behavior of applications and their characteristics then it categorize these characteristics as normal and malicious behavior. Then it puts the two types of characteristics into learning module and generates the signatures for the behavior characteristics, produced by learning module. Then it store the signature in database and compare it with the already existing malware and benign app signatures. It classify the app as benign if the signature matches with already existing benign app' signatures. The solution can extend the signature database dynamically and can provide high detection rate. But it consumes more time while detection process.

2) Taint Analysis

Enck *et al*. [59] proposed *TaintDroid* which provides system-wide information flow tracking for Android. It can simultaneously track multiple sources of sensitive data such as camera, GPS and microphone etc. and identify the data leakage in third party developer apps. It labels the sensitive data and keeps track of that data and app when tainted data leaves moves from the device. It provides efficient tracking of sensitive information but it do not perform control flow tracking. Also it cannot track information that leaves deice and returns in network reply.

3) Emulation Based Detection

Yan *et al*. [60] present Android dynamic analysis platform *DroidScope*, based on Virtual Machine Introspection. As the antimalware detect the presence of malwares because both of them reside in the same execution environment so the malwares also can detect the presence of antimalware. *DroidScope* monitors the whole operating system by staying out of the execution environment and thus have more privileges than the malware programs. It also monitors the Dalvik semantics thus the privilege escalation attacks on kernel can also be detected. It is built upon QEMU. *DroidDream* and *DroidKungFu* [61] were detected with this technique.

Blaising *et al*. [62] proposed Android Application Sandbox (*AASandbox*) which detect the suspicious applications by performing both static and dynamic analysis on them. It first extracts the .dex file into human readable form and then performs static analysis on application. Then it analyzes the low level interactions with system by execution of application in isolated sandbox environment. Actions of application are limited to sandbox due to security policy and do not affect the data on device. It uses Money tool to dynamically analyze the application behavior which randomly generates the user events like touches, clicks and gestures etc. it cannot detect the new malware types.

V. Performance Evaluation & Analysis

In this section, we evaluate the performance of different parameters and provide a comprehensive comparison of different attributes. Table 2 provides the limitations of the static and dynamic approach of the malware detection. The malware detection through static analysis and dynamic analysis is provided in Table 3 and Table 4 respectively.

TABLE II. Limitations of Static and Dynamic Approaches

	Mechanism	Limitations
Static	Signature based detection	Cannot detect unknown malware types.
	Permission based detection	May consider benign app as malicious because of very small difference between permissions requested by both types.
	Dalvik bytecode detection	More power and memory consumption.
Dynamic	Anomaly detection	Incorrect if a benign app shows same behaviors e.g., invoke more API calls or consumes more battery and memory.
	Taint Analysis	Not suitable for real time analysis Reduce performance. 20 times slowdown system
	Emulation based detection	More resource consumption.

On the basis of their working techniques we have deduced major limitations and benefits for each detection mechanism.

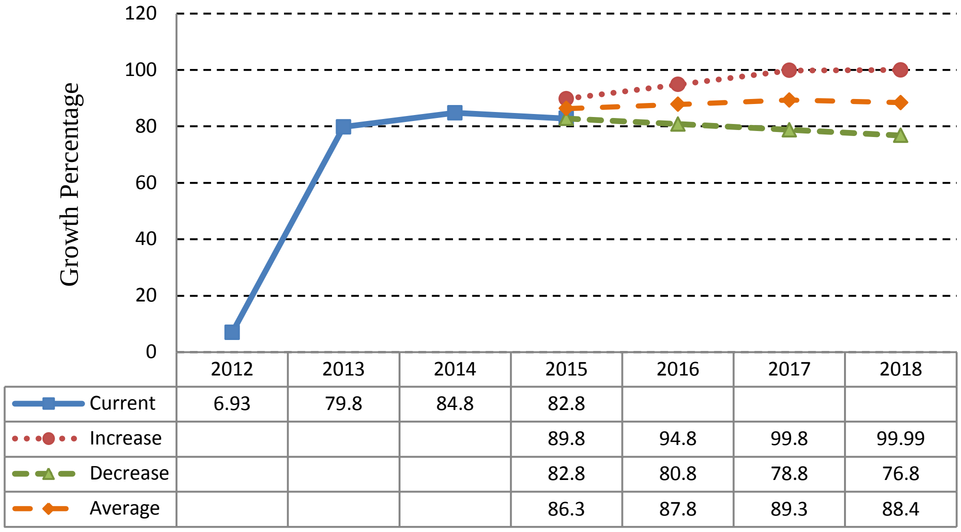
TABLE III. Malware Detection Through Static Analysis

Approach	Name	Goal	Method	Year	Limitations	Benefits
Signature Based Detection	AndroSimilar [26]	Detect unseen and zero day samples of known malwares.	• Creates variable length signature and compares with signature database. • Use fuzzy hashing technique • Differentiates between benign and malicious apps on the basis of similarity percentage.	2013	• Limited signature database • Similarity percentage may classify benign apps as malicious. • Can only detect known malware variants	• Effective against code obfuscation and repackaging.
	DroidAnalytics [27]	Automatic collection, extraction, analysis and association of Android malwares.	• Create 3 level signatures for app on the basis of API calls. • Perform Op-code level analysis (method, class, application). • Correlate application with existing malwares in database via similarity score based on class level signature.	2013	• Similarity score may classify legitimate apps as malicious. • Some level 2 signatures classified as malwares are also used by legitimate apps. • Cannot detect unknown malware types.	• Effective against mutations and repackaged apps. • Associates malware at op-code level • Easy malware and dynamic payload tracking. • Also detect dynamic malware payloads.
Permission Based Detection	Stowaway [28]	Application over privilege detection	• API call tracing through static analysis tool. • Permission map to identify the permissions required by each API cal.	2011	• Cannot resolve complex reflective calls	• Notify about the over privileged applications.
	R.Sato [29]	Malware detection by manifest file analysis.	• Analyze manifest file • Compare extracted information with keyword list. • Calculate malignancy score • Compare malignancy score with threshold values • Classify the app as malware if malignancy score exceeds threshold values.	2013	• Cannot detect adware samples • Generates results only on the basis of manifest file.	• Light weight approach • Low cost • Can detect the unknown malwares. • Can detect the malwares that remain undetectable by signature based detection. • Can be implemented in other security systems for better malware detection.
	C.Y.Haung [31]	Performance evaluation on permission based malware detection.	• Analyze the required and requested permissions for application • Analyze easy to retrieve features • Labels apps as benign or malware using site based, scanner based and mixed labeling • Use machine learning algorithms on three data sets (on the basis of labels) • Evaluate the permission based malware detection performance.	2013	• Performance numbers generated by classifiers are not perfect. • Cannot completely rely on results generated by classifiers. • Ada Boost identifies all apps as legitimate. • Naïve Bayes also do not give précised results.	• Can use different classifiers for different scenarios. • Quick filter for malware detection.
	PUMA [32]	Malware detection	• Analyze extracted permissions • Use the <use permissions> and <use features> tags. • Classify apps by using machine learning algorithms. • Evaluate the performance by k-fold cross validation with k=10.	2013	• High false positive rate • Not adequate for efficient malware detection	• High detection rate
	Tang Wei [34]	Application assessment and analysis to extend android security	• Uses Security Distance Model to measure dangerous level due to combination of requested permissions.	2011	• Applications with threat point between 50 and 100 are difficult to identify as malware or benign apps.	• Provide malware identification during installations. • Can detect unknown malwares

	Kirin [35]	Risk assessment and certification of applications at install time.	• Uses security rules • Compares the security configuration of application with security rules • Certifies the app as malware if app fails to satisfy all the security rules.	2009	• May declare benign app as malware because mostly similar permissions are requested by benign and malicious apps.	• Light weight certification of application at installation time. • Low cost. • Block the malicious applications.
Dalvik Bytecode Detection	SCANDAL [38]	Privacy leak detection	• Extracts bytecode of application as a dalvik executable file • Translates dalvik executable into dalvik core, an intermediate language for efficient analysis	2012	• More time and memory consumption • Needs performance improvement techniques to implement. • Does not support applications that use reflections for privacy leakage • Does not support java native interface libraries	• Saves the data from privacy leakage. • Dalvik bytecode is always available. • Does not need reverse engineering tools
	Karlsen [39]	Dalvik bytecode formalization and control flow analysis	• Provides formal control flow analysis. • Formalizes dalvik bytecode language with reflection features.	2013	• Requires extension in analysis of reflection and concurrency handling.	• Supports reflection and dynamic dispatch features. • Formal control flow analysis easily traces the API calls.
	DroidMOSS [40]	Repackaged malicious app detection	• Extract instructions in app and developer information. • Uses baksmali tool for dalvik bytecode extraction. • Generates fingerprint for each app by applying fuzzy hashing techniques • Measures similarity between apps to detect repackaged apps	2012	• It assumes all the Google Play apps as legitimate apps. • Limited database. • Cannot detect repackaged apps if original app is not present in database.	• Effective detection of repackaged apps.
	DroidAPIMiner [42]	API level Malware detection	• Extract API level features • Apply classifiers for evaluation	2013	• More occurrences of false positives • May generate incorrect classification.	• Better accuracy.
	SCanDroid [43]	Application data flow analysis and security certification	• Analyze data flows in app. • Make decision to classify app as benign or malware on the basis of data flow.	2009	• Cannot be applied to packaged applications.	• Provide security at install time.
	ComDroid [49]	Application communication vulnerability detection	• Extract dalvik executable files • Disassemble DEX files using dedexer tool. • Keep logs of the communication vulnerabilities	2011	• Does not verify the existence of malware • Require users to manually investigate the warnings	• Issue warnings about threats.

TABLE IV. MALWARE DETECTION THROUGH DYNAMIC ANALYSIS

Approach	Name	Goal	Method	Year	Limitations	Benefits
Anomaly Detection	CrowDroid [53]	Detect anomalously behaving malicious applications	• CrowDroid client app installed on user' device. • Strace tool perform system calls tracing. • Creates a log file and send to remote server. • Dynamic analysis is performed on the data at server side. • Consider that malicious apps invoke more system calls.	2011	• Requires the installation of CrowDroid client application to perform detection. • Results incorrect if legitimate app invokes more system calls.	• Provides deep analysis.
	Andromly [55]	Malware detection	• Continuously monitor the features and events e.g., battery level, data packets transferred through Internet, CPU consumption and running processes. • Apply machine learning classifiers to discriminate between benign and malicious applications.	2012	• Only four artificially created malware instances were used for testing the system • Battery drainage issue.	• Can detect the continuous attacks. • Alerts the user about detected anomaly.
	AntiMalDroid [56]	Malware detection through characteristic learning and signature generation.	• Monitor the behavior of applications and their characteristics • Categorize the characteristics into normal behavior and malicious behavior • Put these characteristic types into learning module • Generate behavioral characteristics. • Generate the signatures for these behavioral characteristics • Store these signatures to database. • Compares a signature with the signatures in the database. • Declares as a malware if signature matches with malware signature in database.	2011	• More time consumption.	• Can detect unknown malwares and their variants in runtime. • Extends malware database dynamically. • Higher detection rate • Low cost and better performance.
Taint Analysis	TaintDroid [57]	Data flow analysis and leakage detection	• Automatically labels the data. • Keeps track of the data. • Records the label of the data, source and destination device if the data moves out of the device.	2010	• Only track data flows and do not track control flows. • Cannot track information that leaves the device and return in network reply.	• Efficient tracking of sensitive information
Emulation Based Detection	DroidScope [58]	Android malware analysis	• System calls tracking • Built upon QEMU (quick emulator) • Monitors the OS and Dalvik semantics • Perform virtual machine introspection based dynamic analysis	2012	• Limited code coverage	• Can detect privilege escalation attacks on the kernel.
	AASandbox [60]	Malware detection	• Extracts a class.dex file and decompiles it into human readable form. • Performs static analysis on application. • Executes the application in sandbox and perform dynamic analysis • Uses Monkey tool to analyze the malicious behavior of app.	2010	Cannot detect new malwares	Can be used to improve the efficiency of the antimalware programs for Android OS



	2012	2013	2014	2015	2016	2017	2018
Current	6.93	79.8	84.8	82.8			
Increase				89.8	94.8	99.8	99.99
Decrease				82.8	80.8	78.8	76.8
Average				86.3	87.8	89.3	88.4

Fig. 5. Expected future trends of android OS market share

VI. Discussion

The popularity of Android operating system is increasing tremendously. The yearly records, presented by IDC [3], show that Android OS market shares in second quarter (2Q) of 2015 are 82.8%, which is 2% decrease from the 2Q 2014. If the value remains the same till the end of year and keep on decreasing every year with the same rate then we can expect that in 2018, the Android market shares will drop to 76.8%. According to same record, the Android shares have increased 5% in 2014 from previous year. If it keep on increasing with the same rate and increases up to 89.8% till the end of 2016 then we can say that the Android shares will grow up to 99.9% in 2018. Furthermore, it is predicted that the market shares of the Android will be on average 88.4% in 2018. The estimations and future predictions of the Android market are computed and plotted in Figure 5. It should be noted that with the increased usage of the Android based devices, the number of malwares attacking Android is increasing at an exponential rate. In 2015, number of Android malwares spiked to 7.10 million. This figure is 2.84 million more than the previous year [8][9]. If the malware growth keeps on increasing with the same ratio, it is expected that this number will be increased up to 15.8 million in 2018**.** The malware growth trends are predicted and estimated values are provided in Figure 6.

In contrast to malwares, the antimalware have been designed and developed in a wide range in order to protect the devices. It is inferred that an antimalware using static approach is less efficient in detecting the malicious contents that are loaded dynamically from remote servers. Although, the dynamic approach is efficient as it keeps on monitoring the application and able to detect the malicious content at execution time. However, the portions of malicious code that are not executed remain undetected. It is believed that any single security solution in Android cannot provide full protection against the vulnerabilities and malwares. It is better to deploy more than one solution simultaneously for example, a hybrid of two approaches, i.e. static and dynamic. The hybrid approach will first statically analyze the application and will then perform dynamic analysis. This hybrid solution may be an expensive method to apply because of the limited available resources such as battery, memory etc. However, the limitation of this hybrid solution can be addressed in twofold. Firstly, the static analysis can be performed locally on the Android device; and afterwards, the dynamic analysis could be performed in a distributed fashion by sending the malicious activity or event in the form of a log file to a remote server. The remote server can perform the dynamic analysis quickly and efficiently as the server will have enough resources to perform dynamic analysis and can generate rapid responses against the application behavior and the user can be instantly notified. However, this hybrid solution needs more investigation and is subject to the design tradeoffs. The future works will focus to develop such hybrid antimalware to provide better security for android devices.

VII. Conclusion

In this paper, the malwares and their penetrations techniques have been thoroughly analyzed. The antimalware are categorized on the basis of detection methods they use. A detailed performance evaluation of these antimalware

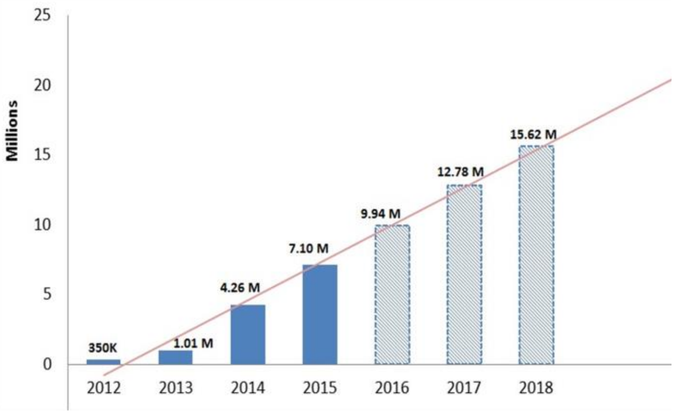


Figure 6: Future Trends of Android Malware Growth

techniques is also provided and the benefits and limitations of these antimalware are deduced comprehensively. At the end, a concept of hybrid antimalware is presented which will address the limitations of existing static and dynamic approaches. In future, it is aimed to implement the proposed hybrid solution which will be a generic antimalware that will provide better security for Android devices by firstly statically analyzing the Android applications on local device and then it will perform dynamic analysis on a remote antimalware server. This will consume very small amount of memory space on the device and the battery consumption will also be low as all dynamic analysis will be performed at the remote server.

References

[1] "Eric Schmidt: 'There Are Now 1.3 Million Android Device Activations Per Day.'" [Online]. Available: http://techcrunch.com/2012/09/05/eric-schmidt-there-are-now-1-3-million-android-device-activations-per-day/. [Accessed: 28-Oct-2015].

[2] "Gartner Says Annual Smartphone Sales Surpassed Sales of Feature Phones for the First Time in 2013." [Online]. Available: http://www.gartner.com/newsroom/id/2665715. [Accessed: 28-Oct-2015].

[3] "IDC: Smartphone OS Market Share 2015, 2014, 2013, and 2012." [Online]. Available: http://www.idc.com/prodserv/smartphone-os-market-share.jsp. [Accessed: 08-Dec-2015].

[4] "Number of available Android applications - AppBrain." [Online]. Available: http://www.appbrain.com/stats/number-of-android-apps. [Accessed: 28-Oct-2015].

[5] "Android and Security - Official Google Mobile Blog." [Online]. Available: http://googlemobile.blogspot.in/2012/02/android-and-security.html. [Accessed: 28-Oct-2015].

[6] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner, "A survey of mobile malware in the wild," Proc. 1st ACM Work. Secur. Priv. smartphones Mob. devices - SPSM '11, pp. 3 – 14, 2011.

[7] R. Fedler, J. Schütte, and M. Kulicke, "On the Effectiveness of Malware Protection on Android," p. 36, 2013.

[8] "Mind the (Security) Gaps: The 1H 2015 Mobile Threat Landscape - Security News - Trend Micro USA." [Online]. Available: http://www.trendmicro.com/vinfo/us/security/news/mobile-safety/mind-the-security-gaps-1h-2015-mobile-threat-landscape. [Accessed: 08-Dec-2015].

[9] "The Mobile Landscape Roundup: 1H 2014 - Security News - Trend Micro USA." [Online]. Available: http://www.trendmicro.com/vinfo/us/security/news/mobile-safety/the-mobile-landscape-roundup-1h-2014. [Accessed: 08-Dec-2015].

[10] R. Raveendranath, V. Rajamani, A. J. Babu, and S. K. Datta, "Android malware attacks and countermeasures: Current and future directions," 2014 Int. Conf. Control. Instrumentation, Commun. Comput. Technol., pp. 137–143, 2014.

[11] Y. Zhou and X. Jiang, "Dissecting Android Malware: Characterization and Evolution," 2012 IEEE Symp. Secur. Priv., no. 4, pp. 95–109, 2012.

[12] "Security Alert: Zsone Trojan found in Android Market | Lookout Blog." [Online]. Available: https://blog.lookout.com/blog/2011/05/11/security-alert-zsone-trojan-found-in-android-market/. [Accessed: 15-Dec-2015].

[13] L. Davi, A. Dmitrienko, C. Liebchen, and A.-R. Sadeghi, "Over-the-Air Cross-platform Infection for Breaking mTAN-based Online Banking Authentication," Black Hat Abu Dhabi, pp. 1–12, 2012.

[14] "root exploits." [Online]. Available: http://www.selinuxproject.org/~jmorris/lss2011_slides/caseforseandroid.pdf. [Accessed: 15-Dec-2015].

[15] "Trojan: Android/DroidKungFu.C Description | F-Secure Labs." [Online]. Available: https://www.f-secure.com/v-descs/trojan_android_droidkungfu_c.shtml. [Accessed: 15-Dec-2015].

[16] Y. Zhou, Z. Wang, W. Zhou, and X. Jiang, "Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets," Proc. 19th Annu. Netw. Distrib. Syst. Secur. Symp., no. 2, pp. 5–8, 2012.

[17] "contagio mobile: Backdoor.AndroidOS.Obad.a." [Online]. Available: http://contagiominidump.blogspot.in/2013/06/backdoorandroidosobada.html. [Accessed: 28-Oct-2015].

[18] C. a Castillo, "Android Malware Past , Present , and Future," McAfee White Pap. Mob. Secur. Work. Gr., pp. 1–28, 2011

[19] "Android.Fakedefender.B | Symantec." [Online]. Available: https://www.symantec.com/security_response/writeup.jsp?docid=2013-091013-3953-99. [Accessed: 15-Dec-2015].

[20] "Riskware | Internet Security Threats." [Online]. Available: http://usa.kaspersky.com/internet-security-center/threats/riskware#.Vm-5IUp97IU. [Accessed: 15-Dec-2015].

[21] "Trend Micro Q2 Security Roundup Report | Androidheadlines.com." [Online]. Available: http://www.androidheadlines.com/2015/08/trend-micro-q2-security-roundup-report.html. [Accessed: 08-Dec-2015].

[22] "A Look at Repackaged Apps and their Effect on the Mobile Threat Landscape." [Online]. Available: http://blog.trendmicro.com/trendlabs-security-intelligence/a-look-into-repackaged-apps-and-its-role-in-the-mobile-threat-landscape/. [Accessed: 15-Dec-2015].

[23] "NotCompatible Android Trojan: What You Need to Know | PCWorld." [Online]. Available: http://www.pcworld.com/article/254918/notcompatible_android_trojan_what_you_need_to_know.html. [Accessed: 15-Dec-2015].

[24] New Threats and Countermeasures in Digital Crime and Cyber Terrorism. IGI Global, 2015.

[25] A. Aiken, "Apposcopy : Semantics-Based Detection of Android Malware Through Static Analysis," Fse 2014, pp. 576–587, 2014.

[26] P. Faruki, V. Ganmoor, V. Laxmi, M. S. Gaur, and A. Bharmal, "AndroSimilar: Robust Statistical Feature Signature for Android Malware Detection," Proc. 6th Int. Conf. Secur. Inf. Networks, pp. 152–159, 2013.

[27] M. Zheng, M. Sun, and J. C. S. Lui, "DroidAnalytics : A Signature Based Analytic System to Collect , Extract , Analyze and Associate Android Malware," 2013.

[28] Android Permissions Demystified." [Online]. Available: https://www.truststc.org/pubs/848.html. [Accessed: 06-Nov-2015].

[29] R. Sato, D. Chiba, and S. Goto, "Detecting Android Malware by Analyzing Manifest Files," pp. 23–31, 2013.

[30] "Weka 3 - Data Mining with Open Source Machine Learning Software in Java." [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed: 16-Dec-2015].

[31] C.-Y. Huang, Y.-T. Tsai, and C.-H. Hsu, "Performance evaluation on permission-based detection for android malware," Adv. Intell. Syst. Appl. - Vol. 2, vol. 21, pp. 111–120, 2013.

[32] S. Ben-david, Understanding Machine Learning : From Theory to Algorithms. 2014.

[33] B. Sanz, I. Santos, C. Laorden, X. Ugarte-Pedrero, P. G. Bringas, and G. Álvarez, "PUMA: Permission usage to detect malware in android," Adv. Intell. Syst. Comput., vol. 189 AISC, pp. 289–298, 2013.

[34] W. Shin, S. Kiyomoto, K. Fukushima, and T. Tanaka, "Towards formal analysis of the permission-based security model for Android," 5th Int. Conf. Wirel. Mob. Commun. ICWMC 2009, pp. 87–92, 2009.

[35] W. Tang, G. Jin, J. He, and X. Jiang, "Extending android security enforcement with a security distance model," 2011 Int. Conf. Internet Technol. Appl. iTAP 2011 - Proc., 2011.

[36] W. Enck, M. Ongtang, and P. McDaniel, "On lightweight mobile phone application certification," Proc. 16th ACM Conf. Comput. Commun. Secur. - CCS '09, pp. 235–245, 2009.

[37] D.-J. Wu, C.-H. Mao, T.-E. Wei, H.-M. Lee, and K.-P. Wu, "DroidMat: Android Malware Detection through Manifest and API Calls Tracing," 2012 Seventh Asia Jt. Conf. Inf. Secur., pp. 62–69, 2012.

[38] L. Kozma, "k Nearest Neighbors algorithm (kNN)," 2008.

[39] "androguard - Reverse engineering, Malware and goodware analysis of Android applications ... and more (ninja !) - Google Project Hosting." [Online]. Available: https://code.google.com/p/androguard/. [Accessed: 01-Dec-2015].

[40] J. Kim, Y. Yoon, and K. Yi, "S CAN D AL : Static Analyzer for

Detecting Privacy Leaks in Android Applications.”

[41] E. R. Wognsen, H. S. Karlsen, M. C. Olesen, and R. R. Hansen, “Formalisation and analysis of Dalvik bytecode,” Sci. Comput. Program., vol. 92, no. December 2012, pp. 25–55, 2014.

[42] W. Zhou, Y. Zhou, X. Jiang, and P. Ning, “Detecting repackaged smartphone applications in third-party android marketplaces,” Proc. Second ACM Conf. Data Appl. Secur. Priv. - CODASKY ’12, pp. 317–326, 2012.

[43] “[Utility][Tool][Windows] Baksmali / Smali Ma… | Android Development and Hacking.” [Online]. Available: http://forum.xda-developers.com/showthread.php?t=2311766. [Accessed: 22-Dec-2015].

[44] Y. Aafer, W. Du, and H. Yin, “DroidAPIMiner: Mining API-Level Features for Robust Malware Detection in Android,” Secur. Priv. Commun. Networks, vol. 127, pp. 86–103, 2013.

[45] A. P. Fuchs, A. Chaudhuri, and J. S. Foster, “SCanDroid : Automated Security Certification of Android Applications.”

[46] W. Enck, D. Octeau, and P. Mcdaniel, “A Study of Android Application Security,” no. August, 2011.

[47] D. Octeau, S. Jha, and P. McDaniel, “Retargeting Android applications to Java bytecode,” in Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering - FSE ’12, 2012, p. 1.

[48] A. Bartel, J. Klein, M. Monperrus, and Y. Le Traon, “Dexpler: Converting Android Dalvik Bytecode to Jimple for Static Analysis with Soot,” 2012.

[49] “A framework for analyzing and transforming Java and Android Applications.” [Online]. Available: http://sable.github.io/soot/. [Accessed: 07-Nov-2015].

[50] “Main Page - WalaWiki.” [Online]. Available: http://wala.sourceforge.net/wiki/index.php/Main_Page. [Accessed: 07-Nov-2015].

[51] E. Chin, A. Felt, K. Greenwood, and D. Wagner, “Analyzing inter-application communication in Android,” Proc. 9th …, pp. 239–252, 2011.

[52] “Dedexer user’s manual.” [Online]. Available: http://dedexer.sourceforge.net/. [Accessed: 08-Nov-2015].

[53] M. Egele, T. Scholte, E. Kirda, and C. Kruegel, “A survey on automated dynamic malware-analysis techniques and tools,” ACM Comput. Surv., vol. 44, no. 2, pp. 1–42, 2012.

[54] I. You and K. Yim, “Malware obfuscation techniques: A brief survey,” Proc. - 2010 Int. Conf. Broadband, Wirel. Comput. Commun. Appl. BWCCA 2010, pp. 297–300, 2010.

[55] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, “Crowdroid: Behavior-Based Malware Detection System for Android,” Proc. 1st ACM Work. Secur. Priv. smartphones Mob. devices - SPSM ’11, p. 15, 2011.

[56] “strace download | SourceForge.net.” [Online]. Available: http://sourceforge.net/projects/strace/. [Accessed: 22-Dec-2015].

[57] A. Shabtai, U. Kanonov, Y. Elovici, C. Glezer, and Y. Weiss, “‘Andromaly’: a behavioral malware detection framework for android devices,” J. Intell. Inf. Syst., vol. 38, no. 1, pp. 161–190, 2012.

[58] M. Zhao, F. Ge, T. Zhang, and Z. Yuan, “AntiMalDroid: An efficient SVM-based malware detection framework for android,” Commun. Comput. Inf. Sci., vol. 243 CCIS, pp. 158–166, 2011.

[59] W. Enck, P. Gilbert, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth, “TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones,” Osdi ’10, vol. 49, pp. 1–6, 2010.

[60] L. Yan and H. Yin, “Droidscope: seamlessly reconstructing the os and dalvik semantic views for dynamic android malware analysis,” Proc. 21st USENIX Secur. Symp., p. 29, 2012.

[61] F. Wu, H. Narang, and D. Clarke, “An Overview of Mobile Malware and Solutions,” J. Comput. Commun., vol. 2, no. 2, pp. 8–17, 2014.

[62] T. Bläsing, L. Batyuk, A. D. Schmidt, S. A. Camtepe, and S. Albayrak, “An android application sandbox system for suspicious software detection,” Proc. 5th IEEE Int. Conf. Malicious Unwanted Software, Malware 2010, pp. 55–62, 2010.

An Improved Transformer for LLC Resonant Inverter for Induction Heating Applications

Amira ZOUAOUI-KHEMAKHEM
Laboratoire Systèmes Electriques
(L.S.E.-LR11ES15)
Ecole Nationale d'Ingénieurs de Tunis, Université Tunis El Manar
BP.37-1002-Tunis le Belvédère, Tunisie

Hamed belloumi
Laboratoire Systèmes Electriques
(L.S.E.-LR11ES15)
Ecole Nationale d'Ingénieurs de Tunis, Université Tunis El Manar
BP.37-1002-Tunis le Belvédère, Tunisie

Ferid KOURDA
Laboratoire Systèmes Electriques
(L.S.E.-LR11ES15)
Ecole Nationale d'Ingénieurs de Tunis, Université Tunis El Manar
BP.37-1002-Tunis le Belvédère, Tunisie

Abstract*—A new trend in power converters is to design a planar transformer that aims for low profile. However, at high frequency, the planar transformer AC losses become significant due to the proximity and skin effects. In this paper the most important factors for planar transformer (PT) design including winding loss, core loss, leakage inductance, and stray capacitance have individually been investigated. The tradeoffs among these factors have to be analysed to achieve optimal parameters. We will show a strategy to reduce losses in the primary coil of the transformer. The loss analysis of th PT was verified by the Finite Element Method (FEM) Simulations (**ANSOFT MAXWELL FIELD SIMULATOR 2D) **and could be utilized to optimize the transformer design procedure. Finally, the proposed PT has been integrated into an LLC Resonant Inverter for Induction Heating Application to test at high signal levels.

Keywords—planar transformer; induction heating; Finite Element Method; skin effect; diffusive representation

List of Symbols:

b : Width of the plate
C: Compensation capacitor
F_0: Resonant frequency
h : Height of the plate
I : Primary current
I_s: Secondary current
K: Coefficient of additional losses
L_s : Secondary inductance
L : Load inductance
m : Number of layers
N: Primary turns of the transformer
p : Diameter of the cross section of the conductor
P: Power
R: Load resistance
R_{ac}: AC resistance of the winding
R_{dc}: DC resistance of the winding
S : Area of the cross section of the conductor
S_{rec}: Rectangular section
V_{in}: Fundamental voltage
$\boldsymbol{V_b}$**:** Primary voltage
Φ : Phase shift between fundamental voltage and current
δ: Skin depth
ρ: Copper resistivity
ξ : Normalized value of the diameter of a strand
μ: Permeability

I. Introduction

High-frequency power transformers (HFPT) are largely used in many electrical and electronic designs. Switching at higher frequencies is an interesting option to reduce passive components and isolation transformers. They allow adapt different levels of voltage and provide electrical and galvanic isolation between both windings. However, there is not a standardized procedure for design transformers, and every manufacturer keeps confidential his method. Optimized design can save power, volume, weight and money for the manufacturer and for the customer. Most of researchers studied the theory of HFPT and its usage in power systems.

Due to eddy current effects, winding losses in high-frequency magnetic components are appreciably greater than those Calculated from the dc resistance of the winding. These losses must then be computed from the AC resistance of the winding which includes the DC resistance and a factor of excess loss related to skin and proximity effects [1].

We have shown in [2] two options to reduce high-frequency copper losses without significantly increasing the space occupied by these connections. Both options rely on the parallelization of copper plates. In the first, plates are interchanged to equalize the flux experienced by them. In the second, plates are interleaved to decrease the magnetic energy around the conductor.

PCB implementation of planar windings replicates the structure of conventional litz wires to reduce the AC losses. This structure consists of a set of fine trace electrically connected in parallel which transposed using changes of both trajectory and layer. An appropriate transposition strategy is essential to reduce the winding AC losses. However PCBs with a high number of vias could result expensive and not reliable [3].

A quick and simple method derived for obtaining the AC leakage impedance of a transformer, from the calculated values of DC winding resistance and DC leakage inductance.

It is rather unfortunate that calculating the DC leakage-inductance components accurately is a fairly lengthy process, although it mainly consists of substituting numbers in the given formulas [4].

[5] Explained a new method for core selection, based on the estimation of current density. The final algorithm presented allows to choice the magnetic core quicker, and designs a more efficient device. Furthermore, the method allows design a transformer with a broad range of frequency.

An improved high-frequency magnetic-integrated planar transformer was demonstrated in [6]. Compared with conventional planar integrated magnetic, the improved structure has better leakage inductance adjustment. The transformer was fabricated with two common EE shape magnetic cores along with magnetic insertions placed between the primary and secondary windings outside it. A three-dimensional finite-element model is used to investigate the eddy current losses and the adjustment of the leakage inductance.

A compact planar transformer with an improved winding configuration has been designed and fabricated in [11]. The meander-type design was engraved over the surface of a planar ferrite core, so that the primary and secondary windings can fit into the engraved design. This structure, was covered with another ferrite core, thus forming a compact planar transformer. Apical type AV polyimide film of 0.025 mm thickness is used between the windings to prevent short circuit. A coupling coefficient of >0.93 achieved, while the primary and secondary inductance was ~1600 nH. A coupling coefficient for air core transformer with the windings of identical geometry was >0.35 and the inductance of the windings near 160 nH. Fabricated transformer prototype has significantly a higher coupling coefficient and inductance of the windings in comparison with the previously developed planar transformer of the similar design.

In this paper, an HFPT for induction heating device is presented, thus its losses analysis. To achieve our HFPT, The choice and sizing of the primary and secondary of the transformer are necessary. Hence a transformer has been designed in order to adapt the voltage and current levels between the converter and the resonant tank.

An analytical method for the calculation of losses in HFPT followed by a simulation is perfomed. The simulation results show the good performance of proposed system and verify the mathematical analysis. Additionally, this new topology provides higher power rating on utility voltage at the load side.

II. Operation of the Proposed Inverter

Fig. 1 illustrates the power circuit of proposed induction heating diagram that employs an *LLC* resonant inverter configuration for induction-heating applications. The inverter consists of two switches with antiparallel diodes, a compensation capacitor (C), a series inductor (*Ls*), and an induction coil that comprises a series combination of a resistor (*R*eq) and an induction coil inductor (*L*coil).

You need to use half bridge capacitor inverter that eliminates the DC component and ensures a zero average current, and voltage in the primary of the transformer.

If you want to work at full power, you only have to operate the two half bridge, otherwise the half-bridge is sufficient for a power of 5kW.

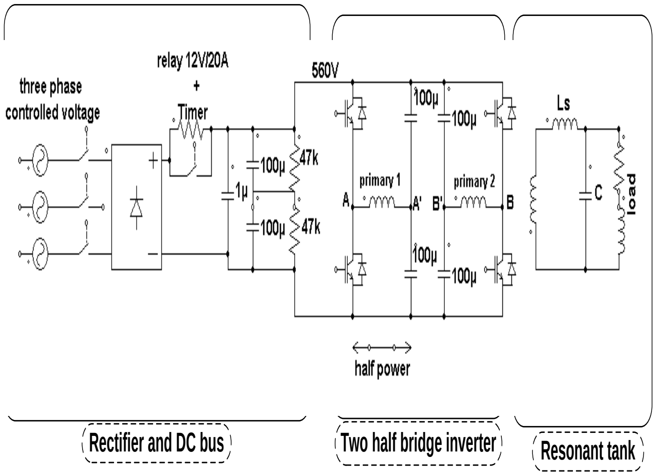


Fig. 1. Electrical diagram of the induction heating system

To ensure the compensation of the reactive, the commands must coincide with the zero crossing of the current Ic in the compensation capacity C.

Fig. 2 displays the current conduction through the inverter during different modes of the power transfer cycle.

During Mode 1, a positive voltage is applied to the gate of the high active power MOSFET. As a consequence, high MOSFET is turned-on. mosfet is conducting and transfer power to the resonant tank through the coupling transformer in our circuit. In Mode 2, the mosfets are transitioning, and the upper mosfet turns off slightly before the bottom one turns on. Here, current is conducted through the free-wheeling diode of the lower mosfet. In Mode 3, the lower mosfet turns on, and the resonant tank throws the power back through mosfet. In Mode 3, both mosfets are off during the transition, and the upper mosfet's free-wheeling diode conducts the current.

III. Losses Analysis of High Frequency Planar Transformer

When power converters operate at high frequency, the design difficulty for the transformer becomes much higher. The current density redistribution inside the winding wires (skin and proximity effects) strongly increases the copper losses.

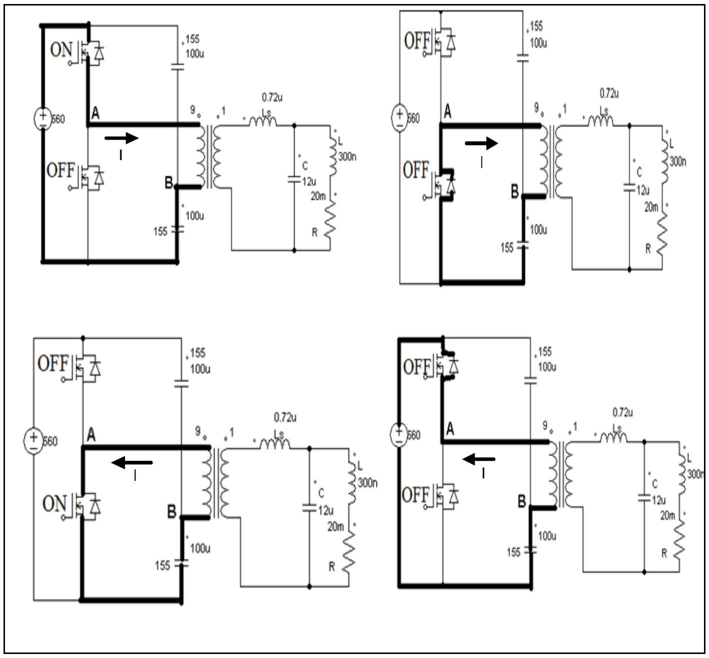


Fig. 2. Analysis of the inverter

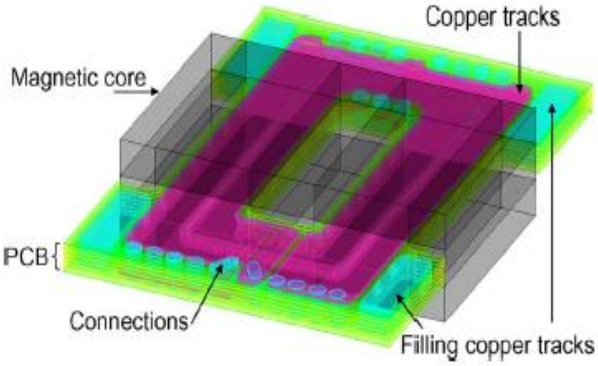


Fig. 3. Planar transformer

A. *Choice and sizing of the primary of the transformer*

Determining the number of primary turns:

Installation must be given to the resonance frequency defined by:

Resonance frequency

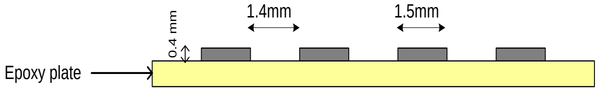
$$F_0 = \frac{1}{2\pi\sqrt{(Ls//L)C}} = 100khz \qquad (1)$$

It sets the Φ switching angle 20°

Data: V_b = 560V ; Vin= 504V ; L= 0.3 µH ; R = 20 mΩ ;

P = 10kW

The 14 turns will be divided over 7 epoxy plates as follows:



A coil consisting of four tracks

The thickness of the planar turn has to be two times the skin depth.

For 100 kHz : δ = 209 µm ; $\rho = \frac{1}{58\ 10^6}$

$R_{coil\ DC} = \rho\frac{L}{S} = 0.028\Omega$;

$R_{total} = 0.0072\Omega$;

For a length of 100mm: $R_{total} = 0.00072\Omega$

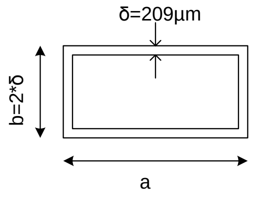
Coppers losses in a single turn in DC are given by:

$$P_{cu} = R_{total} * I^2 = 0.317W$$

Copper primary losses roundtrip in DC mode:

$$P_{copper} * 14 * 2 = 8.89W$$

Now to calculate the AC losses in a track



$$S_{AC} = S_{ext} - S_{int} = (a * b) - [(a - 2\delta)(b - 2\delta)]$$
$$= 2\delta(a + b - 2\delta)$$

Or b=2δ

$$S_{AC} = 2\delta * a = 2 * 0.209 * 1.5 = 0.6mm^2$$

We can therefore deduce that

$$S_{AC} = S_{DC} \text{ and } P_{copper\ AC} = P_{copper\ DC}$$

In the high frequency, the fact of increasing the thickness of a conductor does not reduce its apparent resistance, since the current density is confined in a skin thickness. A second solution to reduce losses is to connect two conductors in parallel.

In practice, the current density will be distributed as if both drivers were a single thicker conductor, so it must wait no significant improvement from the standpoint of losses.

The current distribution in the tracks is not uniform, and this is owed to the skin effect, the current is concentrated in the extremities. We note that when conductors are side-by side, the high concentration of the flux in the region causes high current densities in this small part of the conductor and, as a consequence, high copper losses.

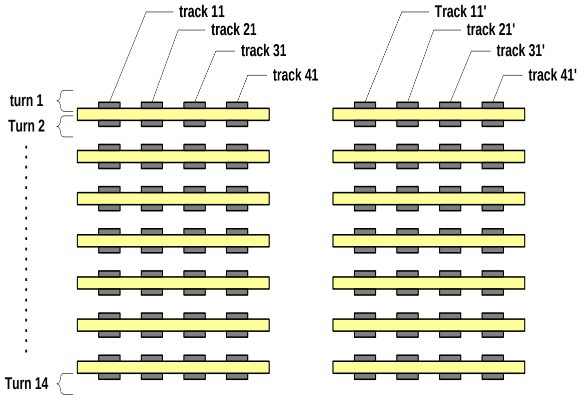


Fig. 4. 14 primary turns

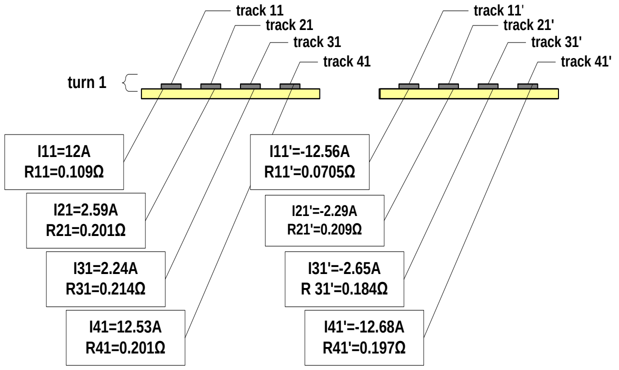


Fig. 5. Current distribution in a track

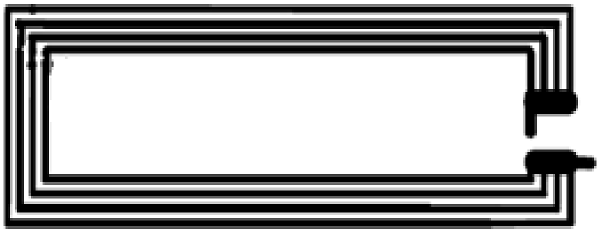


Fig. 6. 1 turn with 4 tracks

To mitigate skin effect and excessive losses due to the large rectangular tracks, another solution proposed by [7] is to use the planar Litz wire. Inspired by the Litz round wire, Litz planar structure has been proposed to reduce the strength of the alternative regime of a planar conductor.

It proposes a new form of planar litz wire. The proposed idea is to translate the turns with each other to force the current to pass around the turns and not be limited to the edges of the rectangular turns. The proposed solution is to divide each turn in four small tracks which leave together in parallel to form the required number of turns, then perform an intersection between the four tracks of each turn.

Resistance in each of the turns, deducted by finite element method (FEM), is nearest value. But we note that the resistance decreases significantly at the 14 turn. This can be explained by the proximity of the turn of the secondary winding traversed by a high current (fig.7).

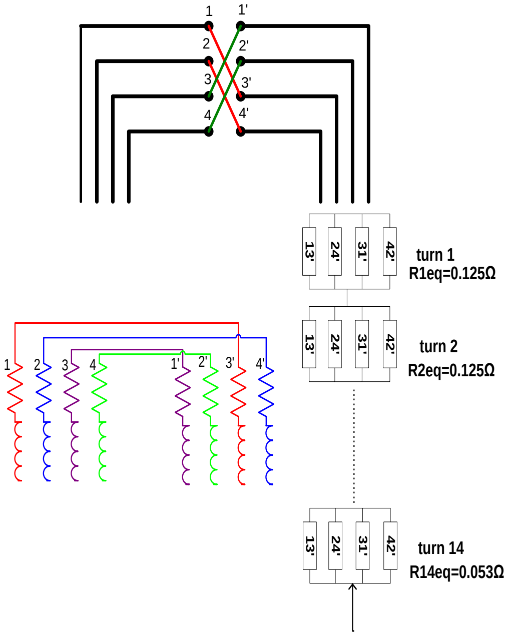


Fig. 7. Distribution of primary turns

Fig. 8. Design of a new turn on ISIS

B. The secondary of the transformer

The shape of the secondary single-turn essentially depends on the ferrite constituting the transformer. In fact the latter will consist of six blocks of ferrites.

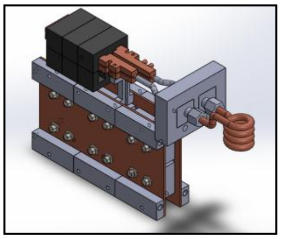
We must be adapted a solution to cool it when the shape and dimensions of the single-coil are fixed.

The idea is simple; it is to design an internal channel that will serve for the flow of coolant. The section of this channel must be maximum; however it must not interfere with the feasibility of the mono-spiral. For this we set the wall thickness to at least 3 mm. To optimize the shape of the single turn an analytical calculation followed by a simulation by the finite element method (FEM) was conducted.

To determine the optimum dimensions of the canalization of the single turn, we compared the total losses Joules effect for the different forms section.

IV. Analytical Method for the Calculation of Losses

Winding losses in transformers increase dramatically with high frequency due to eddy current effects. Eddy current losses, including skin- and proximity-effect losses, seriously impair the performance of transformers in high-frequency power conversion applications. Both the skin effect and the proximity effect cause the current density to be non-uniformly distributed in the cross section of the conductor and thus cause a higher winding resistance at higher frequency.

	
global schema of the transformer, capacity and the inductor	The single turn of the secondary of the transformer

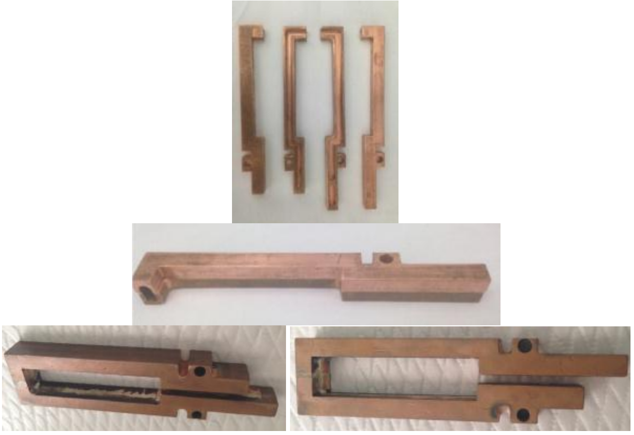


Fig. 9. The single turn of the secondary of the transformer

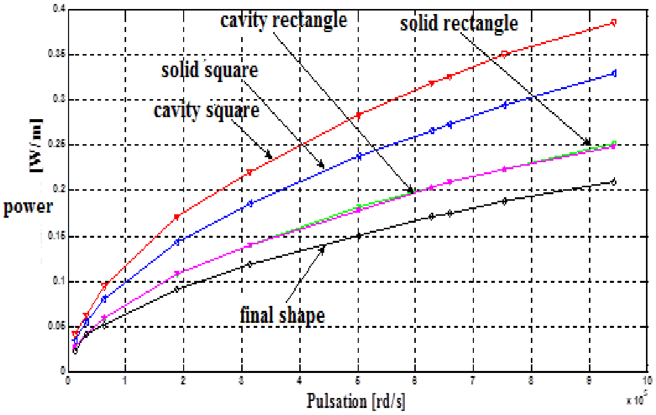


Fig. 10. Justification for the choice of the form of canalization of the single turn

The most commonly used equation that characterizes winding losses is Dowell's equation [4]–[8]

$$R_{ac,m} = R_{dc,m}\frac{\xi}{2}\left[\frac{sinh\xi+sin\xi}{cosh\xi-cos\xi} + (2m-1)^2\frac{sinh\xi-sin\xi}{cosh\xi+cos\xi}\right] \qquad (2)$$

$$\text{With } \xi = \frac{\sqrt{\pi}}{2}\frac{d}{\delta}$$

A. *Empirical formula*

Circular conductor forms:

For this particular form, the calculations are less complex and more accurate results.

The only parameter which is generally considered is the ratio K= $\frac{R_{ac}}{R_{dc}}$ which indicates a form of inadequate conductor when its value deviates too much from the unit.

Several empirical formulas have been proposed; that of Levasseur [9] is particularly simple and leads to errors of less than 2%:

$$K= \sqrt[6]{\frac{3}{4}^6 + (\frac{S}{p\delta})^{6}}+0.25$$

$$R_{ac} = R_{dc} * \sqrt[6]{\frac{3^6}{4} + (\frac{S}{p\delta})^6}+0.25 \qquad (3)$$

Rectangular form of the Conductor:

In this case, the calculations are much more complex and remain unclear because of the assumptions used about the distribution of the magnetic field; authors who have tried these calculations have often completed their work with delicate experiments.

According to estimates [10], a coefficient of 1.15 seems most probable correct errors due to non-compliance of the distribution of the field.

$$R_{ac} = 1.15 * (R_{dc} * \sqrt[6]{\frac{3^6}{4} + (\frac{S}{p\delta})^6}+0.25) \qquad (4)$$

TABLE I. COMPARATIVE TABLE OF LOSSES CALCULATED BY DIFFERENT METHODS

Losses calculated by the empirical formula	losses calculated by the corrected empirical formula	Losses calculated by (FEM)
177W	204 W	227 W
Error 21%	Error 9%	

Frequency	1kHz		10kHz		100kHz	
losses calculated by the corrected empirical formula	23 W	5%	66 W	6%	204 W	9%
Losses calculated by (FEM)	22 W		63 W		227 W	

As the frequency increases, the error increases.

B. *Bessel Formula*

The conductor is carrying a current I_S (rms) = 304A

The complex power due to the skin effect is given by:

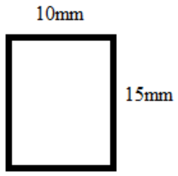
$$P_{skin}=\frac{\sqrt{jw\mu\sigma}}{2\pi r\sigma}\frac{I_0(r\sqrt{jw\mu\sigma})}{I_1(r\sqrt{jw\mu\sigma}}II^* \qquad (5)$$

$$R = Real\frac{\sqrt{jw\mu\sigma}}{2\pi r\sigma}\frac{I_0(r\sqrt{jw\mu\sigma})}{I_1(r\sqrt{jw\mu\sigma}} \qquad (6)$$

With: $I_0(r\sqrt{jw\mu\sigma})$ et $I_1(r\sqrt{jw\mu\sigma})$ are modified Bessel functions of the first kind of order zero and one.

r: radius of a circle
σ: resistivity

In the case of a single rectangular conductor, we first treated the rectangular shape to an equivalent circle keeping the same section.



$$S_{rec} = \pi r^2$$

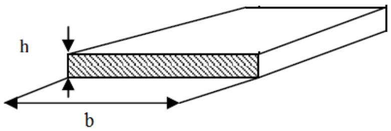
$$r = \sqrt{\frac{S_{rec}}{\pi}} = \sqrt{\frac{150}{\pi}} = 6.9mm$$

FEM formula	Analytical formula	corrected analytical formula
227W	178W	204W
Error	21%	9%

C. *Calculating the internal impedance owed to the skin effect and proximity of a rectangular conductor*

$$R_{recP} = \frac{1}{\sigma b\delta} * \frac{\sinh\left(\frac{2h}{\delta}\right)+\sin\left(\frac{2h}{\delta}\right)}{\cosh\left(\frac{2h}{\delta}\right)-\cos\left(\frac{2h}{\delta}\right)} \quad (7)$$

$$L_{recP} = \frac{1}{\delta\omega\sigma b} \frac{\sinh\left(\frac{2h}{\delta}\right)-\sin\left(\frac{2h}{\delta}\right)}{\cosh\left(\frac{2h}{\delta}\right)-\cos\left(\frac{2h}{\delta}\right)} \quad (8)$$



With b: width of the plate h: height of the plate

Knowing that the width should be strictly greater than the height that this formula is true so we take our case b = 19mm and h = 12mm.

TABLE II. COMPARISON BETWEEN FEM AND THE ANALYTICAL METHOD

	FEM	Analytical
Losses	461	406
Resistance	0.0049	0.0043
Inductance	2.135^{E-007}	6.9109e-009

The error between the two powers is 11%.

We compared the analytical expressions obtained from different methods. In fact the first case (empirical calculation) has an acceptable convergence of results by finite element method (FEM) and analytical losses. This convergence is only true when adding a correction factor of 1.15. This coefficient is valid for the second case too which replaces a rectangular conductor with a circular conductor with the same cross section and the error is minimal. As can be seen, the theoretical loss calculations are in agreement with the results obtained with the Finite Element simulations. As the two dimensional finite element analysis allows the estimating of the fringing effect at the end of the conductors, the simulated losses may exceed the calculated transformer losses a little.

To study the internal current constraint in our transformer during operation, we have been led to make a digital model.

Achieving diffusive admittance was identified. Diffusive symbol approached is shown in Fig.11.

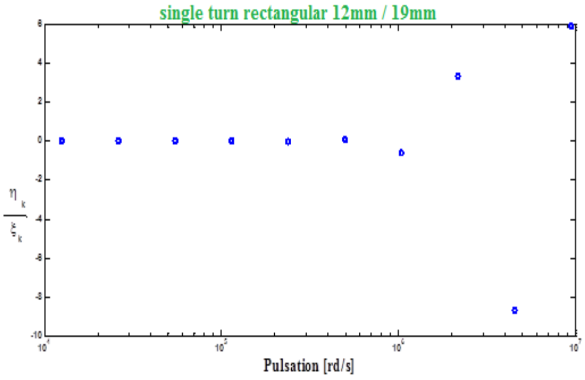


Fig. 11. Diffusive approached symbol of admittance $\frac{H(P)}{l}$

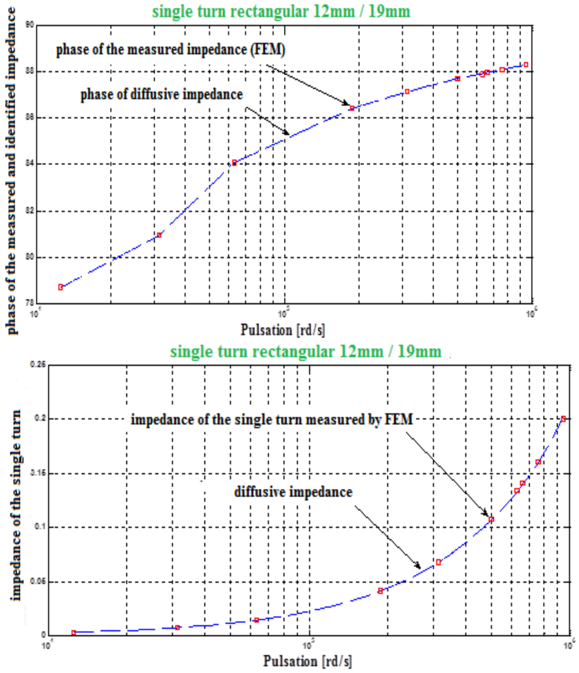


Fig. 12. Comparison of the estimated admittance by FEM and identified

In Fig.12, the frequency response of the admittance thus identified is compared to the response obtained using the parameters determined by FEM simulation. The agreement is quite satisfactory.

V. SIMULATION RESULTS

To validate the loss estimation mentioned above, the detailed losses in a high frequency planar transformer are studied using the Finite Element Method(FEM) in detail.

Finite Element Method (FEM) can be used for calculating transformer characteristics and stray parameters.

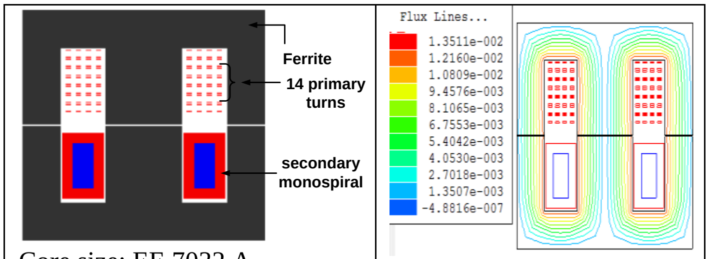
2-D simulations perform an analysis on the transformer based on its symmetry axes.

Determining losses by Joule effect is using the Maxwell computer through the formula of losses:

$$P = \frac{1}{2\sigma}\int_V J.J^*dV \quad (15)$$

To realize the finite element modeling, we impose for the primary winding a current of 30A peak for the four parallel strands that form our proposed Litz wire, and the secondary we impose a current of 432A peak for the single turn. This finite element modeling of the copper losses was performed at a frequency of 100 kHz.

The parameters of the transformer studied are shown in Fig 13. For this planar transformer, the red plates on the top in the model represent the 14 primary turns and the red one in the other side of the transformer represent the secondary turn.



Core size: EE 7032 A
Material: ferrite
Primary turns: 14
Secondary turns: 1
Effective core: 2049 mm^2

Magnetic field flux Distribution at F=100 kHz

Fig. 13. Shows the magnetic flux distribution in frontal model of the planar transformer

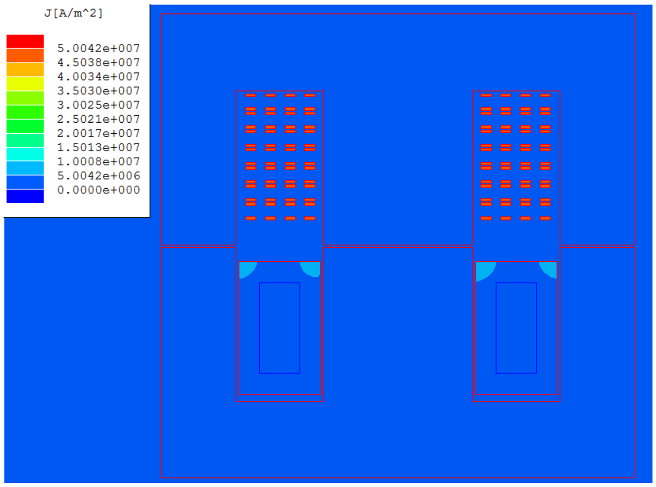


Fig. 14. FEM simulation at F=100 Hz

For each frequency of 100Hz and 100kHz, we simulate the contour plot of current density.

The loss estimation takes account of switching-type waveforms encountered in power supplies, inclusive of high frequency skin and proximity effects in windings.

This planar transformer is used in a half bridge converter and its associated voltage and current waveforms are shown in Fig.16. The voltage across the primary winding is a square wave with peak values of-560/2 V and 560/2 V. The current through the primary winding is also a square wave with peak values of -38A and 38A.

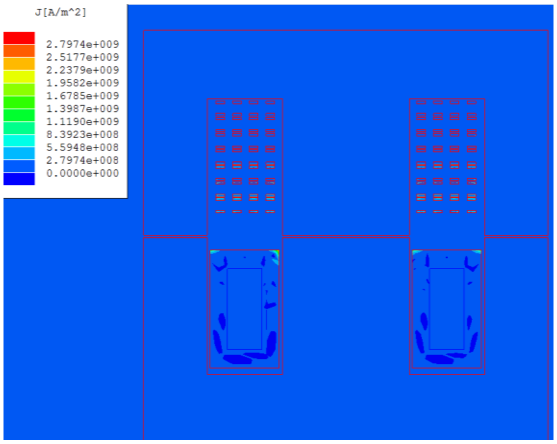
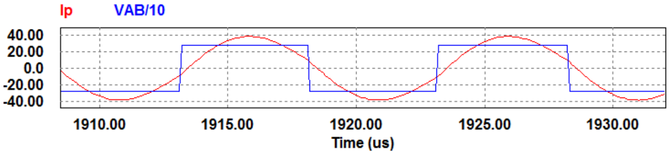


Fig. 15. FEM simulation at F=100 kHz

(b) Half bridge converter for induction heating



(a) key waveforms

Fig. 16. The half bridge converter

VI. Experimental Results

The experimental results presented are for a 5 kW furnace.

Fig. 16 displays the voltage waveforms V_{DS}, of the high MOS and the resonant current. When the MOS is turned-on, the current is initially negative, evidence that circulates in the internal diode of the high MOS. Meanwhile, the MOS is turned-on at zero voltage. The current becomes positive until the order of his blocking, reducing the voltage V_{ab} =+ E between its terminals with no overvoltage thanks to CALCS. The test was carried under an input voltage E = 400V, and a current IpM = 12A.

The temperature of the work piece is measured by an optical pyrometer and the Curie temperature at 875 ° C is achieved as desired.

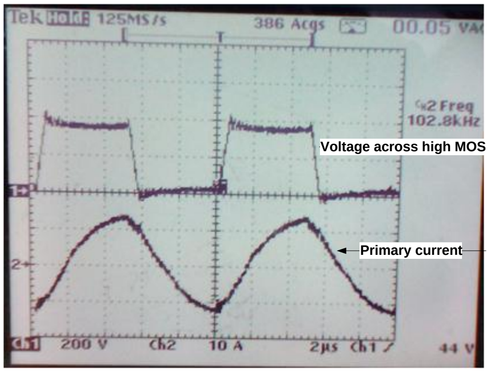


Fig. 17. Diagram viewing the output voltage of the MOS high and the primary current

As a conclusion, these waveforms match up with the theoretical expected ones, and verify the proper operation of the proposed converter. Fig. 18 shows a photo of the implemented hardware prototype.

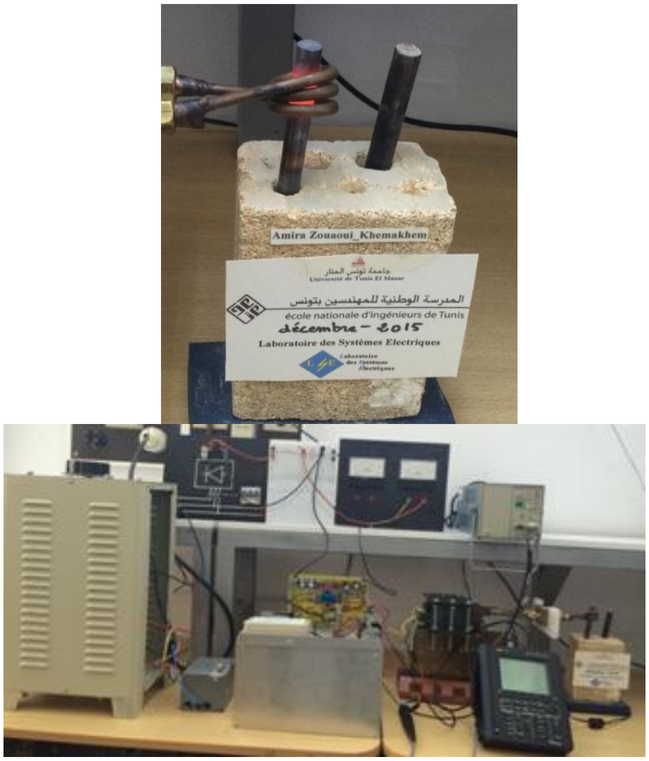

Amira Zouaoui_Khemakhem
جامعة تونس المنار
Université de Tunis El Manar
المدرسة الوطنية للمهندسين بتونس
école nationale d'ingénieurs de Tunis
décembre - 2015
Laboratoire des Systèmes Electriques

Fig. 18. experimental model

VII. Conclusion

Planar transformers are more and more used in electronic power structures since they have major interests in congestion, performance and industrial production method. In the first part of this paper, we have sought to reduce the joule losses planar transformer due to the skin effect and proximity effect.

We have presented all the steps for building a high-frequency transformer model. Some parameters such as resistance and inductance of the single turn are affected by the variation of the frequency (the effects of skin and proximity) as well as the variation of the shape of the conductor. Such a phenomenon can model by adopting the diffusive representation. The different model parameters are estimated by the finite element method (FEM). It has been shown that it is possible to use with certain circumstances analytical methods to provide an order of magnitude of some parameters.

A 5-kW prototype has been designed and implemented to validate the analytical and simulation results. The experimental measurements validate the feasibility of the proposed converter.

Acknowledgment

This work was supported by the Tunisian Ministry of Higher Education and Research under Grant LSE-ENIT-LR11ES15

References

[1] F. Tourkhani, P. Viarouge « Accurate Analytical Model of Winding Losses in Round Litz Wire Windings », IEEE transactions on magnetics, vol. 37, no. 1, january 2001

[2] B. Cougo, J. W. Kolar, G. Ortiz, « Strategies to Reduce Copper Losses in Connections of Medium-Frequency High-Current Converters », 7th Annual Conference of the IEEE Industrial Electronics Society (IECON 2011), Melbourne, Australia, November 7-10, 2011.

[3] I. Lope, J. Acero, J. Serrano, C. Carretero, R. Alonso and J.M. Burdio, « Minimization of Vias in PCB Implementations of Planar Coils with Litz-Wire Structure », IEEE 2015

[4] P.L Dowell, « Effects Of Eddy Currents In Transformer Windings » Proc Iee Vol 113, NO 8, August 1966

[5] Xavier Duran Reus, Foo Javier Quiros And Daniel Montesinos Miracle, « Design Of A High Frequency Transformer For An Induction Heating System » SSD2014

[6] Wayne Water And Junwei Lu, « Improved High Frequency Planar Transformer For Line Level Control (LLC) Resonant Converters », IEEE Magnetics Lettters, Volume 4 (2013)

[7] B. Cougo, J. W. Kolar, G. Ortiz « Strategies to Reduce Copper Losses in Connections of Medium-Frequency High-Current Converters » Proceedings of the 37th Annual Conference of the IEEE Industrial Electronics Society (IECON 2011), Melbourne, Australia, November 7-10, 2011.

[8] A.Ducluzaux, « Extra Losses Caused In High Current Conductors By Skin And Proximity Effects », Cahier Technique N° 83, edition 1977

[9] K. Ben Smida M. Elleuch, « Analyse Des Effets De Peau Et De Proximite Sur L'impedance Interne D'un Enroulement Statorique »

[10] Alexander W. Barr , « Calculation of Frequency-Dependent Impedance for Conductors of Rectangular Cross Section » AMP Journal of Technology Vol. 1 November, 1991

[11] Djuric S.M, Stojanovic G.M. "A Compact Planar Transformer With an Improved Winding Configuration" , IEEE Transactions on Magnetics (Volume:50 , Issue: 11) Nov. 2014

[12] Ignacio Lope, Claudio Carretero, Jesus Acero, Rafael Alonso, and Jose M. Burdio "Frequency-Dependent Resistance of Planar Coils in Printed Circuit Board With Litz Structure" IEEE transactions on magnetics, vol. 50, no. 12, december 2014

[13] Ignacio Lope, jesus Acero, Claudio Carretero, "Analysis and Optimization of the Efficiency of Induction Heating Applications with Litz-Wire Planar and Solenoidal Coils" IEEE transactions on power electronics, September 2015

An Empirical Investigation of Predicting Fault Count, Fix Cost and Effort Using Software Metrics

Raed Shatnawi
Software Engineering Department,
Jordan University of Science and Technology,
Irbid, Jordan

Wei Li
Computer Science Department,
University of Alabama in Huntsville,
Huntsville, AL, USA

***Abstract*—Software fault prediction is important in software engineering field. Fault prediction helps engineers manage their efforts by identifying the most complex parts of the software where errors concentrate. Researchers usually study the fault-proneness in modules because most modules have zero faults, and a minority have the most faults in a system. In this study, we present methods and models for the prediction of fault-count, fault-fix cost, and fault-fix effort and compare the effectiveness of different prediction models. This research proposes using a set of procedural metrics to predict three fault measures: fault count, fix cost and fix effort. Five regression models are used to predict the three fault measures. The study reports on three data sets published by NASA. The models for each fault are evaluated using the Root Mean Square Error. A comparison amongst fault measures is conducted using the Relative Absolute Error. The models show promising results to provide a practical guide to help software engineers in allocating resources during software testing and maintenance. The cost fix models show equal or better performance than fault count and effort models.**

Keywords—Software metrics; fault prediction; fix cost; fix effort; regression analysis

I. Introduction

Predicting faults in modules is important to assess software quality and to direct software engineers' effort to spend more time on more trouble-prone modules. Software metrics are surrogates for fault measures such as fault-proneness, fault count, fault-fix cost, and effort. Software metrics measure the complexity of software and can be used to identify the faulty modules using statistical and machine-learning techniques. These techniques can be used to build prediction models such as fault count, fix cost, and fix effort to predict which modules are likely to have these problems. Software systems are becoming larger and larger and contain thousands of modules that are investigated in testing and maintenance phases. However, the cost of testing and maintenance are growing with the size of systems. This growing trend leads to either very costly system or compromised quality. Software engineers can use prediction models to prioritize modules to focus the testing and maintenance activities on the modules that are either have more faults, more costly to fix or demand more efforts to fix. Hence, detecting and ranking faulty modules is an important engineering task for improving system quality and reducing cost. There are usually two measures of module quality: fault count or fault-proneness. In most systems, a small number of modules have faults and the majority of modules have zero faults. Researchers use fault-proneness by using binary coding of modules (zero for no faults and one if there are faults in a module) to build prediction models that are usually easy to interpret [1][2][3][4][5][6][7]. However, the binary coding does not explore all information available about faults. Fault count is an indicator of quality in a module but may not provide enough information about the fix cost or effort. Therefore, regression and machine-learning models are used to identify complex modules by considering fault count, fix cost and effort. In this paper, five regression and machine-learning techniques are used to predict the three fault measures. Twenty procedural metrics used as independent variables in the prediction models. The models are trained and tested on three data sets provided by NASA. Overall, fifteen models were built for each data set using 10-fold cross-validation. The results for the three fault measures have shown similar results, but the cost-fix models are slightly better. These models can help in allocating resources for software testing and maintenance. The results of the models are used to rank the modules based on the fault measures, and the results are promising and commensurate with previous works [8][9].The performance of the three fault measures is compared to find the best ranking. The results show similar results for the three measures with some advantage for fault count and fix cost over fix effort.

The rest of the paper is organized as follows: related work to the three fault measures are discussed in Section 2. In section 3, the study design is discussed which includes a description of the dependent, independent variables and regression models used in this paper. The data analysis is presented in Section 4, which also evaluates the predictions of the fault measures. Validity threats to the study are discussed in Section 5. The study is concluded in Section 6.

II. Related Work

Fault prediction has been discovered in many previous research in two major themes: fault-proneness and fault count. Studies on fault proneness categorized software classes into groups. Usually, classes are divided into two groups: faulty classes that had one or more faults in the current release, and non-faulty classes. Software metrics have shown significant relations with fault-proneness using many machine learning and statistical techniques [1][2][3][5][6][7][10]. Many research studies used the NASA fault data to build fault-proneness models. For example, Pai and Dugan [11] conducted a Bayesian analysis of fault count and fault proneness. The study produced statistical significant results using linear, Poison, and binomial logistic regression. The modeling of the results have shown 20:60 relationship when classes were ranked using module-fault order. Catal and Diri [12] used the NASA data sets to predict fault-prone modules and proposed an artificial immune system (semi-supervised approach) that uses a recent algorithm called YATSI. Gondra [13] also used the NASA's Metrics Data Program data to build prediction models of fault-proneness of modules using two machine l++-earning techniques: Artificial Neural Networks (ANN) and Support Vector Machines (SVM). Zheng [14] used four datasets from NASA projects to compare the effect of cost-sensitive boosting algorithms on the performance of neural networks for predicting fault-prone parts. In other studies on fault measures , Ohlsson and Alberg [15] noted that in commercial products, the average cost of fixing an operational fault was $7000. Biyani and Santhanam [16] found correlation between the number of faults found in development and the number of faults remaining in operation. Ostrand et al. [17] developed a negative binomial regression model to predict the number of faults in each file for many consecutive releases of a software. Khoshgoftaar and Gao [9] used two statistical models: Poisson regression model and the zero-inflated Poisson to predict fault count in two industrial case studies. The zero inflated model showed better performance than poison regression model. Other researchers focused on other fault measures such as fix cost and effort. For instance, [18] used the KC1 data to build faults fix cost using Neural Networks. Panjer [19] proposed to build machine-learning models to predict fault-fix time. (Khoshgoftaar and Gao [9] proposed to use a program module-order models to explore the relationship between %modules and %faults as a more practical model that is based on the predictions resulting from machine learning models. Khoshgoftaar et al. found that 80% of faults are found in the top 20% of files when ordered by faults predicted by models [9]. In a recent study, Hamill and Goseva-Popstojanova [20] studied the relationship between faults and failure of 21 large-scale software components extracted from a safety-critical NASA mission. However, the study focused more on fault types.

Fault prediction models are reported frequently in previous works as reported in surveys on software fault prediction [21] [22]. This study provides an exploration of the added dimension for the relationships between software metrics and fault measures such as fix cost and fix effort. In addition, the module-order models proposed in Briand et al. and Khoshgoftaar and Gao [8][9] are used to prioritize modules according to models predicting fix cost and fix effort.

III. Study Design

Fault data are becoming more available on many repositories such as PROMISE [23], Eclipse Bug Data [24], and NASA fault data [25]. The NASA data provides more details on the costs and efforts of fixing software faults, which are the focus of this research. Three data sets, KC1, KC3 and KC4 report the cost of fault fixes in terms of person hours and effort measured in Source Line of Code (SLOC) modified to accomplish the fix. Table 1 shows a summary of the three data sets. All these projects were built in similar software development environments and analyzed by the same set of software product metrics. These data sets are available publicly and other researchers can repeat and verify this study's results. The MDP is funded by NASA's Software Independent Verification & Validation (IV&V) facility. These systems met the requirements to support NASA mission [26].

TABLE I. A Summary of Data Sets

Data set	Description	Language	#instances	#faulty instances	%faulty instances
KC1	is a system implementing storage management for receiving and processing ground data	C++	2107	278	13%
KC3	Storage management for ground data	Java	458	25	5.5%
KC4	a ground-based subscription server	Perl	125	60	48%

A. Research Questions

Given the information available on fault count, fix cost and fix effort, this research aims to find answers for the following research questions.

RQ1*: Can software metrics predict fault count?*

Fault count is defined as the number of faults fixed in a module. This question is already answered in previous research as explained in more details in the related work section. However, this study adds the evaluation of faults prediction using other machine learning techniques. Fault prediction is important to assess the complexity of software modules. Five prediction models are conducted to answer this question. The results of the prediction models are used to rank the modules by sorting according to the predicted fault count. The models can be used to allocate resources efficiently to identify for instance the 20% modules that have the most faults.

RQ2: Can software metrics predict fix cost as measured in man-hours?

Fix cost is defined as the total number of hours the developers spent to fix all faults in a module. For each module, the cost of fault fixes are aggregated. The fix cost in hours is an indicator of the complexity of code. A positive relationship is expected between the studied metrics and fix cost, i.e., more complex modules cost more than less complex modules. To put the cost prediction models in practical use, the results of prediction models are used to sort the modules by the predicted fix cost. The models can be used to allocate resources efficiently to identify for instance the 20% modules that have the most fix cost.

RQ3: Can software metrics predict fix effort as measured in SLOC modified?

Fix effort is defined as the actual number of SLOC added or modified to fix all faults in a module. In this study, the aggregation of all modified SLOC for a particular module is used to investigate the relationship between the fix effort and the complexity of modules. To put the effort prediction in practical use, the results of the prediction models are used to sort the modules by the predicted fix effort. The models can be used to allocate resources efficiently to identify for instance the 20% modules that need the most fix effort.

The results of the three quality predictions are compared using the relative absolute error to find which models are better.

B. *Dependent variables*

NASA MDP has many projects, but only three of these projects have details on fault fixes, cost and effort. For each module, the number of faults (fault content), the total fix hours, and the total SLOC changed or added are aggregated. Table 2 provides a summary of the fault measures used in this study. The scale for fix cost and effort are larger than the fault count. The scale has effect on the performance measures used in evaluating the prediction models and the comparison should be based on unbiased performance measures. Relative absolute error is used to evaluate models besides the root mean square error.

C. *Independent variables - software metrics*

The software metrics under investigation are procedural metrics for three systems collected by NASA MDP. The metrics collection were applied to the lowest level functional unit, procedures. The data were stored in a structured format. For example, a file named KC1_static_defect_data.csv, keeps all information related to faults including severity, priority, fix hours, the actual number of SLOC changed or added. Another file includes all the static metrics for each module and recognized using a unique variable, MODULE_ID. These files are then combined together into one file using the MODULE_ID, which is an identifier of module records in all files.

The NASA MDP data needs preprocessing as reported in [27]. Therefore, we use only those metrics that were reported by [27] which had 21 metrics as reported in Table 3. The LOC_BLANK metric is deleted because it is not meaningful and its interpretation is not clear. These metrics were originally proposed in [28][29]. The McCabe and Halstead measures are module-based where a module is the smallest unit of functionality. McCabe argued that code with complicated pathways are more error prone. Halstead considered the code readability as indicator of fault proneness. Halstead metrics measure software complexity by counting the number of concepts in a module [26].

TABLE II. DESCRIPTIVE STATISTICS FOR THE THREE FAULT MEASURES (FAULT COUNT, FIX HOURS, SLOC MODIFIED)

Fault count	**Min**	**Max**	**Mean**	**stdev**	**Total**
KC1	0	11	0.30	0.991	631
KC3	0	3	0.114	0.50	52
KC4	0	23	2	3.60	248
Fix cost	**Min**	**Max**	**Mean**	**stdev**	**Total**
KC1	0	397	5.99	26.7	12629
KC3	0	190	6.62	29.365	3032
KC4	0	498	28.6	62.43	3548
Fix effort	**Min**	**Max**	**Mean**	**stdev**	**Total**
KC1	0	1016	14.57	57.59	30713
KC3	0	512	7.63	44.00	3496
KC4	0	467	19.24	62.70	7176

D. *Regression Models*

We propose to use a set of data mining techniques to predict the value of a numerical variable (e.g., fix cost) by building a model based on many software metrics. This research uses the following regression techniques to predict fault count, fix cost and fix effort.

Regression Decision Trees (M5P): Decision tree is used to build regression models in the form of a tree structure using the M5 algorithm [30]. The algorithm constructs a decision tree for regression different from classification by using Standard Deviation Reduction instead of Information Gain. A dataset is continuously partitioned into smaller subsets while the standard deviation is larger than zero.

Multiple Linear regression (MLR): Multiple linear regression (MLR) is a well-known statistical technique used to model the linear relationship between a count variable and many independent variables. MLR is based on calculating ordinary least squares (OLS), the model is fit such that the differences between actual and predicted instances are minimized.

k Nearest Neighbors (kNN): The kNN algorithm is an instance-based method that is not used to build a model from training data; rather, it keeps the training instances with the intention of analyzing future instances. The kNN algorithm searches the training instances to find the closest instances to a new unknown instance to be analyzed. The search starts by finding the distance with all other instances using the Euclidean Distance. The kNN algorithm selects the average of the closest group of k objects in the training set [31].

TABLE III. SOFTWARE METRICS USED IN THE EMPIRICAL WORK

Metrics	description or formula
LOC_CODE_AND_COMMENT:	The number of lines which contain both code and comment in a module
LOC_COMMENTS	The number of lines of comments in a module
LOC_EXECUTABLE	The number of lines of executable code for a module (not blank or comment)
LOC_TOTAL	The total number of lines for a given module
BRANCH_COUNT	Branch count metrics
CYCLOMATIC_COMPLEXITY:	The cyclomatic complexity of a module $v(G) = e - n + 2$
DESIGN_COMPLEXITY:iv(G)	The design complexity of a module
ESSENTIAL_COMPLEXITY:ev(G)	The essential complexity of a module
NUM_OPERATORS:N1	The number of operators contained in a module
NUM_OPERANDS:N2	The number of operands contained in a module
NUM_UNIQUE_OPERATORS:µ1	The number of unique operators contained in a module
NUM_UNIQUE_OPERANDS:µ2	The number of unique operands contained in a module

HALSTEAD_CONTENT:μ	The halstead length content of a module μ = μ1 + μ2
HALSTEAD_LENGTH:N 2	The halstead length metric of a module N = N1 + N
HALSTEAD_LEVEL:L	The halstead level metric of a module L = (2∗μ2)/μ1∗N2
HALSTEAD_DIFFICULTY:D	The halstead difficulty metric of a module D = 1/L
HALSTEAD_VOLUME:V	The halstead volume metric of a module V = N ∗ log2(μ1 + μ2)
HALSTEAD_EFFORT:E	The halstead effort metric of a module E = V/L
HALSTEAD_PROG_TIME: T	The halstead programming time metric of a module T = E/18
HALSTEAD_ERROR_EST: B	The halstead error estimate metric of a module B = $E^{2/3}$/1000

Multi-layer Perceptron - Backpropagation algorithm: The multi-layer perceptron (MLPRegressor) is similar to the organization of the brain neurons. Artificial neurons are arranged in layers (i.e., input layer, hidden layers and output layer). Connections between the neurons provide the network with the ability to learn patterns. In MLP, each neuron in the hidden layer uses a combination of weighted outputs of the neurons from the previous layer. In the final hidden layer, neurons are combined to produce an output, which is compared to the correct output and the difference between the two values (the error) is fed back to update the network [13].

Support Vector Machine (SMOreg): SMOreg implements the support vector machine for regression. SMOreg is more complicated to be taken into consideration than the classification version. However, both aim to minimize error, i.e., individualizing the hyperplane which maximizes the margin while error is tolerated [32].

E. *Regression performance evaluation*

The models are trained and tested using 10-fold cross-validation, in which data is partitioned into ten equal sample sizes. Nine partitions are used for training while the last partition is used for testing. This process is repeated ten times to use all partitions in testing. The performance of regression models is usually evaluated using the Root Mean Squared Error (RMSE) as defined in Eq. (1). RMSE is frequently used to measure the difference between predicted and actual values. RMSE is calculated as follows.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \qquad (1)$$

In this research the dependent variables have different units and to be able to compare models on different units, the Relative Absolute Error (RAE) is used as defined in Eq. (2). [32]. RAE is calculated as follows.

$$\text{RAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{\sum_{i=1}^{n}|\bar{a} - a_i|} \qquad (2)$$

In both measures, a is the actual value, p is the predicted value, and $\bar{a}$ is the mathematical mean.

IV. Data Analysis

In the following, the evaluation of the prediction performance for fault measures are reported using RMSE and then compared using RAE.

A. *Evaluation of fault count prediction*

Five prediction models are built for fault count using twenty metrics under investigation. The performance of fault prediction is calculated and summarized in Table 4. The results of the five models do not differ from each other when compared within any data set. However, the LR models look better in two data sets, while KNN models are also better in two data sets as marked in bold. However, the differences in the performance among the models are not enough to provide ranking of the machine learning techniques. The MLP can be considered the worst in performance among all.

TABLE IV. Fault Count Regression Models

Fault Count	LR	kNN	M5P	SMOreg	MLPRegressor
KC1	**0.90**	0.92	0.93	1.00	1.06
KC3	**0.46**	**0.46**	0.48	0.47	0.63
KC4	3.17	**2.60**	2.78	3.16	3.69

To put models in practice, the results of the models are depicted using Alberg diagrams as proposed in [15]. In Figure 1, modules are sorted in decreasing order by the predicted faults. The plot shows the percentage of modules (x-axis) against the percentage of actual faults after sorting the instances. Figure 1 shows the results of fault count prediction in KC1. These results are taken from running the models in the 10-fold cross-validation. The figure can be used as follows, for example at X=20 the value of the curve is 60, which means 20% of modules (369 modules) with highest predicted fault count constitute of 60% of faults. It can be noticed that the top 30% of modules has 70% of actual faults. This behavior is similar in all models.

We also plot the same graph for KC3 and KC4 prediction models in Figure 2 and 3. In Figure 2, we observe similar results for the top 20% modules, i.e., about 60% of faults are found in the top 20% of modules in all prediction models. In Figure 3, we observe similar results for KC4 data in kNN model. Other models show 20:50 relationship, i.e., 50% of faults are found in the top 20 modules. We can conclude that software metrics can be used to predict fault count and models can be used in practice to rank modules based on predicted fault count. Therefore, RQ1 is answered in this research. When planning for quality inspection during the software development process, we can make a trade-off between the resources spent on inspection and the effectiveness of inspections [8]. The prediction models can be used to put the modules in a priority list for more investigation such as testing and maintenance. We can use the graph in Figure 1 to determine the percentage of faults that are expected in the system by inspecting a certain percentage of the system modules. For example, the top 20% modules can be investigated first if allocated resources are only available for investigating such number of modules.

The graphs in Figures 1-3 have shown similar behavior to works in [33][8][11]. For instance, Briand et al. [8] found that the first 20% of classes have 52% of faults in the system. They also suggested that such curves can be used in practice if they appear to be constant across projects. Software managers can use fault prediction models to allocate more resources on the parts of the code that were predicted to be more fault-prone [5][34].

B. *Evaluation of fix-cost prediction*

We repeated the same experiment to predict fix cost using all metrics and the results are shown in Table 5. We notice no significant differences among the models except MLP, which is again the worst modelling technique. M5P regression trees can be considered the best among all models, while others have almost equal performances.

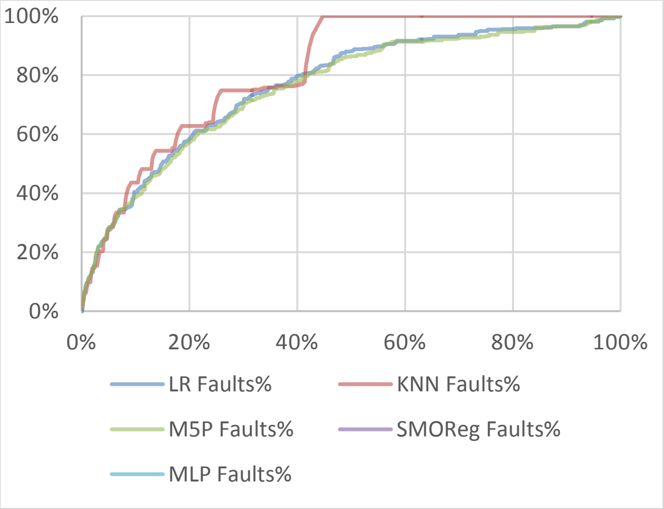


Fig. 1. Alberg diagram for five prediction models of KC1

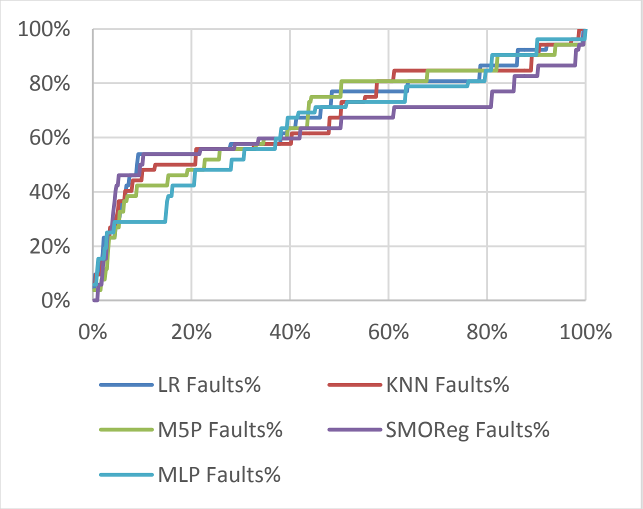


Fig. 2. Alberg diagram for five prediction models of KC3

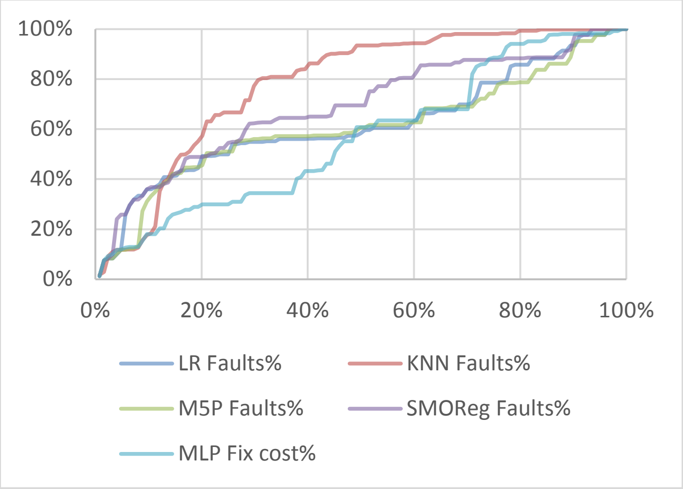


Fig. 3. Alberg diagram for five prediction models of KC4

TABLE V. FIX-COST REGRESSION MODELS

Fix cost	LR	kNN	M5P	SMOreg	MLPRegressor
KC1	24.70	25.05	24.70	25.71	29.70
KC3	25.98	26.3	27.71	24.47	36.38
KC4	59.88	50.83	50.00	55.47	73.21

The fix cost can be used in practice to order modules based on cost prediction. We plot the percentage of modules (x-axis) and the percentage of actual costs after sorting the instances in decreasing order by the predicted fix cost. Figure 4 shows the results of the five prediction models for fix-cost prediction in KC1. The figure can be used, at X=20 the value of the curve is 60% in three models whereas in two models (LR and MLP) is about 50%. This result means 20% of modules (369 modules) with highest predicted fix cost incurred 60% of the spent person hours on fixing cost. It can be noticed that the top 30% of modules ordered by the prediction model has 60-70% of actual fix cost.

We plot the Module-Cost graph for KC3 and KC4 in Figure 5 and 6.

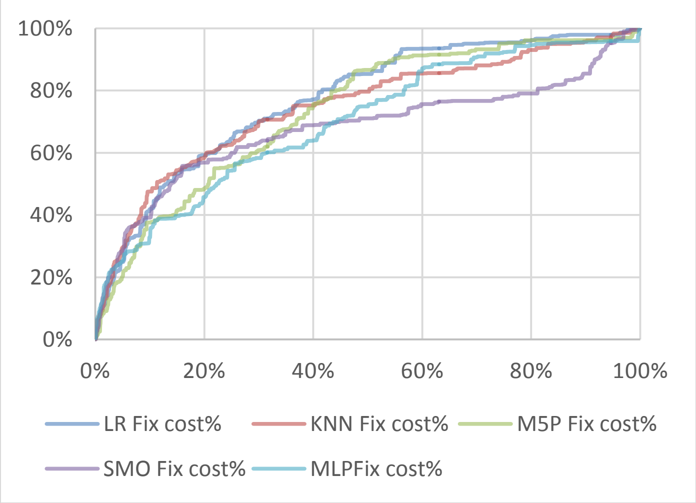


Fig. 4. Alberg diagram for five prediction models of KC1

The graphs show a 10:60 relationship, i.e., 10% of modules incurred 60% of the fix cost in both KC3 and KC4 in four models except the MLP prediction models which shows 20:40 relationship. These results are better than the models in KC1. In addition, the use of fix cost seems more efficient than the use of fault count in two data sets: KC3 and KC4, which show 20:60 relationship. Therefore, RQ2 is answered in this research as well. Fix cost can be predicted using software metrics and models can be used in practice to rank modules based on predicted fix cost. Fix cost can be used to allocate resources in software testing and maintenance activities.

C. Evaluation of fix effort prediction

The SLOC modified in a module is also studied as a fault measure and the results are presented in Table 6. The results are not conclusive in identifying the best model. The MLP models are again the least in performance among all. We plot the percentage of modules (x-axis) and the percentage of actual SLOC modified to fix faults in each module after sorting modules in decreasing order by the predicted effort as shown in Figure 7,8 , and 9. The results of the five prediction models do not show consistent results in all data sets. Almost all models show 20:60 relationship in KC1, but are different in KC3 and KC4 for different models. However, the results of the models on KC4 are similar to the models in KC1. While the models obtained from KC3 do not show promising results. These results show that RQ3 is answered. Fix effort as measured using SLOC can be used in practice to order the modules based on fix effort. However, the fault count and fix cost in person hours can be more beneficial to software managers.

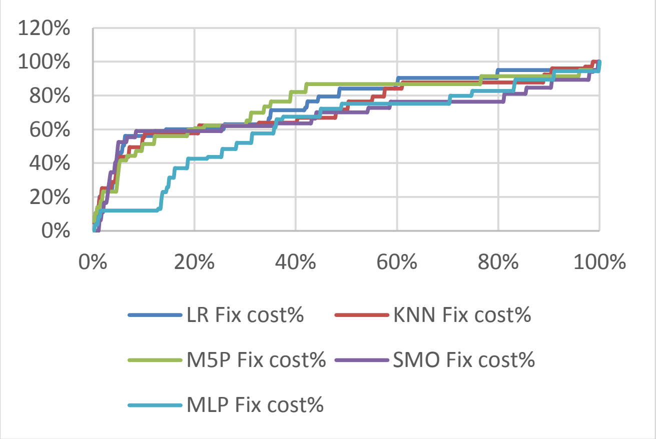


Fig. 5. Alberg diagram for five prediction models of KC3

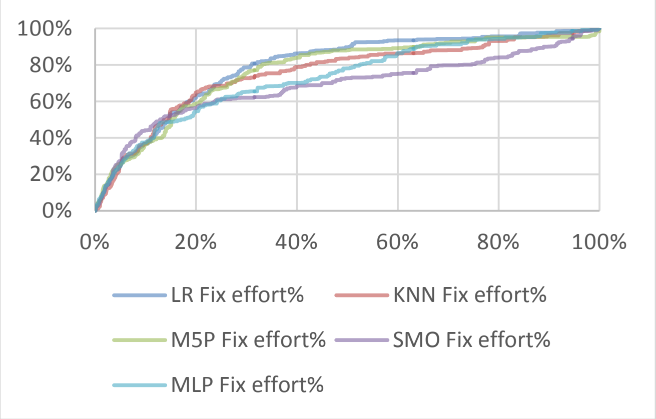


Fig. 6. Alberg diagram for five prediction models of KC4

TABLE VI. FIX EFFORT REGRESSION MODELS

Fixed SLOC	LR	kNN	M5P	SMOreg	MLPRegressor
KC1	53.05	56.32	53.58	56.66	61.66
KC3	36.91	38.00	35.49	34.58	49.85
KC4	109.63	92.66	90.41	103.79	129.00

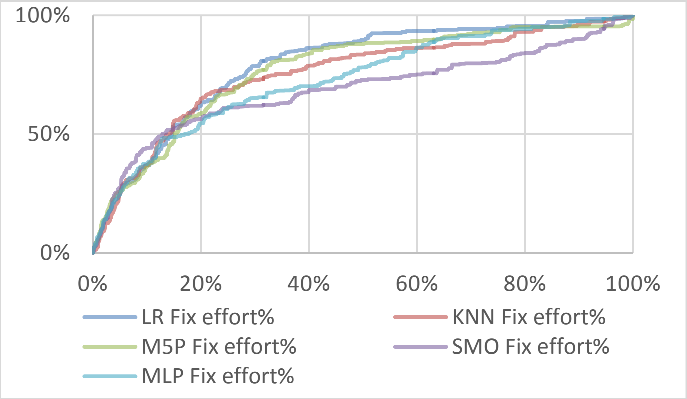


Fig. 7. Alberg diagram for five prediction models of KC1 data

D. Comparison of models performance

The RMSE results cannot be used to compare the results across the three fault measures because of the differences in measurement units. Therefore, we use another measure, the Relative Absolute Error (RAE), to analyze the results among the fault measures. The results of the models performance in RAE are reported in Table 7, where we find the following observations. In KC1, the Fault count models are the best in most models except one. In KC3, the fix cost models are the best except for two models. In KC4, the Fault count models are again the best. Therefore, for the systems under investigation, we can observe that prediction models based on fault count are slightly better in performance than other studied models. However, we do not observe large differences among the three fault measures under investigation. These results help the software engineers to consider other quality factors related to fault discovery and fix processes. The regression models for the fix cost and fix effort can be used similarly to fault count models.

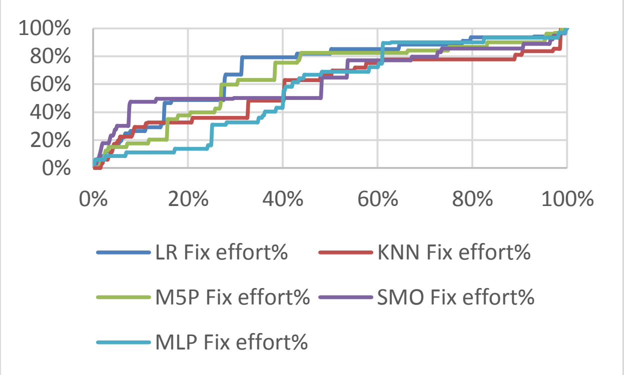


Fig. 8. Alberg diagram for five prediction models of KC3 data

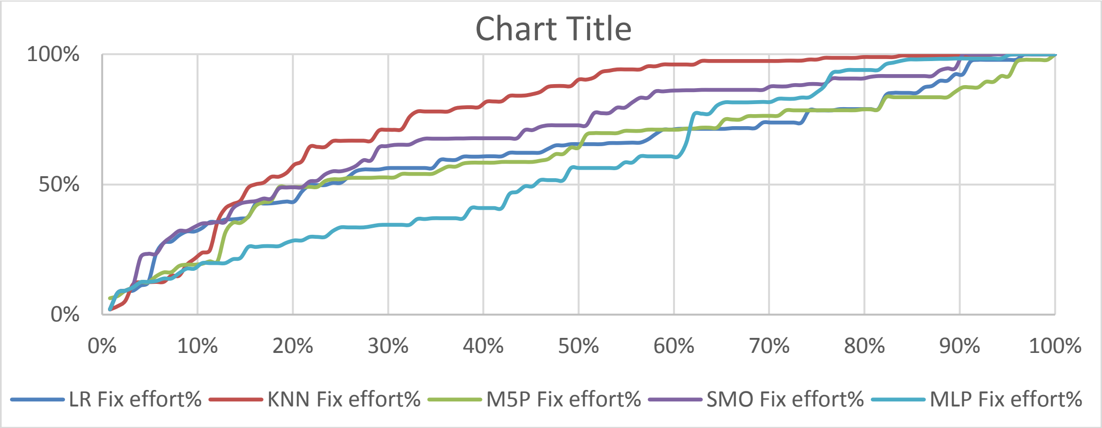


Fig. 9. Alberg diagram for five prediction models of KC4 data

TABLE VII. THE RAE RESULTS FOR FIVE MODELS

	KC1			KC3			KC4		
	Fault count	Fix cost	Fix effort	Fault count	Fix cost	Fix effort	Fault count	Fixcost	Fix effort
LR	**84.2**	91.3	89.7	90.6	**89.8**	101.2	**92.0**	102.2	97.9
KNN	**72.2**	76.4	78.4	71.4	**69.3**	73.5	**67.1**	77.1	73.4
M5P	**85.8**	88.2	87.5	**97.7**	96.1	99.2	**79.9**	84.6	82.7
SMO	58.2	**57.7**	58.4	51.6	51.6	**50.1**	74.9	70.4	**69.3**
MLP	**106**	114.1	109.7	131	**125.4**	137.7	**103**	122.3	114.1

The three quality factors can be used in practice to allocate resources, but it is important to know which models are consistent and always useful. The results of the practical implementation of the models for the three factors when 20% of modules are selected for further investigation are summarized in Table 8. The results show that both fault count and fix cost are more consistent than fix effort. In some cases, the cost models show better results. Furthermore, the use of fix cost in allocating resources provide more insights about the person hours spent to fix faults and can be considered a stronger indicator of where difficulties in code may appear.

V. VALIDITY THREATS

In the following, we address two kinds of possible threats that may affect the conducted research.

Construct Validity Threats: Construct validity refers to the degree to which the dependent and independent variables in this research measure the intended targets. Fix cost as measured in person hours are estimated by the developers and there is no detailed information about how developers estimate the fix cost. However, the data comes from a well-reputed organization, NASA, and their work is focused on quality of data and quality of work. The metrics in this study are well-studied metrics and recommended by many researchers to measure modules at procedural level.

Internal validity threats: internal validity is the degree to which conclusions can be drawn from the proposed data sets. This study depends on data from other organization and there is not enough information available about the development process followed in developing the three applications under study. However, the studied systems were considered in many other research papers and recommended to use by NASA.

External validity threats: External validity is concerned with the degree to which the results can be generalized to other research settings. The results of this study is based on only three data sets published by NASA MDP. We need more data sets to be able to generalize the results of this study into other systems. In addition, the systems are measured at procedural levels and conclusions may not be applicable for other paradigms like object-oriented paradigm.

VI. CONCLUSIONS AND FUTURE WORK

The fault prediction models are surrogates for the software quality. The assessment of faults in modules can be used to direct the efforts of software engineers in assuring software quality. Five well-known regression models were used to predict fault count, fix cost, and fix effort. The results of regression models for three data sets were reported. The results were not conclusive to find the best models in each data set and all regression models had similar performance. The prediction of fault count had a better performance in most models in KC1 and KC4 data sets. We found the prediction of fix cost is the best in KC3 only. Engineers may not have enough time to explore the quality of all modules in large software systems. It is vital to show the value of using these models in doing cost-effective quality assurance, e.g., prioritizing modules for further investigation. We have modeled the results of the prediction models by plotting the relationship between %modules and %faults after sorting the modules by faults predicted.

TABLE VIII. THE RELATIONSHIPS RESULTING FROM IMPLEMENTATION OF THE PREDICTION MODELS

	KC1			KC3			KC4		
	Fault count	Fix cost	Fix effort	Fault Count	Fix cost	Fix effort	Fault count	Fix cost	Fix effort
LR	20:58	20:59	20:62	20:54	20:60	20:49	20:49	20:60	20:43
KNN	20:58	20:58	20:63	20:50	20:58	20:33	20:57	20:58	20:58
M5P	20:59	20:48	20:58	20:48	20:60	20:38	20:46	20:60	20:49
SMO	20:58	20:57	20:56	20:54	20:59	20:50	20:49	20:59	20:49
MLP	20:58	20:44	20:54	20:42	20:43	20:14	20:30	20:43	20:29

We have also plotted the relationship between (%modules, %fix cost) and (%modules, %fix effort). The plots have shown that the 20:60 rule can be applied for the three measures.

These results are important to conclude that we can use the same metrics to predict different fault measures, i.e., answering the three research questions. The software engineers can have alternative methods to select software modules for further verification and validation from different perspectives. In future, we plan to expand this study to more diverse data sets.

REFERENCE

[1] V. R. Basili, L. C. Briand, and W. L. Melo, "A validation of object-oriented design metrics as quality indicators," IEEE Trans. Softw. Eng., vol. 22, pp. 751–761, 1996.

[2] K. El Emam, W. Melo, and J. Machado, "The prediction of faulty classes using object-oriented design metrics," J. Syst. Softw., vol. 56, no. 1, pp. 63–75, Feb. 2001.

[3] T. M. Khoshgoftaar and N. Seliya, "Comparative assessment of software quality classification techniques: An empirical case study," Empir. Softw. Eng., vol. 9, no. 3, pp. 229–257, 2004.

[4] T. Gyimothy, R. Ferenc, and I. Siket, "Empirical validation of object-oriented metrics on open source software for fault prediction," IEEE Trans. Softw. Eng., vol. 31, no. 10, pp. 897–910, 2005.

[5] Y. Zhou and H. Leung, "Empirical analysis of object-oriented design metrics for predicting high and low severity faults," IEEE Trans. Softw. Eng., vol. 32, no. 10, pp. 771–789, 2006.

[6] R. Shatnawi, W. Li, J. Swain, and T. Newman, "Finding software metrics threshold values using ROC curves," J. Softw. Maint. Evol. Res. Pract., vol. 22, no. 1, pp. 1–16, 2010.

[7] E. Arisholm, L. C. Briand, and E. B. Johannessen, "A systematic and comprehensive investigation of methods to build and evaluate fault prediction models," J. Syst. Softw., vol. 83, no. 1, pp. 2–17, 2010.

[8] L. C. Briand, J. Wust, J. W. Daly, and D. Victor Porter, "Exploring the relationships between design measures and software quality in object-oriented systems," J. Syst. Softw., vol. 51, pp. 245–273, 2000.

[9] T. M. Khoshgoftaar and K. Gao, "Count models for software quality estimation," IEEE Trans. Reliab., vol. 56, no. 2, pp. 212–222, 2007.

[10] R. Shatnawi, "Empirical study of fault prediction for open-source systems using the Chidamber and Kemerer metrics," IET Softw., vol. 8, no. 3, pp. 113–119, 2013.

[11] G. J. Pai and J. B. Dugan, "Empirical Analysis of Software Fault Content and Fault Proneness Using Bayesian Methods," IEEE Trans. Softw. Eng., vol. 33, no. 10, pp. 675–686, 2007.

[12] C. Catal and B. Diri, "A fault prediction model with limited fault data to improve test process," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2008, vol. 5089 LNCS, pp. 244–257.

[13] I. Gondra, "Applying machine learning to software fault-proneness prediction," J. Syst. Softw., vol. 81, no. 2, pp. 186–195, 2008.

[14] J. Zheng, "Cost-sensitive boosting neural networks for software defect prediction," Expert Syst. Appl., vol. 37, pp. 4537–4543, 2010.

[15] N. Ohlsson and H. Alberg, "Predicting Fault-Prone Software Modules in Telephone Switches," IEEE Trans. Softw. Eng., vol. 22, no. 12, pp. 886–894, 1996.

[16] S. Biyani and P. Santhanam, "Exploring defect data from development and customer usage on software modules over multiple releases," in Proceedings Ninth International Symposium on Software Reliability Engineering (Cat. No.98TB100257), 1998, pp. 316–320.

[17] T. J. Ostrand, E. J. Weyuker, and R. M. Bell, "Predicting the location and number of faults in large software systems," IEEE Trans. Softw. Eng., vol. 31, no. 4, pp. 340–355, 2005.

[18] H. Zeng and D. Rine, "Estimation of software defects fix effort using neural networks," in Proceedings - International Computer Software and Applications Conference, 2004, vol. 2, pp. 20–21.

[19] L. D. Panjer, "Predicting eclipse bug lifetimes," in Proceedings - ICSE 2007 Workshops: Fourth International Workshop on Mining Software Repositories, MSR 2007, 2007, p. 29.

[20] M. Hamill and K. Goseva-Popstojanova, "Exploring fault types, detection activities, and failure severity in an evolving safety-critical software system," Softw. Qual. J., pp. 1–37, 2014.

[21] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell, "A systematic literature review on fault prediction performance in software engineering," IEEE Trans. Softw. Eng., vol. 38, no. 6, pp. 1276–1304, 2012.

[22] D. Radjenović, M. Heričko, R. Torkar, and A. Živkovič, "Software Fault Prediction Metrics:A Systematic Literature Review," Inf. Softw. Technol., vol. 55, no. 8, pp. 1397–1418, 2013.

[23] G. Boetticher, T. Ostrand, and T. Menzies, "Promise repository of empirical software engineering data." 2007.

[24] T. Zimmermann, R. Premraj, and A. Zeller, "Predicting defects for eclipse," in Proceedings - ICSE 2007 Workshops: Third International Workshop on Predictor Models in Software Engineering, PROMISE'07, 2007, p. 9.

[25] NASA M.D.P., "NASA Independent Verification and Validation facility," 2014. [Online]. Available: http://mdp.ivv.nasa.gov.

[26] T. Menzies and J. S. Di Stefano, "How good is your blind spot sampling policy," in Eighth IEEE International Symposium on High Assurance Systems Engineering, 2004. Proceedings., 2004, pp. 129–138.

[27] M. Shepperd, Q. Song, Z. Sun, and C. Mair, "Data quality: Some comments on the NASA software defect datasets," IEEE Trans. Softw. Eng., vol. 39, no. 9, pp. 1208–1215, 2013.

[28] T. J. McCabe, "A Complexity Measure," IEEE Trans. Softw. Eng., vol. 4, no. 2, pp. 308–320, 1976.

[29] M. Halstead, Elements of Software Science. 1977.

[30] J. R. Quinlan, "Learning with continuous classes," in Machine Learning, 1992, vol. 92, pp. 343–348.

[31] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," Mach. Learn., vol. 6, no. 1, pp. 37–66, 1991.

[32] S. Sayed, An Introduction to Data Mining. 2014.

[33] L. Briand, J. Wust, S. Ikonomovski, and H. Lounis, "A Comprehensive Investigation of Quality Factors in Object-Oriented Designs: An Industrial Case Study." 1998.

[34] Y. Zhou, B. Xu, L. Chen, and L. Hareton, "An in-depth study of the potentially confounding effect of class size in fault prediction," Trans. Softw. Eng. Methodol., vol. 23, no. 1, p. 51, 2014.

Issues Elicitation and Analysis of CMMI Based Process Improvement in Developing Countries Theory and Practice

A Case of Pakistan

Shahbaz Ahmed Khan Ghayyur
Department of Computer Sciences,
Preston University,
Islamabad Kohat Campus, Pakistan

Muhammad Daud Awan
Department of Computer Sciences,
Preston University,
Islamabad Kohat Campus, Pakistan

Ahmed Noman Latif
System Analyst at South Taranaki District Council, Taranaki,
Wanganui & Manawatu,
New Zealand

Malik Sikandar Hayat Khiyal
Department of Computer Sciences,
Preston University,
Islamabad Kohat Campus, Pakistan

***Abstract*—Researchers have tried to find out the pattern of rising and fall of Pakistani software industry and also the reasons for what is going exactly wrong with this industry. Different studies have witnessed that in Pakistan, the software industry is not following the international standards. Another surprising fact, being observed in the past analysis, is the companies which have initiated CMMI-based SPI program have not even achieved the higher levels of CMMI from past three years, which is an alarming sign of the declining attitude of the industry. Therefore, it has become mandatory to look for the weak points or critical barriers or issues which are actually, the reason for this slow progress of CMMI-based SPI in Pakistan software industry. This study has identified that the issues for CMMI based SPI in Pakistan are much different than what it reported in the literature. Giving proper attention to the root of the problem can help solve many of the problems in this regard.**

Keywords—CMMI-based SPI; process improvement; inefficiency; Pakistani software industry

I. Introduction

Pakistan software industry is relatively new in the region, and its roots are not very deep. Therefore, the level of maturity is not yet attained. Lack of effort at government level and lack of presence of central controlling authorities or organization has caused slow growth for this industry. In 1995, it was felt by the Government of Pakistan that to guide and nourish the development of the software industry, ministries, departments and government agencies in the country, an organization needs to be created as no such existing department was providing the said services.

The formation of PSEB is also to make sure that it assists the software market of Pakistan regarding process improvement with the help of coping with the CMMI levels, human resource development, international marketing, making an effective strategy and promoting research providing innovation and enhancing technologies.

PSEB has successfully completed two different projects on CMMI so far. The first project, "Standardization of Pakistani IT industry (CMM Pakistan 2003)" was aimed at bringing the top five Pakistani IT companies to CMMI Level 3 or above, was completed with a cost of Rs. 39.9M. The second project, "Standardization of Pakistani IT industry on CMMI" was aimed at bringing at least 18 companies to CMMI Level 2 was completed with a cost of Rs. 39.3M. The objective of running these two projects was to initiate CMMI activities in Pakistan, and create a base for CMMI implementation at the mass level.

On close observation, we come to know that not many companies in Pakistan have initiated the CMMI-based SPI program which is obvious from the fact that out of more than 500 software companies, only 23 have initiated CMMI-based SPI program and out of these only one has achieved level-5. And this is why this research has become even more of an importance which will reveal these de-motivating factors in the industry. Avoidance and mitigation of these issues will result in a quick and smooth implementation of CMMI-based SPI in Pakistan. Since some demotivates are well known in literature for causing friction in CMMI based process improvement in developing countries this study aims to investigate if there is a gap between theory and practice when it comes to resisting the de-motivation factors. The real question is: are we fighting off the right issues for SMMI based SPI?

II. Literature Review

Software industry in Pakistan has a lot of potential to grow and compete with the other software industries in the region. We have brilliant resources and creative people who are driving this industry. This industry can play a vital role in the economic growth of Pakistan. Quality of work produced by the Pakistani professionals is highly recognized by the European, American and African markets [1]. We have a lot of work that is flooded towards Pakistani market. "*From its* *nascent beginnings in the late 1980s, the industry has successfully arrived to a point where its value proposition has been validated over and over again. The largest members are grossing 15-25 million dollars in revenues, and receiving 100 million dollar valuations. Most tech companies are growing in excess of 30% a year annually. The industry as a whole is doing over 2 billion dollars a year in revenue, up from less than a billion dollars a few years ago*" [1]. But over the years it has been noticed that the productivity of Pakistani software industry is not stable. There could be multiple factors that are causing such inconsistency. The most important factor that is hearting our industry very badly is poor process adoption. "*Remember that Pakistan is a country, which has only recently recognized the importance of ISO certification; although ISO has been around for much longer than CMM*" [2]. "*One of the most prominent human aspects is that software practitioners are de-motivated to deal with SPI initiative in their organizations* [3]. Despite having such huge potential and manpower, we are unable to produce even a single" organization like WIPRO, TATA and GOOGLE of the world. "*The famous Indian firm named TATA Consultancy has thirteen centers in India - it has 12 CMM Level-5 and 1 CMM Level-4 certifications to its credit; WIPRO has three Level-5 certifications to its credit*" [2].

Lack of adoption of software process improvement programs- like CMMI, is causing frequent closures or losses of software companies. "*Recent times have seen many Pakistani companies go bellies up - although I am not implying that lack of CMM initiative is the reason for their debacle- but this definitely was a contributing factor*" [2]. Processes are not adopted in their true spirits. Management is not aware of software improvement processes. In General, management of these companies wants ROI without spending money on the stability and continuation of the adopted processes. "*There are two challenges that a software development firm faces. First, to come up with reliable, efficient and pragmatic Official processes. Second, to make these processes a part of the company's culture i.e. to make the Official process the same as the Actual Process*" [2]

The adoption of software process improvement program has proved its value. Countries, where software industry have adopted such kind of standards, are in the fore front. Their software industry is contributing huge revenue segment. "*The **Actual Process** is what you do, with all its omission, mistakes, and oversights. The **Official Process** is what books say you are supposed to do"* [4].

The MPS.BR project in Brazil was initiated in December 2003. This program was to propose the software process improvement model for the small to medium software companies according to Brazilian software industry needs. "*The MPS Model is a software process improvement and assessment model, mainly oriented to the small and medium-size enterprises (SMEs). This model aims at: i) to fulfill the 'business need' of these firms; ii) to be recognized, locally and internationally, as an applicable model to organizations which develop or acquire software*" [5].

Through survey in different times, issues of software process improvement adoption, in general, and CMMI adoption, in particular, have also been identified. Through multiple reports published by PSEB, P@SHA, and other agencies that out of 500 software companies only 33 are CMMI certified for the different levels of CMMI. Most of the lot are at CMMI Level-3 or below. Reports are evident that most haunting factors out of the lot, which have been proved as issues are Time, Cost and ROI of CMMI.

Keeping in view the above discussion this study aims to find the difference between theory and practice when it comes to issues of CMMI-based SPI in Pakistan. Hypothesizing that this difference can be the vital factor that is causing the main hindrance in growth and process improvement of companies adopting CMMI and not reaping its fruits.

III. Research Method Details

Considering the limitations of this research and the differences in the survey methods, Personal Method in Interviews and Mail Method in Questionnaire have been selected for this research.

Questionnaire Design

The questionnaire designed for final or comprehensive survey has three sections out of which:

Section 1: Includes Company's information

Section 2: Includes Respondent's information

Section 3: Includes issues of CMMI-based SPI.

Data Collection Techniques and Methodologies

Questionnaires were floated via email and on personal contacts in different CMMI initiating organizations. It was made sure that the questionnaire reaches the maximum number of organizations with a pass-on strategy which suggests that while receiving the questionnaire the SPI practitioners were asked to forward it to other CMMI initiating companies if they have any personal contact with that particular organization. A total of 35 software companies were visited. A total of 33 companies were chosen to provide the research project with a cross-section of company maturity levels, company types, and sizes.

Sample

Whole population (all 33 companies SPI involved) employees were the sample size for this study. Since sample size was limited, a regular follow-up with the respondents was set up via emails, telephone calls and meeting them in personal. Some incentives were also introduced to get the maximum number of responses such as providing them with free discount coupons and scratch cards.

Identification of SPI Practitioners

In this survey, it is ensured that not just SPI Practitioners but software practitioners also responded to the questionnaire. For this purpose, some companies were visited personally, and respondents were contacted via phone and emails. It was ensured that this survey includes the whole sample of SPI practitioners and related persons.

Identification of SPI practitioners is achieved using the following criteria:

1) SPI practitioner is currently working/has already worked in a CMMI initiating organization.

2) SPI practitioner is currently working/has already worked in an organization where SPI is achieved using the model, which is similar to CMMI e.g., ISO standards etc.

3) SPI practitioner's willingness of being available for the interview.

Compilation of Issues

Once the feedback was received from the respondents, it was compiled into a spread sheet using SPSS and Excel categories-wise. When the data was successfully compiled, a list of issues was extracted based on agreed responses of the respondents.

Interviews to resolve open issues

Interviews had been conducted to resolve some of the open issues which couldn't be addressed in the questionnaire. A total of four unstructured interviews with four SPI professionals regarding the extracted issues had been conducted. The opinions are included in conclusion.

Identification of Renowned SPI Practitioners

Selection of renowned SPI practitioners had been done on the following criteria:

1) At least five years of SPI related experience OR

2) Worked in CMMI implemented organization for more than two years OR

3) Taken or conducted CMMI training in past five years OR

4) Achieved SEI/CMMI or related certifications.

Compilation of Data (Interviews & Survey Results)

A total of 33 software companies were visited. Participating companies were selected from a larger sample of companies who responded to the final questionnaire giving information about the problems they are facing regarding SPI intuitive in their organizations. The companies were chosen to provide the research project with a cross-section of company maturity levels, company types, and sizes. Since the questionnaires used for this research had both open and closed ended questions, therefore, the analysis of this research can be categorized in to qualitative and quantitative.

Type of Analysis used

There are two major types of analysis in survey research namely Quantitative Analysis and Qualitative Analysis. In this research, both of these analysis techniques were used out of which the quantitative analysis was focused more.

Quantitative Analysis

Quantitative research is best to be opted in a scenario which requires the numeric figure to be estimated from the questionnaire or in which numbers answer the questions rather than the answers to the questions. Usually, this type of research require a large amount of data which cannot be estimated or analyzed by qualitative research or a very time-consuming activity, the quantitative research can draw a meaningful result from this. *"The main beneficial aspect is that it provides the means to separate out the large number of confounding factors that often obscure the main qualitative findings. Quantitative analytical approaches also allow the reporting of summary results in numerical terms to be given with a specified degree of confidence."*

Key Aggregate Statistics On Responses

Total Number of Software Companies Surveyed	33
Companies achieved CMMI Level 5:	03
Companies achieved CMMI Level 4:	01
Companies achieved CMMI Level 3:	10
Companies achieved CMMI Level 2:	19
Companies working on Offshore Development:	21
Companies working on In-house Development:	10
Companies working on Both:	02
Companies of Age (1-4 Yrs):	08
Companies of Age (4-7 Yrs):	09
Companies of Age (7-Above Yrs):	16
Project Based Companies:	11
Product Based Companies:	17
Hybrid Companies:	05
Small-Medium Companies:	23
Large Companies:	10
Total Number of Software Practitioners Contacted	90
Total Number of Software Practitioners Responded	48
Senior Manager/Director:	12
Manager/Team Lead/Senior Executive:	17
Software Engineer/Developer/Junior Executive:	19
Respondent's Experience (1-3 Yrs):	11
Respondent's Experience (4-6 Yrs):	21
Respondent's Experience (7-10 Yrs):	16
Total Number of Questions in Questionnaire	50
Mandatory Questions:	48
Optional Questions:	02
Total Number of Agreed Responses:	18
Total Number of Neutral Responses:	17
Total Number of Disagreed Responses:	13

IV. Analysis & Results

Authors in an earlier study have already identified and classified the issues for CMMI-based Software process improvement and have organized the issues into categories. The same demotivates are used in this survey for practice identification in Pakistan Industry and then the results as shown below are used for identifying dependencies among dependent and independent variables in literature and compared with survey and for the creation of a de-motivator mitigation model.

Frequency Distribution of Overall Feedback from Respondents

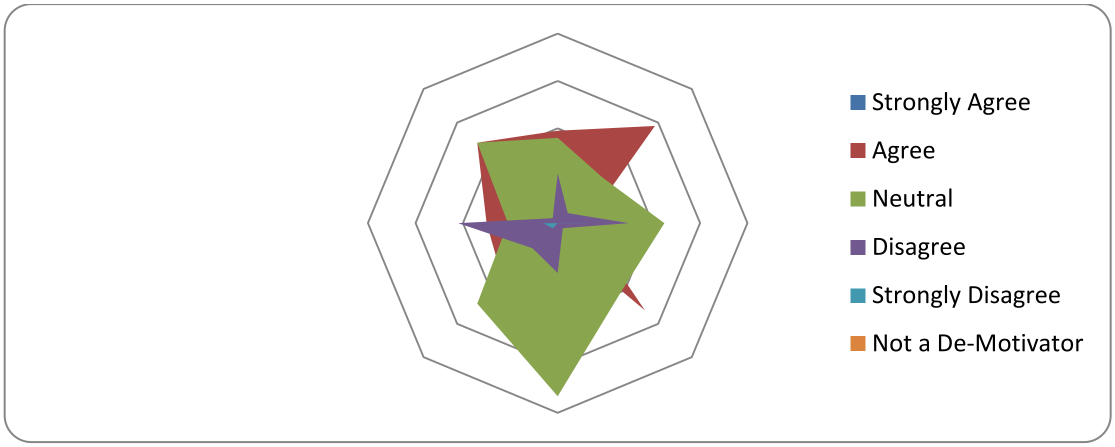


Fig. 1. Frequency Distribution Chart Overall Response Rate

TABLE I. OVERALL RESPONSE RATE

SN	De-Motivator Categories	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Not a De-Motivator
1	Communication Management	0%	39%	36%	21%	3%	0%
2	Cost Management	9%	58%	27%	6%	0%	0%
3	HR Management	0%	12%	45%	30%	12%	0%
4	Integration Management	6%	52%	39%	3%	0%	0%
5	Quality Management	0%	6%	73%	21%	0%	0%
6	Risk Management	0%	33%	48%	15%	3%	0%
7	Scope Management	0%	30%	21%	42%	6%	0%
8	Time Management	0%	48%	48%	3%	0%	0%

A. *Issues Frequency Distribution, Trend Agreement-Disagreement, Severity Measurement*

In this section, independent variables are analyzed with the frequency of responses, the trend of agreement and disagreement levels in de-motivator categories corresponding to different independent variables elaborated, and severity levels of the de-motivator categories are also listed.

1) Response Rate vs. CMMI-Levels

According to the feedback from all the respondents, it has been observed that nearly 60% of respondents are associated with the software companies having CMMI-Level 2 which therefore can be considered as the largest population among all the CMMI implemented Organizations.

Statistics show that it is difficult or time taking activity for the organizations to climb to higher levels of CMMI.

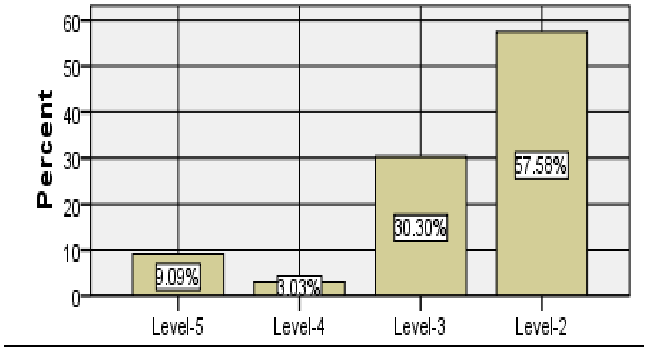


Fig. 2. Bar Chart: Response Rate & CMMI-Levels

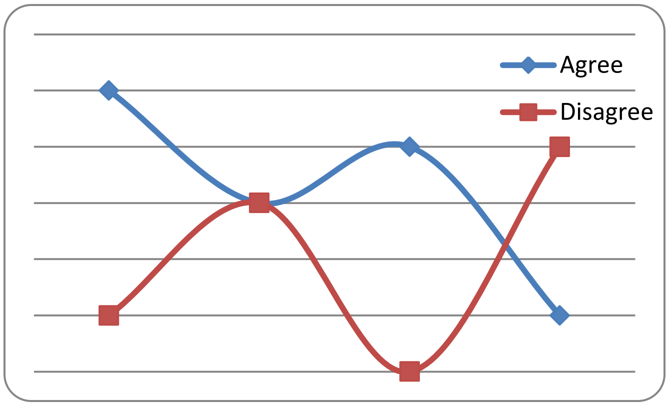


Fig. 3. Linear Graph: Response & CMMI-Levels

TABLE II. DE-MOTIVATOR CATEGORIES WITH SEVERITY LEVELS ACCORDING TO COMPANY SIZES

High	Medium	Low
Communication Management Cost Management Integration Management	Quality Management Risk Management Time Management	Scope Management HR Management

Above figure depicts that companies with CMMI Level-2 are in agreement with the Issues which then reduce to some extent in Level-3 organizations. Organizations with Level-4 have very less disagreements which then increase steeply in the Level-5 Organizations.

So it can be concluded that there is a mixed trend of agreements and disagreements for different De-motivator categories in comparison with CMMI-Levels.

Likert Scale i.e., "Agree-Neutral-Disagree" has been translated to "High-Medium-Low" according to our understanding for measuring the severity of De-motivator categories.

As per above mentioned criteria, following De-motivator categories have been listed according to their severity levels measured in the table 2.

2) Response Rate vs. Company Size

The frequency of responses according to company size is varied across different sizes. As shown in the graph below, the only type of enterprise that has minimal responses is micro-enterprise. Therefore, it is evident from the fact that the sample of this survey mainly contains the Small, Medium and Large Enterprises.

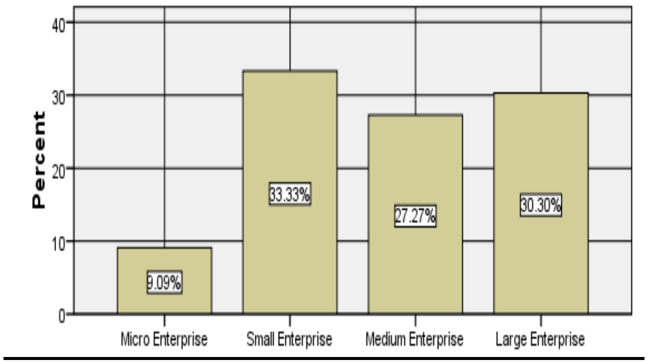


Fig. 4. Bar Chart: Response Rate & Company Size

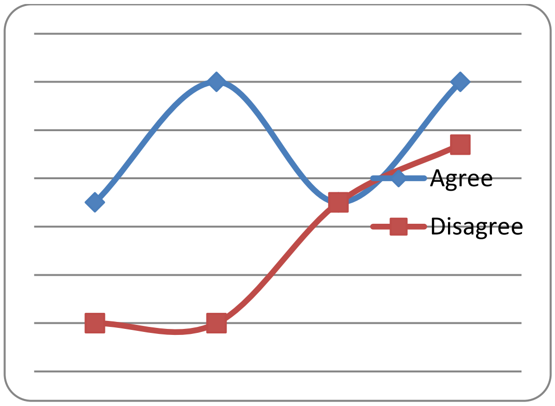


Fig. 5. Linear Graph: Response Rate & Company Size

TABLE III. DE-MOTIVATOR CATEGORIES WITH SEVERITY LEVELS ACCORDING TO CMMI LEVELS

High	Medium	Low
Integration Management Cost Management Time Management	Quality Management Time Management	Communication HR Scope

Above mentioned linear graph depicts that the agreement level of respondents increases as we move from micro to small enterprises, which then tends to drop equally in the medium enterprises and disagreement level increases steeply equal to four times the previous level. In the large enterprises, there is an equal trend of both agreement and disagreement for the Issues. The analysis confirmed that following De-Motivator categories are at the High, Medium, and Low severity levels according to company sizes.

3) Response Rate vs. Company Age

Statistics shows that companies older than seven or more years have a vast majority reaching almost 80% to all responses. The difference can also be seen in the graph below:

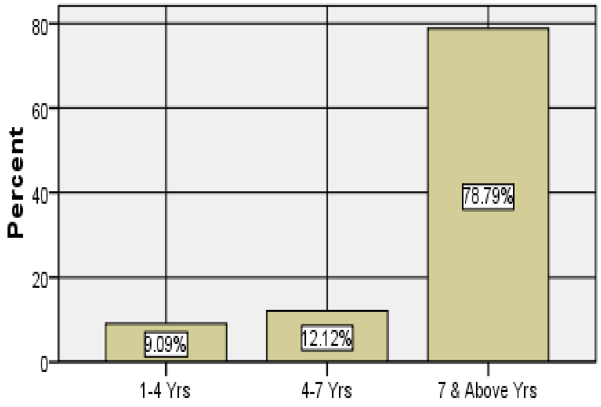


Fig. 6. Bar Chart: Response & Company Age

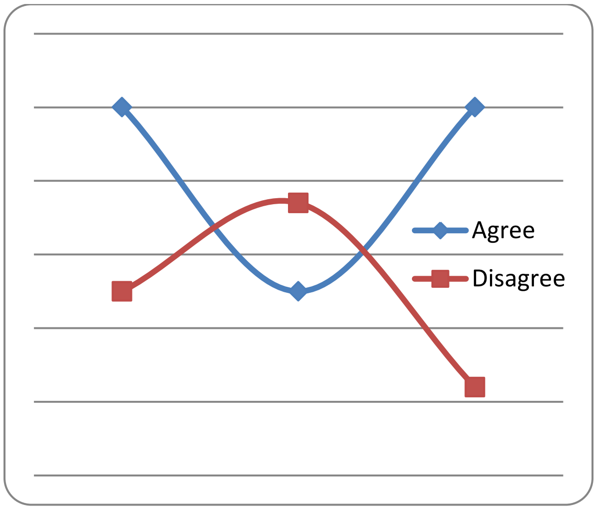


Fig. 7. Linear Graph: Response & Company Age

Rate of differences in agreements and disagreements can be observed in the linear graph mentioned above. As we can see that the agreement level is higher than the disagreement level in the organizations not older than four years which than decreases in the organizations older than 4 years but less than seven years, and the disagreement level increases gradually. And for the organizations older than seven years have a significant increase in agreement than disagreement.

The analysis confirmed that following De-Motivator categories are at the High, Medium, and Low severity levels according to Company Age.

4) Response Rate vs. company Type

Statistics have confirmed that Companies with Hybrid (Project & Product based) type are in the majority in the sample equals 42 % of total population. It is also evident that Product Based Companies are not very far behind Hybrid Type equaling almost 40 % of total population. So it can be said that the sample size of this survey contains mainly Product- Based and Hybrid Type Companies.

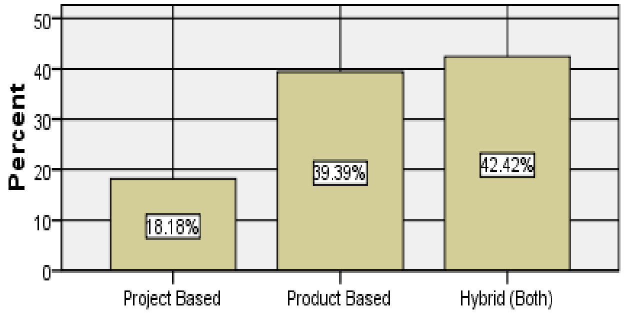


Fig. 8. Bar Chart: Response & Company Type

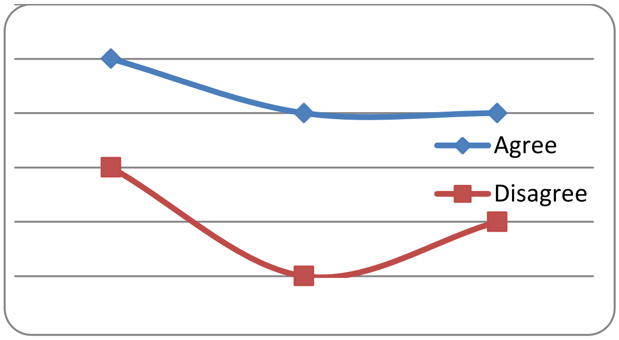


Fig. 9. Linear Graph: Response & Company Type

TABLE IV. DE-MOTIVATOR CATEGORIES WITH SEVERITY LEVELS ACCORDING TO COMPANY AGE

High	Medium	Low
1. Communication Management 2. Cost Management 3. Integration Management 4. Time Management	1. Quality Management 2. Risk Management	1. Scope Management 2. HR Management

The linear graph mentioned above states that the level of agreement is more or less constant in all type of companies except Project Based in which it reaches to forty percent agreement. On the other end, the disagreement level gets high in Project Based and Hybrid Type Companies.

The analysis confirmed that following De-Motivator categories are at the High, Medium, and Low severity levels according to Company Type.

5) Response Rate vs. Company's Type of Business

Our statistics have confirmed that Companies involved in Offshore Development are higher population rate than the rest of the companies equaling almost fifty percent of total population.

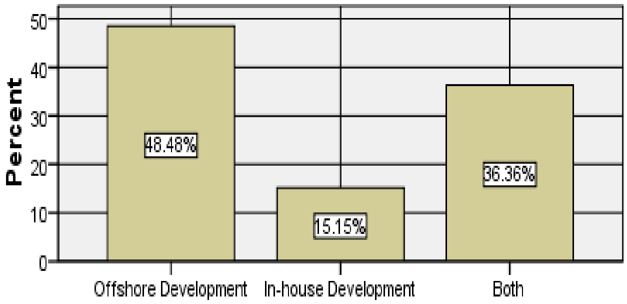


Fig. 10. Bar Chart: Response & Type of Business

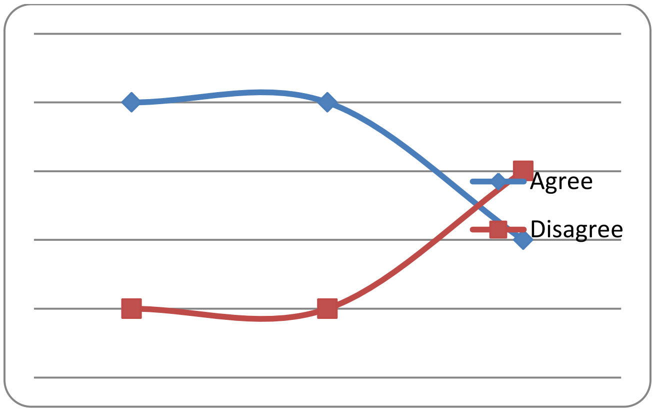


Fig. 11. Linear Graph: Response & Type of Business

TABLE V. DE-MOTIVATOR CATEGORIES WITH SEVERITY LEVELS ACCORDING TO COMPANY TYPE

High	Medium	Low
1. Communication Management 2. Cost Management 3. Integration Management	1. Quality Management 2. Risk Management 3. Time Management	1. Scope Management 2. HR Management

The above-mentioned linear graph displays that there is a constant increase in agreement level among both Offshore and In-house Company types. There is a sudden decrease in the agreement level and steep increase in the disagreement level in the companies having both types of businesses.

The analysis confirmed that following De-Motivator categories are at the High, Medium, and Low severity levels according to Company Types

6) Response Rate vs. Respondent's Experience

Statistics have confirmed that majority of the respondents population in this survey is of the software practitioners having experience more than four years and less than six years equaling almost fifty percent of total population. Practitioners having experience more than seven years are on the second number in the list having a population of almost thirty-three percent.

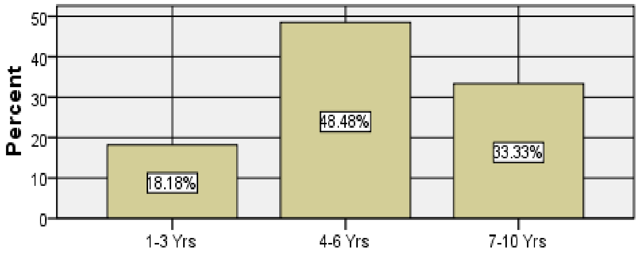


Fig. 12. Bar Chart: Response & Respondent's Experience

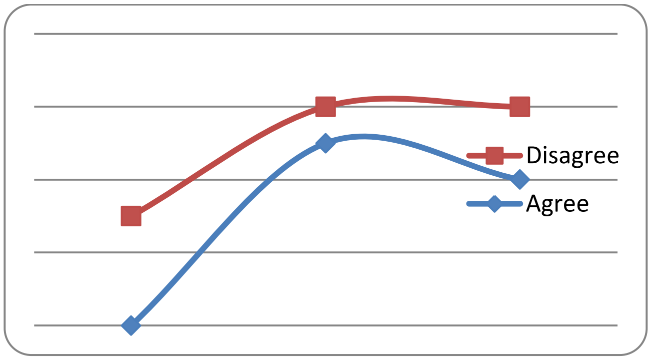


Fig. 13. Linear Graph: Response & Respondent's Experience

TABLE VI. DE-MOTIVATOR CATEGORIES WITH SEVERITY LEVELS ACCORDING TO COMPANY TYPE

High	Medium	Low
1. Communication Management 2. Cost Management 3. Time Management	1. Quality Management 2. Risk Management	1. HR Management 2. Scope Management

The above-mentioned graph explains that practitioners with one to three years of experience have almost higher disagreement level than the agreement level. The agreement level then increases among the respondents having experience of more than three years while disagreement level remains constant.

The analysis confirmed that following De-Motivator categories are at the High, Medium, and Low severity levels according to Respondents' Experience.

7) Response Rate vs. Respondent's Job Function

Statistics have confirmed that there are almost an equal number of respondents with different designations, whereas Software Engineers/Developers/Junior Executives are in the majority equaling almost forty-two percent. Manager/Team Lead/Senior Executive comes second with proportion equals to almost thirty-six percent.

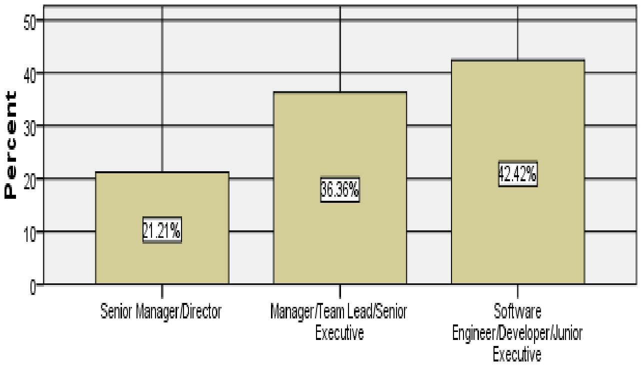


Fig. 14. Bar Chart: Response & Respondent's Job Function

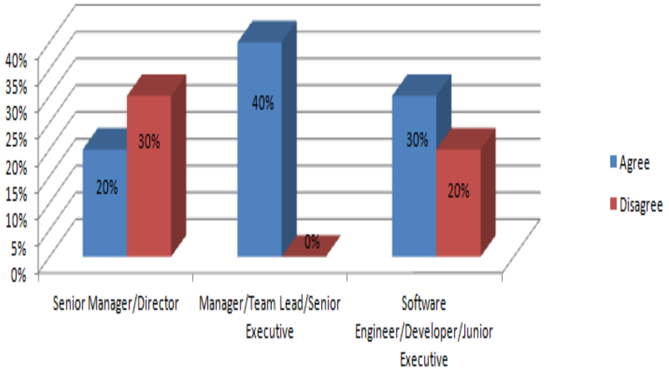


Fig. 15. Bar Chart: Response & Respondent's Job Function

TABLE VII. DE-MOTIVATOR CATEGORIES WITH SEVERITY LEVELS ACCORDING TO RESPONDENTS' JOB FUNCTIONS

High	Medium	Low
1. Cost Management 2. Time Management 3. Integration Management	1. Risk Management 2. Quality Management	1. HR Management 2. Scope Management

The above-mentioned bar graph shows that middle management is mostly agreed to the issues whereas the top management is mostly disagreed with them. However, in case of lower management, the agreement level and disagreement level among issues are equally divided. The analysis confirmed that following De-Motivator categories are at the High, Medium, and Low severity levels according to Respondents' Job Functions.

8) Issues Mitigation Model "A":

DMODEL with Job Function as "Independent Variable"

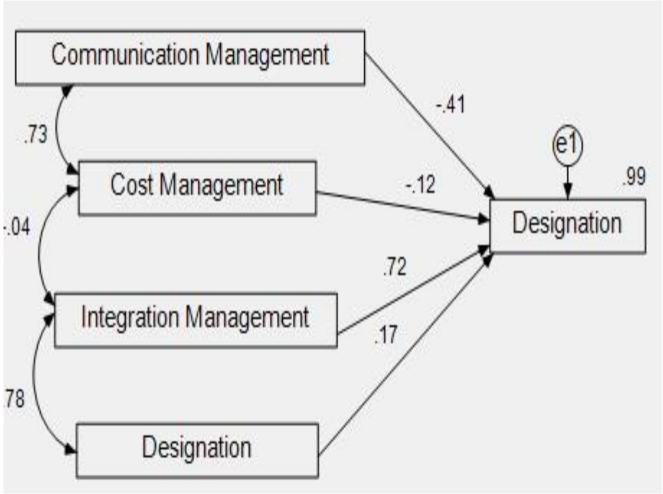
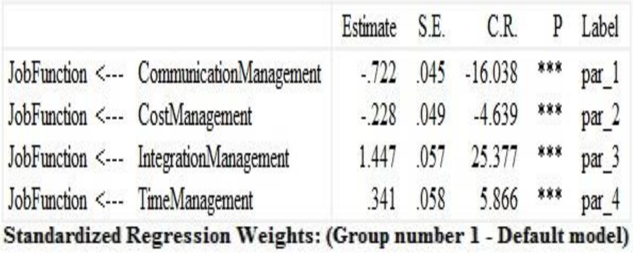


Fig. 16. Issues Mitigation Model-A

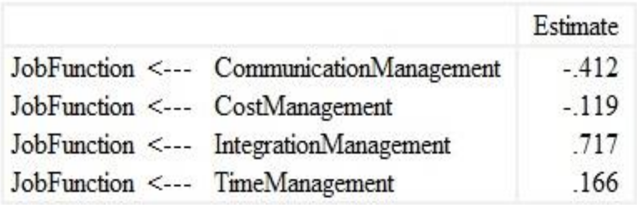
Regression Weights

Regression Weights

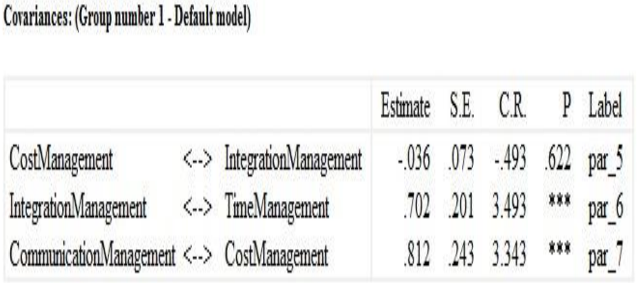


			Estimate	S.E.	C.R.	P	Label
JobFunction	<---	CommunicationManagement	-.722	.045	-16.038	***	par_1
JobFunction	<---	CostManagement	-.228	.049	-4.639	***	par_2
JobFunction	<---	IntegrationManagement	1.447	.057	25.377	***	par_3
JobFunction	<---	TimeManagement	.341	.058	5.866	***	par_4

Standardized Regression Weights: (Group number 1 - Default model)

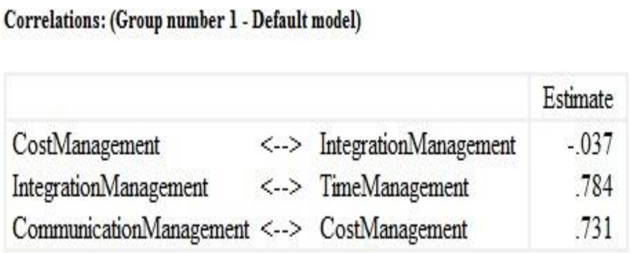


			Estimate
JobFunction	<---	CommunicationManagement	-.412
JobFunction	<---	CostManagement	-.119
JobFunction	<---	IntegrationManagement	.717
JobFunction	<---	TimeManagement	.166



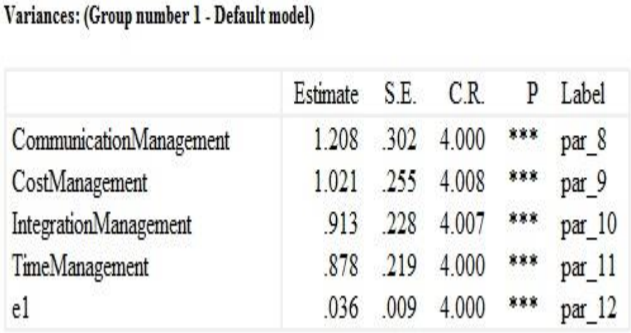
Covariances: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
CostManagement	<-->	IntegrationManagement	-.036	.073	-.493	.622	par_5
IntegrationManagement	<-->	TimeManagement	.702	.201	3.493	***	par_6
CommunicationManagement	<-->	CostManagement	.812	.243	3.343	***	par_7



Correlations: (Group number 1 - Default model)

			Estimate
CostManagement	<-->	IntegrationManagement	-.037
IntegrationManagement	<-->	TimeManagement	.784
CommunicationManagement	<-->	CostManagement	.731



Variances: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
CommunicationManagement	1.208	.302	4.000	***	par_8
CostManagement	1.021	.255	4.008	***	par_9
IntegrationManagement	.913	.228	4.007	***	par_10
TimeManagement	.878	.219	4.000	***	par_11
e1	.036	.009	4.000	***	par_12

TABLE VIII. MODEL "A" RELIABILITY TEST

Model Dependencies	Cronbach's Alpha
Job Function ⟷ Communication Management	.811
Job Function ⟷ Cost Management	.864
Job Function ⟷ Integration Management	.917
Job Function ⟷ Time Management	.949

9) Issues Mitigation Model "B"

Model with Company Size as "Independent Variable"

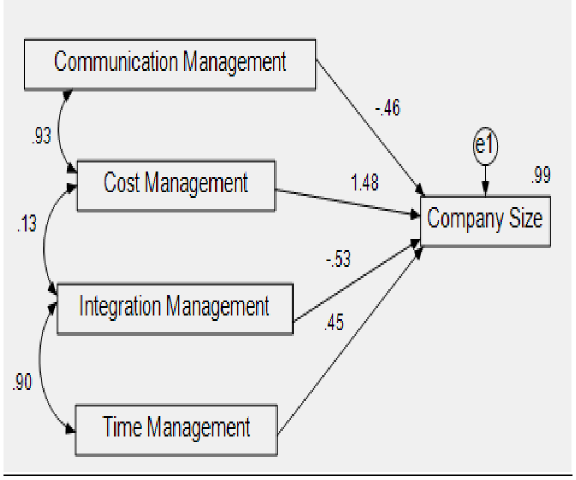
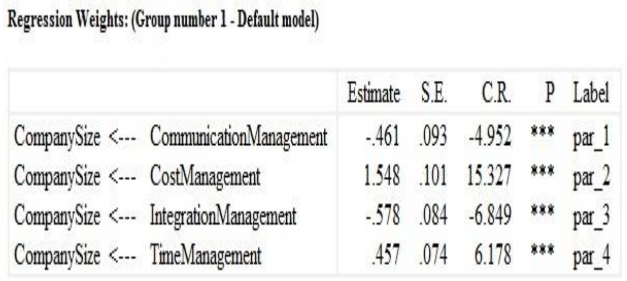
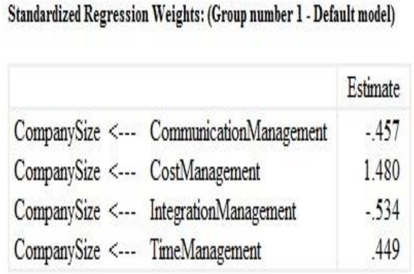


Fig. 17. Issues Mitigation Model-B



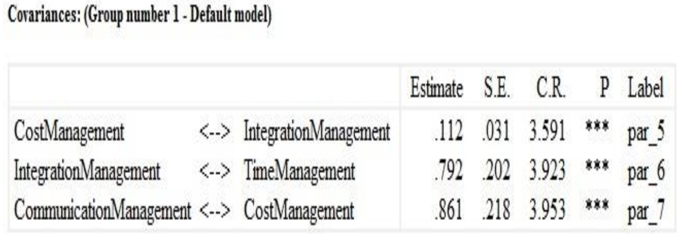
Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
CompanySize	<---	CommunicationManagement	-.461	.093	-4.952	***	par_1
CompanySize	<---	CostManagement	1.548	.101	15.327	***	par_2
CompanySize	<---	IntegrationManagement	-.578	.084	-6.849	***	par_3
CompanySize	<---	TimeManagement	.457	.074	6.178	***	par_4



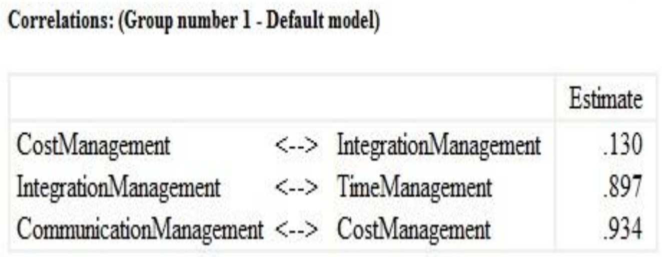
Standardized Regression Weights: (Group number 1 - Default model)

			Estimate
CompanySize	<---	CommunicationManagement	-.457
CompanySize	<---	CostManagement	1.480
CompanySize	<---	IntegrationManagement	-.534
CompanySize	<---	TimeManagement	.449



Covariances: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
CostManagement	<-->	IntegrationManagement	.112	.031	3.591	***	par_5
IntegrationManagement	<-->	TimeManagement	.792	.202	3.923	***	par_6
CommunicationManagement	<-->	CostManagement	.861	.218	3.953	***	par_7

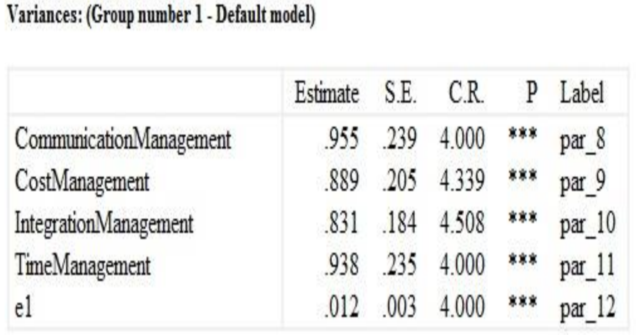


Correlations: (Group number 1 - Default model)

			Estimate
CostManagement	<-->	IntegrationManagement	.130
IntegrationManagement	<-->	TimeManagement	.897
CommunicationManagement	<-->	CostManagement	.934

TABLE IX. MODEL "B" RELIABILITY TEST

Model Dependencies	Cronbach's Alpha
Company Size ⟷ Communication Management	.961
Company Size ⟷ Cost Management	.815
Company Size ⟷ Integration Management	.933
Company Size ⟷ Time Management	.923



Variances: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
CommunicationManagement	.955	.239	4.000	***	par_8
CostManagement	.889	.205	4.339	***	par_9
IntegrationManagement	.831	.184	4.508	***	par_10
TimeManagement	.938	.235	4.000	***	par_11
e1	.012	.003	4.000	***	par_12

10)Variables Dependency

Literature vs. Findings

TABLE X. DEPENDENCIES AMONG DEPENDENT AND INDEPENDENT VARIABLES IN LITERATURE AND COMPARED WITH SURVEY

Dependencies from Literature	*Dependencies in this Survey*	*Participants Agreement /Disagreement with Literature*
1. It is quite difficult for any SME to choose an improvement approach and to apply it in their organization without the help of external consultants or substantial investment in the time of their software manage *[10]*.	• *SME and Consultancy: (17/33=51%)*	Mostly Agreed
2. Cultural issues like resistance to change from the employees or the management areas, who regard the extra work required for quality assurance as a useless and complicated burden put on the developing team *[10]*.	• Developers and Cultural Change: (9/33=27%)	Agreed
3. "According to Biro et al. [6], national culture also affects the process improvement methods."	• Lack of expertise in implementing Cultural Changes (22/33=66%) • Lack of defined SPI implementation methodology (16/33=48%)	Disagreed
4. Kuvaja et al. [44] mentioned that the main problem of the small companies is that they cannot afford to maintain substantial expertise of software process improvement within their companies, but they have to buy it from external sources.	• *Size of the Company (11/33=33%)* • *Budget Constraints (23/33=70%)* • *Balance between Technical efficiency and social considerations (15/33=45%)* • *Small Companies and Incompetent Staff: (6/33=18%)*	Disagreed

	• *Small Companies and Budget Constraints: (8/33=24%)*	
5. Due to budget constraints, the services of a consultant organization to improve the software quality are not possible *[10]*.	• *Consultancy (new factor)(21/33=63%)* • *Budget Constraints (23/33=70%)*	Mostly Agreed
6. People issues come under the umbrella of the ‘‘organizational’’ class and incorporate problems relating to*[9]*: a. Responsibilities, roles, rewards, expectations, blame; b. Staff turnover, retention, recruitment; *c. Lack of expertise in implementing cultural challenges* *d. Balance between technical efficiency and social considerations* e. *Politics*	• Rewards (7/33=21%) • Job Security (26/33=78%) • *Lack of expertise in implementing cultural challenges (22/33=66%)* • *Balance between technical efficiency and social considerations (15/33=45%)* • *Politics (10/33=30%)*	Mostly Agreed
7. The ‘‘tools and technology’’ category is recognized as a ‘‘project’’ problem and is the second most mentioned problem for developers and project managers. It includes issues such as *[9]*: a. implementation of new technologies and tools(including SPI generally and the CMM specifically), productivity, volume of work and pressures that inhibit the use of new tools	• *Balance between technical efficiency and social considerations (15/33=45%)* • *Lack of resources (new factor)(11/33=33%)* • *Tools and technologies with Developers:3/33=10%* • *Tools and technologies with Managers: 6/33=18%*	*Partially Agreed*
8. Documentation is also high on the list of developer problems. Project Managers are also concerned with documentation and state that CMM involves ‘‘too much paperwork. It is not as automated as it should be’’ *[9]*.	• *Documentation and Developers: (8/33=25%)* • *Documentation and PM: (6/33=18%)*	*Partially Agreed*
9. Differences in practitioner group problems. Senior managers have below average concern for project issues such as documentation and tools and technology issues, as they concentrate on problems relating to people and communication. They have above average concern for requirements issues in terms of problem ranking (equal 2nd), but an average concern in terms of percentage of problems. Indeed, further examination of Table 2 reveals that developers devote a higher percentage of overall problems to requirements than senior managers do with 11% and 10% respectively *[9]*.	• *Documentation and Senior Managers: (3/33=10%)* • *Tools & Technologies and Senior Managers: (2/33=6%)* • *People Issues and Senior Managers: (2/33=6%)* • *Communication and Senior Managers: (2/33=6%)* • *Requirements and Senior Managers: (9/33=27%)* • *Requirements and Developers: (4/33=12%)*	*Mostly Disagreed*
10. With experienced staff less rework of the documentation items is required, issues can be resolved quickly, and chances of destruction are reduced (Kautz and Nielsen, 2000; Moitra, 1998).	• *Practitioner's Experience and Documentation: (6/33=18%)*	Disagreed
11. Analysis of the responses based on the company size of the respondents shows that the practitioners working for small-medium size companies are highly motivated to support SPI initiatives by factors such as cost beneficial, job satisfaction, knowledgeable team leaders, and maintainable/easy processes.*[Motivators of SPI: An Analysis of Vietnamese Practitioners]*	• *Small-medium companies and budget constraints: (14/33=42%)* • *Small-medium companies and cumbersome processes: (5/33=15%)* • *Small-medium companies and job security: (19/33=57%)*	*Mostly Agreed*
12. "The large companies' practitioners are more motivated to support SPI initiatives by factors such as career prospects, cost beneficial, eliminate bureaucracy, feedback, job satisfaction, maintainable/easy processes, rewards schemes, shared best practices, task force, top-down commitment, training, and visible success" [7].	• *Large companies and job security: (7/33=21%)* • *Large companies and lack of feedback: (5/33=15%)* • *Large companies and cumbersome processes: (3/33=10%)* • *Large companies and rewards: (2/33=6%)* • *Large companies and commitment: (3/33=10%)* • *Large companies and training: (6/33=20%)*	*Mostly Disagreed*
13. "The larger size or hierarchy of a company, the more time needed to get a commitment from all levels of the organization" [8].	• *Large companies and commitment: (3/33=10%)*	*Disagreed*
14. "Participation, commitment and reasonable expectations are the end result which should be manifested by the organizational staff, if they are willing to contribute to the SPI project" [8].	• *Direction/Commitment/Requirement: (13/33=40%)* • *Incompetent Staff: (13/33=40%)*	*Agreed*
15. "Several respondents mentioned that the SPI project implementation result is also defectively affected if SPI schedule mix up with the ongoing software development project in their companies. The respondents are suggesting that proper and synchronized planning should be done to ensure that the SPI implementation schedule can be carried out harmoniously	• *SPI gets in the way of real work: (19/33=57%)* ○	*Agreed*

with the ongoing software development project" [8].		
16. *"RESULTS – We have identified 6 'high' perceived value SPI stimuli that are generally considered critical for successfully implementing SPI initiatives. These stimuli are: cost beneficial, job satisfaction, knowledgeable team leaders, maintainable/easy processes, shared best practices, and top-down commitment. Our results show that developers are highly motivated by: career prospects, communication, cost beneficial, empowerment, knowledgeable team leaders, maintainable/easy processes, resources, shared best practices, top-down commitment, and visible success; Managers are motivated by: job satisfaction, knowledgeable team leaders, maintainable/easy processes, meeting targets, shared best practices, and top-down commitment. Our results also show that practitioners of small and medium sized companies are highly motivated by: cost beneficial, job satisfaction, knowledgeable team leaders, and maintainable/easy processes; practitioners of large companies are highly motivated by: cost beneficial, reward schemes, shared best practices and top-down commitment" [7].*	• *Developers and job security: (9/33=27%)* • *Developers and Communication: (6/33=18%)* • *Developers and Budget Constraints: (11/33=33%)* • *Developers and Cumbersome Processes: (6/33=18%)* • *Managers and Job Security: (11/33=33%)* • *Managers and Cumbersome Processes: (2/33=6%)* • *Managers and Commitment: (6/33=18%)* • *SMEs and Budget Constraints: (15/33=45%)* • *SMEs and Job Security: (19/33=57%)* • *Large companies and Budget Constraints: (9/33=27%)* • *Large Companies and Rewards: (1/33=5%)* • *Large Companies and Commitment: (3/33=10%)*	*Partially Agreed*
17. There was also a clear link between the amount of Documentation carried out and the size and growth stage of the company; the smaller the company the greater the hostility towards documentation *[11].*	• *Small Companies and Documentation: (6/33=18%)* • *Large Companies and Documentation: (8/33=24%)*	*Disagreed*
18. However, even in the larger organizations, Documentation was regarded as a 'necessary evil' *[11].*	• *Large Companies and Documentation: (8/33=24%)*	*Partially Agreed*
19. In the course of the study interviews, few of the managers concerned expressed any enthusiasm about process or process improvement models. A far greater emphasis was placed on product, with process often believed to be a 'brake' on product development. *The managers believed process to have a significant cost which, in their respective companies, they attempted to keep to a minimum. What the managers perceived as the Cost of Process centred on a number of factors and these are represented as a network diagram in Figure 1 [11].*	• *Managers and Inadequate Matrices: (1/33=5%)* • *Product Based Companies and Cumbersome Processes: (2/33=6%)*	*Disagreed*
20. The interview extracts above demonstrate that many of the managers, far from being process converts, believe that many process activities are not essential and require too much time and resource. One of the process activities that managers consider can often delay, or hinder, product development, is Documentation *[11].*	• *Managers and Time Constraints: (15/33=45%)*	*Agreed*
21. Smaller companies, especially, feared having to allocate people, either to write the Documentation in the first place, or to manage it on an ongoing basis *[11].*	• *Small Companies and Documentation: (6/33=18%)*	*Disagreed*
22. Significantly, the resources required to implement SPI are proportionately much greater in smaller companies, and those smaller companies intent on, firstly, survival and then stability, have many competing and higher priorities than SPI *[11].*	• *Small Companies and Tools & Technologies: (3/33=10%)*	*Disagreed*
23. As all of the study companies, at time of interview, fell into the EU-defined SME category, it is therefore perhaps not surprising that they would reflect greater hostility to SPI models that required them to divert resources from what they would perceive as more deserving activities. For many of the interviewees, SPI creates an additional burden or weight to their development efforts resulting in increased Documentation and Bureaucracy *[11].*	• *SME and Incompetent Staff (16/33=50%)* • *SME and Simultaneous Focus on Many Improvement Areas (10/33=30%)*	*Agreed*
24. Small software companies, in the first instance, focus exclusively on survival. This, in part, explains the success of agile methodologies whose 'light', non-bureaucratic techniques support companies in survival mode attempting to establish good, fundamental software development practices *[11].*	• *Small Companies and Lack of defined SPI Implementation Methodology: (6/33=18%)* • *Small Companies and Inadequate Matrices: (2/33=6%)*	*Disagreed*

11) *Common Critical Issues in Literature and this Survey:*

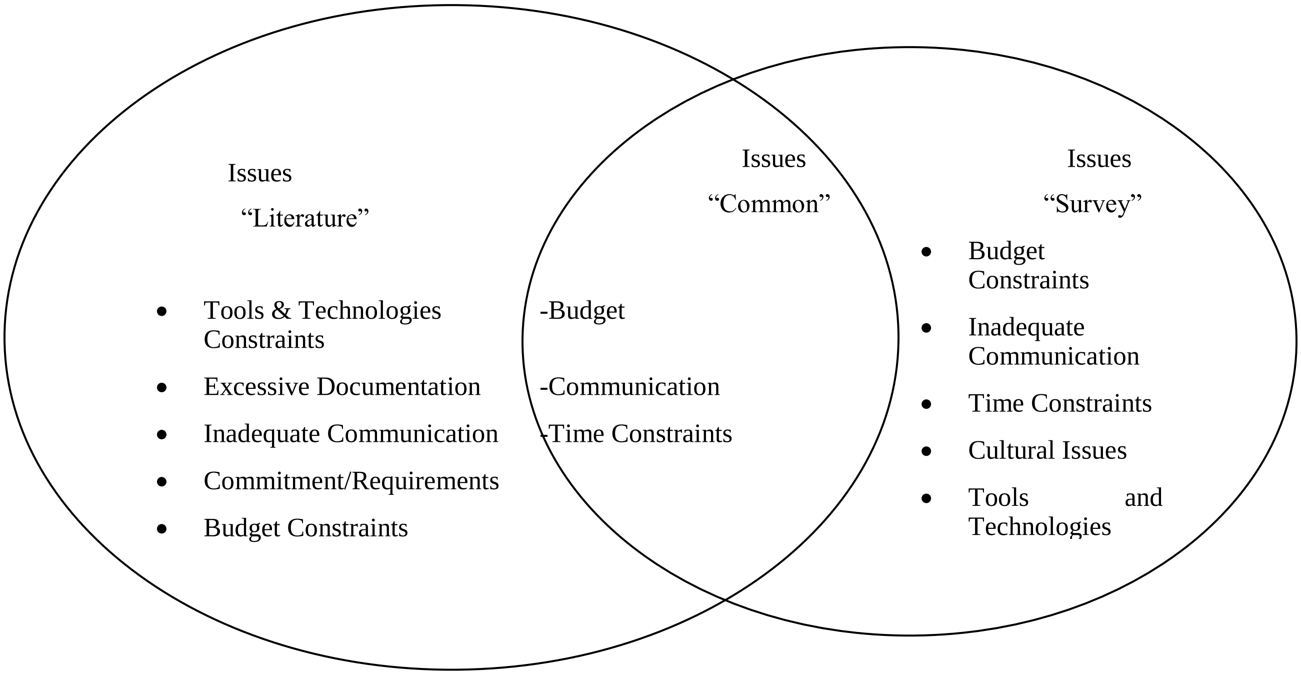


Fig. 18. Common Issues

12)SPI Practitioners perspective on Mitigating Issues

In order to know the perspective of SPI Practitioners on mitigating issues, three-step approach is adopted as follow:

a) Identification of renowned SPI practitioners

b) Conducting interviews of renowned SPI practitioners

c) Compiling and listing interview results

13)Selecting Renowned SPI Practitioners in Pakistan Software Industry

To achieve this, a thorough inspection was performed on Pakistan Software Industry. As a result, total of five renowned SPI practitioners are approached for the interviews according to the criteria already mentioned above. These practitioners are working or have already worked in the organizations i.e, Pakistan Telecommunication Company Limited (PTCL), Siemens, National Database and Registration Authority (NADRA), Teradata and NCR.

14)Conducting interviews

Interviews of the SPI practitioners were conducted with the help of an interview questionnaire attached at Annexure D. The questionnaire is divided in two sections: 1.It contains all the issues which were exposed in the literature and which were confirmed in the initial and comprehensive survey questionnaires. 2. Open ended questions which will reveal the proposed strategy to mitigate these issues.

V. Conclusion & Future Work

This survey is conducted on thirty three software companies of Pakistan who have implemented CMMI. There are a total of 48 software practitioners and 4 SPI professional who have participated in this research. The survey has extended the empirical research and case studies conducted on SPI Issues in other parts of the world. It has revealed more Issues which were missing in past articles. The analysis has revealed that Issues such as Time & Budget Constraints and Cultural Issues are on the top priority among all respondents. Whereas, there have been differences in responses among respondents with different job functions as the top management is mostly disagreed to all Issues. Surprisingly, most of the respondents with Middle and Lower Management agree with the existence of these Issues in the IT Industry and feel the need to mitigate them in order to successfully implement SPI initiative in their organizations.

A Model to tackle with these Issues has also been proposed after analyzing the results of the data gathered from the survey. The model has been tested regarding its reliability using SPSS's AMOS showing Cronbach's Alpha value above 0.7 which is required to keep the reliability intact among the elements in the Model. This Model can assist software practitioners to implement CMMI-based SPI not only in the Pakistan software industry but also in other under-developed countries.

The future work of this research is an implementation of the Model proposed, and the extension of this work i.e., Issues of SPI focusing CMMI Model only in other parts of the world. The differences between this research and the future research will in fact serve as the improvements in the Model and will lead it to perfection.

References

[1] "Annual Review of Pakistan's Software/BPO Industry 2007: Executive Summary".

[2] Kashif Manzoor: The Challenge of Implementing Capability Maturity Model (CMM) in Pakistan- Feb 2002

[3] Nathan Baddoo, Tracy Hall : "Issues for SPI: A Practitioners' View". University of Hertfordshire, UK. 2002

[4] Watts Humphrey Introduction to the Personal Software Process Addison-Wesley 1997 ISBN 352-762-875-3

[5] Gleison Santos, Marcos Kalinowski,Ana Regina Rocha, Guilherme Horta Travassos, Kival Chaves Weber, José Antonio Antonioni :

MPS.BR: A Tale of Software Process Improvement and Performance Results in the Brazilian Software Industry - © 2010 IEEE

[6] Bernard Wong, Sazzad Hasan, May 10, 2008, “Cultural Influences and Differences in Software Process Improvement Programs”. Leipzig, Germany.

[7] Mahmood Niazi, Muhammad Ali Babar, 2007, “Motivators of Software Process Improvement: An Analysis of Vietnamese Practitioners’ Views”. Keele University UK, 11th International Conference on EASE, Springerlink.

[8] Mohd Hairul Nizam Md Nasir, Rodina Ahmad, Noor Hafizah Hassan, 2008, “Resistance Factors in Implementation of Software Process Improvement Project in Malaysia”, ISSN 1549-3636, IEEE.

[9] Sarah Beecham, Tracy Hall & Austen Rainer, Software process improvement problems in Twelve Software Companies: An empirical Analysis, SCB/Problem paper, University of Hertfrodshire,UK.

[10] Deepti Mishra and Alok Mishra, “Software Process Improvement in SMEs: A Comparative View”, Atillim University, Ankara, Turkey.

[11] Rory V O’Conner, Gerry Coleman, “An Investigation of Barriers to the Adoption of Software Process Best Practice Models”, Australasian Conference on Information Systems, Ireland.

Energy Efficient Clustering Using Fixed Sink Mobility for Wireless Sensor Networks

Muhammad Ali Khan
I.T Department, Hazara University,
Mansehra, Pakistan

Noor ul Amin,
I.T Department, Hazara University,
Mansehra, Pakistan

Arif Iqbal Umar
I.T Department, Hazara University,
Mansehra, Pakistan

Shaukat Mehmood,
I.T Department, Hazara University,
Mansehra, Pakistan

Babar Nazir,
Department of CS, COMSATS Institute of IT,
Abbottabad, Pakistan

Kaleem Habib,
I.T Department, Hazara University,
Mansehra, Pakistan

***Abstract*—In this research an efficient data gathering scheme is presented using mobile sink as data collector with Clustering as sensor organizer in a randomly organized sensors in sensing field for wireless sensor network. The scheme not only extends the network lifetime through clustering process but also improves the data gathering mechanism through efficient and simplified mobile sink movement scheme. The cluster heads selection is based on both energy and data weight, which gathering all the data from the nodes within a cluster and then delivers it to the sink. The single mobile sink which visits the cluster heads as per defined path gathers data periodically. The scheme is organized through "Mobile Sink based Data Gathering Protocol (MSDGP)" which combines single message energy efficient clustering and data gathering process. For performance evaluation, the protocol is extensively simulated for different performance metrics, namely Residual energy consumption, Number of dead nodes during variable number of rounds and Network lifetime. The results proves that MSDGP is successful in achieving the defined objectives of energy efficiency and is capable of extending the network lifetime by increasing the number of rounds. Therefore, proposed protocol is more suitable for scenarios where sensor nodes generate variable amount of data.**

***Keywords*—Wireless Sensors Network; Mobile Sink; Clustering; Sensors**

I. Introduction

The current increase in the demand for usage of multimedia applications and their data in the field of Wireless Senor Network (WSN) anticipate an efficient infrastructure [1] for data gathering & delivery mechanism as well as energy efficiency over clustered topology. Such energy efficient infrastructure can be achieved by reducing the number of hops for communications [2], [3]. Considerable research has been focused on developing a robust energy efficient navigation system for the sinks in WSNs. In the recent years, multiple approaches have been proposed for both clustering and sink mobility [4], which broadly falls into two categories of proactive and reactive approach. In proactive approach, the sensed data is stored on the specialized central nodes storage, which is later collected by the sink [5]. On the other hand, in reactive approach, the sensed data is collected directly from the sensing nodes by the mobile sink.

The added motivation behind this research work is the further study of different factors contributing towards the strategies concerning clustering process and mobile sink adaptation in WSN environment and to utilizing these factors to devise simpler and efficient mechanism to improve sensor network efficiency and performance involving energy consumption and network lifetime, and contribute to a better understanding for the feature research activities.

The recent growth in event based and user centric applications which target remote monitoring, surveillance, and other related applications will deeply benefits from this research. This research aims to target the most common issues faced during the clustering of sensors and movement of Mobile Sink [2] for data gathering within a sensor field to provided data gathering scheme with limited constraints.

This paper presents a fixed mobility based reactive protocol named "Mobile Sink based Data Gathering Protocol (MSDGP)". In this protocol, the sensor nodes with highest energy and highest amount of data are selected as cluster heads which gathers data from the normal nodes within the cluster. This data is stored unless the mobile sink (with predefined mobility) comes within the transmission range of the cluster heads and request for the aggregated data. Once the request is received, the cluster heads forward the data to the mobile sink.

To performance of the proposed protocol is evaluated for multiple parameters using intensive simulations. The results show the MSDGP incurs low energy consumption with less dead nodes over a period of time, extending network lifetime. The paper is further organized by presenting core related research readings in Section 2 and proposed research protocol in Section 3. The Section 4 details simulations, findings and results, whereas paper is concluded in Section 5.

II. Related Work

In currently available literature, the researchers have concentrated on the different methodologies in this field of WSNs over the period of few decades both physically and in terms of their working. Enormous schemes have been designed to achieve optimum performance depending upon the purpose and their implementation. Architectural advancements have been done and two basic approaches [1] are mostly being followed; Static nodes with static sink / mobile sink and mobiles nodes with static sink / mobile sink, with experimentation with hybrid approaches as well.

Currently with the advancement in wireless technology, mobile sink or sink mobility is a hot topic and lot of work is being done in this field considering the large / multimedia type data targeting routing techniques for QoS and Network life time with goal of achieving efficient performance both in terms of routing and improved network life time (energy efficiency).

Sensors connectivity is utilized in schemes [5] to determine mobile sink path by locating data collection points at centroid of each cluster to minimizing multi-hop communications using travel salesman scheme for path determination. And schemes [2] where, MS path identification is based on message success rate from node to BS through information flooding over the network to find best path towards BS. Such schemes exert excessive inter-node communication and imbalanced energy usage due to direct communication with nodes.

Sink mobility techniques using dynamic virtually adjustable routing tree [3] using virtual backbone for routing data from CHs to edged circular mobile sink also limits energy efficiency as MS path consists of virtual tree and straight lines whereas cluster heads are selected nearest to the selected path. Each CH may follow multiple paths based on the location of MS causing overhead of path reconstruction. Whereas, direct mobile sink data gathering method based on centroid collection points in clusters [6] may collect data directly from all nodes but increase nodes-MS communication, limiting its implementation for large clusters.

Multiple specialized relays based data transfer schemes using pre-define movement of mobile sink, relays are selected based on point that CHs closest to mobile sink [7]. The constraint fixed trajectory of MS highlight the to-and-fro messaging neglects the energy consumption and multi hop data transfer rises data loss and collision. Similarly, [8] signal strength is considered for next-hop selection, whereas slow MS speed along its trajectory increases delay in data delivery to sink from high storage and energy abundance relays. In [9], protocol is applicable to limited scenarios, because the sink speed is fixed and the overall network lifetime is limited.

In [4], energy-hole problem is discussed through MS based Routing Protocol (MSRP) to increase life of network. CHs gather data from the normal nodes and wait until the sink arrives. On the arrival of the sink, the CHs send their data to sink. The sink considers energy of the CHs to schedule its data collection. Do to so, the sink records the energies information of CHs. As the sink gathers data from CHs and their neighbors to balance the energy consumption, this protocol provides enhanced network lifetime [10].

In [11], protocol called Virtual Circle Combined Straight Routing is proposed, using a combination of circles and straight lines to elect Cluster-Heads (CHs) and create a virtual backbone network. CH nodes collect data for delivering to sink using tree structure. The CH broadcasts request to sensors and form a tree to provide shortest path to the sink. The limitation of this protocol is that the sink follows only a circular path can cause longer delays in data collection. In [12], the authors proposed a protocol which visits subset of nodes to get sensed data. In this protocol, depositing sensed data at nodes and formation of virtual structure may cause an imbalance in energy and storage space. It is because the nodes and their next hops may quickly deplete their energy. Sensor nodes purposively has buffer space for sensed partial data storage, composed of data as well as compressed version of their neighbors. This protocol is suitable for scenarios where real time data is not a primary concern.

In [13], the authors propose a geographic routing protocol based scheme called Integrated Location Service and Routing ensuring MS packet delivery. MS, sensors, neighbor nodes all broadcast their location. MS movement is slow; network is static and MS hang on to at least one node on the network during one complete process. Target nodes near the sink are considered first for any changes in the data. The process of location update is time-stamped.

In [14], propose a protocol for delay-tolerant application. It focuses on self-aware MS for its future trajectories through broadcast of anticipated trajectories set to nodes. Relay nodes are used as mind points to route data to mobile sinks with trajectories information. Relays store all data for delivery to sink on its passing and predict other possible trajectories of the mobile sinks. This protocol provides improved data delivery, but incurs high energy consumption due to retransmissions. Similar concept [15] of mobile CHs which increase WSN lifetime for real time applications. The mobile CHs gathers data from the network before sending it to sink situated at the center of the sensing field. All moving CHs ensure connection with base-station during reporting. Three schemes are adapted for CHs movement to minimize communications and enhance network lifetime. The requirement of resource-rich mobile CHs is the major limitation of this scheme.

In [16], hierarchical cluster based architecture with large number of nodes, their CHs and one mobile sink is proposed. A predefined movement strategy through finding an optimum Hamiltonian Round calculated by sink according to Euclidean distance table provided by sink. Sink broadcasts message for CHs location which is replied with Cartesian coordinates, the sink calculates Euclidean distance between each CH and other CHs and between each CH and itself. In [17], scheme requires a single mobile sink for data collection. Do to so, the MS moves in a straight direction for data collection from nodes. This protocol focuses to minimize communications through efficient path planning with an ultimate goal of energy efficiency. To do so, a message is sent to the sink from where it is transmitted to the nodes and their neighbors. The sensor nodes report data to the sink when event of the similar interest occurs [18]. Moreover, acknowledgment is used to confirm successful packet delivery for each packet. The drawback is the limitation in fixed line path of mobile sink and avoiding delay.

The clustering protocol in [10], basis cluster heads selection on residual energy and centroid positioning of the node. Due to static sink, inter-cluster communication is done by utilizing cluster's overlapping nodes as anchor nodes or guard nodes. Excessive intra and inter cluster communication is done for data gathering.

The DEMC protocol is used to check the performance of MSDGP, and to evaluate if the selection of cluster heads on the basis of residual energy and data volume provides improved network lifetime then scheme of residual energy and the centroid positioning of that specific node.

III. The Proposed Protocol

A WSN consists of a set of nodes V, where v represents a single node and v∈V. Each node is recognized by a unique identifier, whereas all nodes communicate via full-duplex links. Moreover, a cluster is a subgroup of V, and multiple clusters cover the entire network.

The assumptions for our network model are listed as follows:

- Randomly deployed sensor in the sensing field.
- Nodes will be stationary.
- Nodes will use fixed transmission power.
- The sink is mobile with rich resources.
- Nodes unaware of their location information.

MSDGP uses three types of nodes, mentioned as follows:

1) *Normal nodes:* Nodes that sense information from the surroundings.

2) *Cluster-heads:* The nodes accountable to collect data from the normal nodes.

3) *Sink node:* The node that collects data from the cluster-heads and provide it to the outer world.

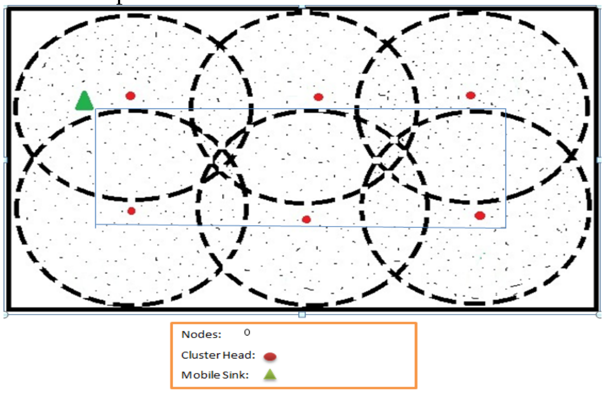


Fig. 1. General Structure of proposed WSN

B. MSDGP Operation

All the sensors are randomly deployed in the target sensing field; clustering is performed where cluster heads are selected through election between sensors on the basis of both data stored and residual energy of involved sensors. Once the cluster heads are selected, the normal nodes associate themselves with the cluster heads and forward their sensed readings to that cluster heads within a cluster. Once the data is gathered, cluster heads forward their collected sensed readings to specialized mobile sink whenever it comes within the vicinity or in their transmission range. The proposed is a fixed mobility based reactive wireless sensor networks with basic objective behind the selection of cluster heads on the basis of energy and data is to minimize the overall intra cluster communication to achieve extended network lifetime.

C. MSDGP cluster formation

The formation of the clusters for the sensor nodes is purely based on selection of their cluster heads. Therefore, cluster heads must be selected in an optimal way. MSDGP selects the cluster heads by calculating residual energy (E) and size of the sensed data (D) of the sensor nodes. To do so, each node calculates a weight (W) considering E and D and makes a decision of becoming a cluster head. The weight can be calculated as:

$$W = (\alpha \times E) + (\beta \times D) + (\mu \times I) \qquad (1)$$

Where,

$$0 < \mu < \beta < \alpha < 1.$$

The weight (W) is a linear combination of E, D, and I, where E has greater proportion in weight calculation followed by D and I, respectively. It is because, we have given energy the highest importance followed by size of the sensed data, whereas I works as a tie breaker incase two nodes have similar energy and data size. Once, every node calculates its weight, a timer (T) (delay) is set for the broadcast of a clustering request message as per the following:

$$T = \frac{1}{W} \qquad (2)$$

In the initialization phase, all nodes assume that they are the cluster-heads, therefore, they set their cluster-head flag to true. However, when the respective timers of the nodes are reached, the nodes broadcast their cluster message as per weight of the node. When the other nodes receive the clustering message, they simply compare their received weight to their local weight. If the received weight is greater, they set the node (from which the message is received) as their cluster-head. Moreover, then mark self-cluster head flag to false, cancelling scheduled clustering request.

D. MSDHP Intra-cluster communication

Once the clustering is performed and completed, the remaining of the nodes associate themselves with the cluster heads and start reporting their sensed data to that respective cluster heads based on saved cluster-head ID. The cluster head on receiving the data performs aggregation and sends that data to available mobile sink, normally when cluster head is within the transmission range to that mobile sink.

E. Mobile Sink Data Collection

In MSDGP, the sink follows rectangular mobility model and covers the entire network crossing the clusters formed. It sends data requests to cluster heads within the transmission range and collects the aggregated data. In this way, the sink collects the aggregated data from the cluster-heads. The data gathering phase depends on the speed of the mobile sink and is long enough in which a mobile sink can traverse throughout the network.

F. MSDGP Algorithm

The following sub section presents the algorithm being employed for MSDGP, It has been divided into four phased namely Initialization phase, Phase1, Phase2, and Phase3. The initial phase is responsible for information gathering required for initial setup and working of the preceding phases. In phase 1, clustering is done, which includes selection of cluster head. Phase 2 carries out data gathering by CH from normal nodes. Phase 3 is responsible for gathering data from CH through movement of mobile sink.

```
Initialization phase:
    Deploy_Nodes()
    I ← Assign_Node_ID()
    E ←Assign_Default_Energy()
    D ←Sense_Data()
    W = (α × E) + (β × D) + (μ × I)
    $T = \frac{1}{W}$
    Is_Cluster_Head =  True
    Schedule_CH_Message (T)
Phase 1:
    IF Timer_Expires()
      Broadcast_CH_Message()
    END IF
    IF CH_Message_Received()
        IF Rec_W >W
            Set_CH(Rec_Node_ID)
            Is_Cluster_Head = False
        END IF
    END IF
Phase 2:
    IF Normal_Node()
        Send_Data (D, CH_ID)
    END IF
Phase 3:
    IF Node_is_Sink()
        Broadcast_Data_Request()
    END IF
    IF Data_Request_Received()
        IF Node_is_CH()
            Send_Data(D, Sink_ID)
        END IF
    END IF
```

IV. Results and Discussion

The comprehensive simulations have been organized to evaluate the performance of MSDGP using OMNET++, a simulator that uses INET [19] framework. INET framework consists of simulation modules that are specially designed for wireless sensor networks. The energy model that is presented in [20] is used based on which, the energy consumed in transmitting a k bit message over a distance d is calculated as:

$$E_{Tx}(k,d) = E_{Tx} - elec(k) + E_{Tx} - amp(k,d)$$
$$E_{Tx}(k,d) = E_{elec} \times k + E_{amp} \times k \times d^2 \tag{3}$$

Similarly, the energy consumed in receiving the message is calculated as:

$$E_{Rx}(k) = E_{Rx} - elec(k)$$
$$E_{Rx}(k) = E_{elec} \times k \tag{4}$$

Where, $E_{Tx}(k,d)$ represents, amount of energy that is needed to send k bit message over distance d meters. Similarly, $E_{Rx}(k)$ represents the amount of energy that is needed to receive k bit message. Moreover, E_{elec} and E_{amp} represents the amount of energy required for using the transceivers and amplifier, respectively.

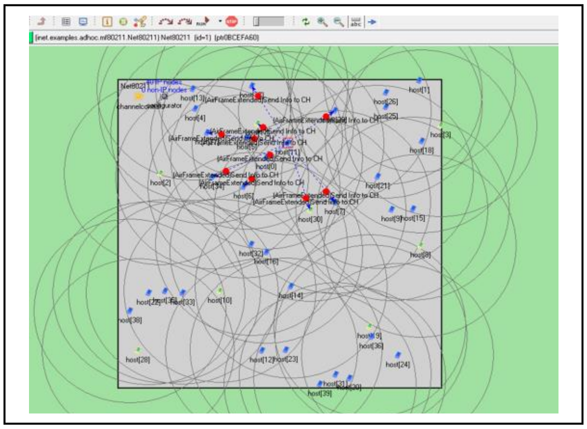


Fig. 2. MSDGP simulation using OMNET++

The simulation setup and other parameters for the simulation are detailed in table 1. All the protocols evaluated during the performance comparison have been setup using these parameters.

TABLE I. Simulation Setup

Type	*Parameter*	*Value*
Network	Field dimensions	1000×1000
	Nodes location	Random
	Sink mobility model	Rectangular
	Sink initial energy	9 J /battery
	Sink speed	Fixed 5 (0-20 m/s)
	Nodes initial energy	3 J /battery
Application	Data packet size	Variable
	Broadcast packet size	25 bytes
	Packet header size	25 bytes

Radio Model	E_{elec}	50nJ/bit
	E_{amp}	0.0012pJ/bit/m^4

A. *Sum of Residual Energy:*

Figure 3, shows the results for the Sum of residual energy consumed by all the sensor nodes with respect to variable number of rounds. As MSDGP and MSDGP-E are designed for static sensor network with mobile sink, the overall energy consumption is recorded to be low as compared to DEMC as DEMC is designed for static sink or base station. However, during clustering phase both MSDGP and MSDGP-E uses single message clustering approach which plays important role in saving energy required for clustering as compared to DEMC which not only uses multi-messaging during clustering but also uses multi hoping for data delivery to base station which is avoided by both MSDGP and MSDGP-E. Considerable amount of energy is saved in both phases. The result shows that for all three schemes the energy consumption spikes during the middle rounds and then slowdowns due to the low concentration of the nodes towards the end.

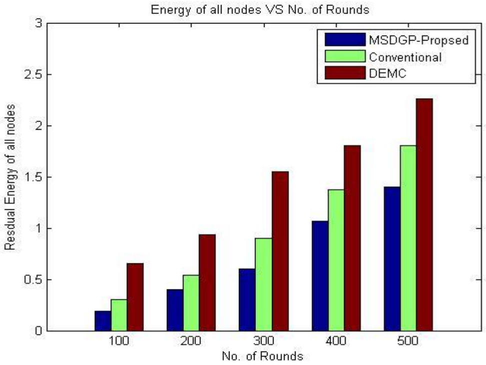


Fig. 3. Sum of residual energy of all nodes

B. *Number of Dead Nodes:*

Figure 4, shows the number of dead nodes with respect to variable number of rounds of the MSDGP, MSDGP-E and DEMC protocols. As projected in result and due to the fact that both MSDGP and MSDGP-E are based on similar sink mobility approach so there is only small difference among their number of dead nodes due to clustering technique difference whereas DEMC being totally different approach shows major difference in number of dead nodes specially towards the end. The results proves that multi-hoping consumes more energy during inter-cluster communication.

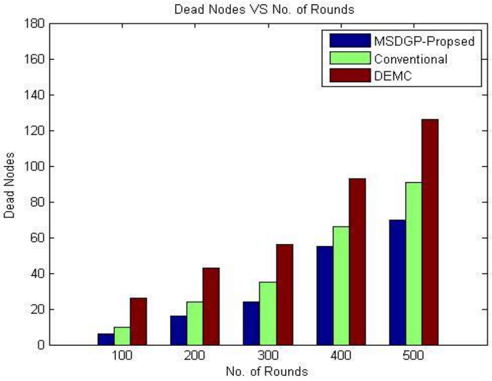


Fig. 4. Number of dead nodes w.r.t rounds

MSDGP uses intelligent clustering mechanism that uses a single message for cluster formation. In addition, the cluster-head communicates with the sink only. Hence, it can be concluded that it also uses optimum number of clustering messages to achieve the required level of packet delivery ratio, which does not exceed one. However, under high mobility of the mobile sink, the packet loss may increases and the packet delivery ratio is affected.

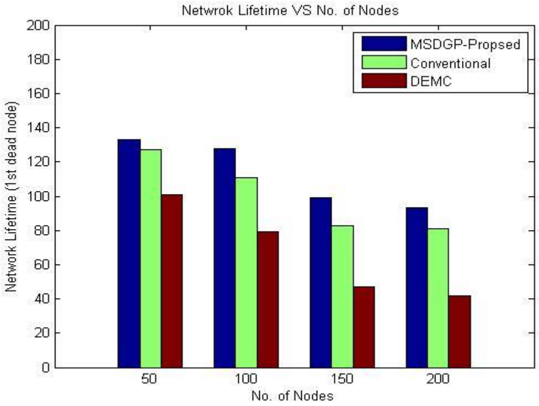


Fig. 5. Network lifetime w.r.t variable nodes

C. *Network Lifetime:*

Figure 5, presents the network lifetime of proposed MSDGP, conventional approach MSDGP-E and base approach of DEMC with respect to variable number of nodes being introduces into the system. As all three schemes uses same fixed transmission power as per the simulation setup, the overall network lifetime is shown as it fluctuate when the sensor nodes increases. Due to the fact that all three schemes select cluster heads using different approaches which make remarkable difference in their overall network lifetime. It is because; when concentration of nodes is less, lesser is the intra-cluster communication w.r.t clustering whereas selection of cluster heads in MSDGP (residual energy and data volume) reduces the overall communications in the network. Consequently, MSDGP provides improved energy efficiency (due to less communications) and extends the network lifetime as compared to DEMC where both intra and inter- cluster communication are being done with additional usage normal nodes as anchor nodes, which also do imbalanced energy consumption .

V. CONCLUSION

This paper presents a novel scheme for wireless sensor networks. It is a fixed mobility based reactive protocol using clustering based on the amount of sensed data and residual energy. The motive behind this cluster formation was to extend network lifetime by reducing the intra-cluster and eliminating inter-cluster communications. Results show that

MSDGP achieved less energy consumption and provided extended network lifetime through implementation of single message CH selection process and by introduction of mobile sink instead of static sink.

VI. Future Work

As a future work, QoS operations and data recovery mechanism are to be implemented into the proposed scheme, and compare it with more protocols and evaluate its performance using new performance metrics involving QoS parameters like throughput, fault tolerance, coverage areas and will also introduction hybrid approach by combing nodes mobility and more fixed Sinks.

References

[1] Khan AW, Abdullah AH, Anisi MH, Bangash JI. A Comprehensive Study of Data Collection Schemes Using Mobile Sinks in Wireless Sensor Networks. Sensors (Basel, Switzerland). 2014;14(2):2510-2548. doi:10.3390/s140202510.

[2] Hyunjo Lee; Miyoung Jang; Jae-Woo Chang;," A New Energy-Efficient Cluster-Based Routing Protocol Using a Representative Path in Wireless Sensor Networks", International Journal of Distributed Sensor Networks Volume 2014, Article ID 527928, 12 pages.

[3] Chen, T.S.; Tsai, H.W.; Chang, Y.H.; Chen, T.C. 2013. Geographic converge cast using mobile sink in wireless sensor networks. Comput. Commun. 2013, 36, 445–458.

[4] Nazir, B.; Hasbullah, H. 2010. Mobile Sink Based Routing Protocol (MSRP) for Prolonging Network Lifetime in Clustered Wireless Sensor Network. In Proceedings of the IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE), Kuala Lumpur, Malaysia, 5–8 December 2010; pp. 624–629.

[5] Alhasanat, A.I., Matrouk, K.D., Alasha'ary, H.A. and Al-Qadi, Z.A. (2014) Connectivity-Based Data Gathering with Path-Constrained Mobile Sink in Wireless Sensor Networks. Wireless Sensor Network, 6, 118-128.

[6] Ilkyu H.; Mamurjon D.; Byoungchul A. "An Energy-Efficient Data Collection Method for Wireless Multimedia Sensor Networks", International Journal of Distributed Sensor Networks, Volume 2014 (2014), Article ID 698452, 8 pages.

[7] Zhongyu Zhang; Zunwen He; Jianguang Jia; Cunxiang Chen, "A new data gathering scheme for trajectory constrained mobile sink in WSN," Wireless Communications & Signal Processing (WCSP), 2012 International Conference on , vol., no., pp.1,6, 25-27 Oct. 2012.

[8] Edgar H, and Callaway J. 2003. Wireless Sensor Networks: Architectures and Protocols: a CRC press company, Auerbach publications.

[9] P. Madhumathy; D. Sivakumar;," Reliable Data Gathering by Mobile Sink for Wireless Sensor Networks", International Conference on Communication and Signal Processing, IEEE April 3-5, 2014, 978-1-4799-3358-7114.

[10] S.Ali,S.Madani "Distributed Efficient Multi Hop Clustering Protocol for Mobile Sensor Networks" , The International Arab Journal of Information Technology 8, no. 3 , 2011.

[11] Kotsilieris, T.C.; Karetsos, G.T. 2013. "Prolonging the lifetime of two-tiered wireless sensor networks with mobile relays". ISRN Sens. Netw. 2013, doi: 10.1155/2013/610796.

[12] Shi, L.; Zhang, B.; Mouftah, H.T.; Ma, J., 2013, DDRP: An efficient data-driven routing protocol for wireless sensor networks with mobile sinks. Int. J. Commun. Syst. 2013, 26, 1341–1355.

[13] Tang, B.; Wang, J.; Geng, X.; Zheng, Y.; Kim, J.U. 2012. "A novel data retrieving mechanism in wireless sensor networks with path-limited mobile sink". Int. J. Grid Distrib. Comput. vol.5, pp.133–140

[14] Aioffi, W.M.; Valle, C.A.; Mateus, G.R.; da Cunha, A.S. "Balancing message delivery latency and network lifetime through an integrated model for clustering and routing in wireless sensor networks". Comput. Netw. 2011, 55, 2803–2820.

[15] Jin Wang; Xiaoqin Yang; Zhongqi Zhang; Liwu Zuo; Jeong-Uk Kim, "Energy Efficient Routing Algorithm for Wireless Sensor Networks Supporting Mobile Sinks ", Advanced Science and Technology Letters Vol.49 (SoftTech 2014), pp.262-268

[16] Tacconi, D.; Miorandi, D.; Carreras, I.; Chiti, F.; Fantacci, R. 2010, Using wireless sensor networks to support intelligent transportation systems. Ad Hoc Netw. 2010, 8, 462–473.

[17] Tian, K.; Zhang, B.; Huang, K.; Ma, J., 2010. Data Gathering Protocols for Wireless Sensor Networks with Mobile Sinks. Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM 2010), pp. 1–6.

[18] Banerjee, T.; Xie, B.; Jun, J.H.; Agrawal, D.P. 2010. Increasing lifetime of wireless sensor networks using controllable mobile cluster heads. Wirel. Commun. Mob. Comput. 2010, 10, 313–336.

[19] Drytkiewicz W; Sroka S; Handziski W; Köpke H; Karl H; " A Mobility Framework for OMNeT++"

[20] Witold Drytkiewicz, Steffen Sroka, Vlado Handziski, Andreas Köpke, and Holger Karl. Mobilty framework website. http://www-tkn.ee.tu-berlin.de/research/ mobility-omnetpp-sim.

[21] Debroy Bijan Kumar, Sheikh Sadi and Md. Al Imran, "An Efficient Approach to Select Cluster Head in Wireless Sensor Networks", Journal of Communication, Vol. 6, No. 7, (October 2011).

[22] A. Kansal, A.A. Somasundara, D.D. Jea, M.B. Srivastava, and D. Estrin. "Intelligent fluid infrastructure for embedded networks".Proceedings of the 2nd International Conference on Mobile Systems, Applications, and Services (MobiSys'04), Boston, MA, USA, pp. 111–124, 2004.

A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools

D. P. Acharjya
School of Computing Science and Engineering
VIT University
Vellore, India 632014

Kauser Ahmed P
School of Computing Science and Engineering
VIT University
Vellore, India 632014

***Abstract*—A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.**

***Keywords*—Big data analytics; Hadoop; Massive data; Structured data; Unstructured Data**

I. Introduction

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets. In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing applications. These are available in structured, semi-structured, and unstructured format in petabytes and beyond. Formally, it is defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi-structured etc. The fourth V refers to veracity that includes availability and accountability. The prime objective of big data analysis is to process data of high volume, velocity, variety, and veracity using various traditional and computational intelligent techniques [1]. Some of these extraction methods for obtaining helpful information was discussed by Gandomi and Haider [2]. The following Figure 1 refers to the definition of big data. However exact definition for big data is not defined and there is a believe that it is problem specific. This will help us in obtaining enhanced decision making, insight discovery and optimization while being innovative and cost-effective.

It is expected that the growth of big data is estimated to reach 25 billion by 2015 [3]. From the perspective of the information and communication technology, big data is a robust impetus to the next generation of information technology industries [4], which are broadly built on the third platform, mainly referring to big data, cloud computing, internet of things, and social business. Generally, Data warehouses have been used to manage the large dataset. In this case extracting the precise knowledge from the available big data is a foremost issue. Most of the presented approaches in data mining are not usually able to handle the large datasets successfully. The key problem in the analysis of big data is the lack of coordination between database systems as well as with analysis tools such as data mining and statistical analysis. These challenges generally arise when we wish to perform knowledge discovery and representation for its practical applications. A fundamental problem is how to quantitatively describe the essential characteristics of big data. There is a need for epistemological implications in describing data revolution [5]. Additionally, the study on complexity theory of big data will help understand essential characteristics and formation of complex patterns in big data, simplify its representation, gets better knowledge abstraction, and guide the design of computing models and algorithms on big data [4]. Much research was carried out by various researchers on big data and its trends [6], [7], [8].

However, it is to be noted that all data available in the form of big data are not useful for analysis or decision making process. Industry and academia are interested in disseminating the findings of big data. This paper focuses on challenges in big data and its available techniques. Additionally, we state open research issues in big data. So, to elaborate this, the paper is divided into following sections. Sections 2 deals with challenges that arise during fine tuning of big data. Section 3 furnishes the open research issues that will help us to process big data and extract useful knowledge from it. Section 4 provides an insight to big data tools and techniques. Conclusion remarks are provided in section 5 to summarize outcomes.

II. Challenges in Big Data Analytics

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet search indexing includes ISI, IEEE Xplorer, Scopus, Thomson

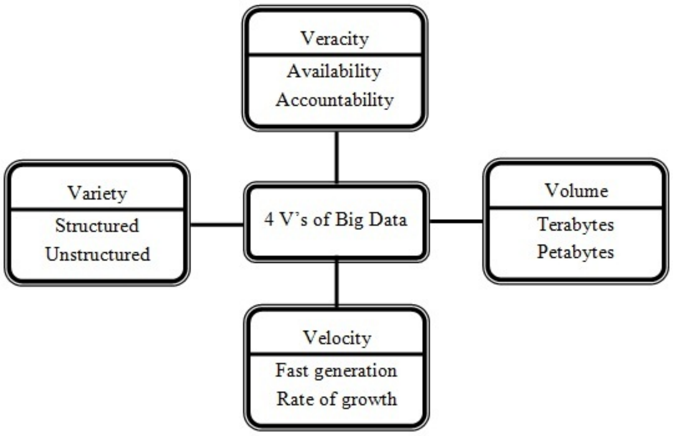


Fig. 1: Characteristics of Big Data

Reuters etc. Considering this advantages of big data it provides a new opportunities in the knowledge processing tasks for the upcoming researchers. However oppotunities always follow some challenges.

To handle the challenges we need to know various computational complexities, information security, and computational method, to analyze big data. For example, many statistical methods that perform well for small data size do not scale to voluminous data. Similarly, many computational techniques that perform well for small data face significant challenges in analyzing big data. Various challenges that the health sector face was being researched by much researchers [9], [10]. Here the challenges of big data analytics are classified into four broad categories namely data storage and analysis; knowledge discovery and computational complexities; scalability and visualization of data; and information security. We discuss these issues briefly in the following subsections.

A. Data Storage and Analysis

In recent years the size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally becuase there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the top priority for the knowledge discovery and representation. The prime reason is being that, it must be accessed easily and promptly for further analysis. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the avialable storage technologies cannot possess the required performance for processing big data.

Another challenge with Big Data analysis is attributed to diversity of data. with the ever growing of datasets, data mining tasks has significantly increased. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets. This presents an unprecedented challenge for researchers. It is becuase, existing algorithms may not always respond in an adequate time when dealing with these high dimensional data. Automation of this process and developing new machine learning algorithms to ensure consistency is a major challenge in recent years. In addition to all these Clustering of large datasets that help in analyzing the big data is of prime concern [11]. Recent technologies such as hadoop and mapReduce make it possible to collect large amount of semi structured and unstructured data in a reasonable amount of time. The key engineering challenge is how to effectively analyze these data for obtaining better knowledge. A standard process to this end is to transform the semi structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge. A framework to analyze data was discussed by Das and Kumar [12]. Similarly detail explanation of data analysis for public tweets was also discussed by Das et al in their paper [13].

The major challenge in this case is to pay more attention for designing storage sytems and to elevate efficient data analysis tool that provide guarantees on the output when the data comes from different sources. Furthermore, design of machine learning algorithms to analyze data is essential for improving efficiency and scalability.

B. Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issue in big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. There are several tools for knowledge discovery and representation such as fuzzy set [14], rough set [15], soft set [16], near set [17], formal concept analysis [18], principal component analysis [19] etc to name a few. Additionally many hybridized techniques are also developed to process real life problems. All these techniques are problem dependent. Further some of these techniques may not be suitable for large datasets in a sequential computer. At the same time some of the techniques has good characteristics of scalability over parallel computer. Since the size of big data keeps increasing exponentially, the available tools may not be efficient to process these data for obtaining meaningful information. The most popular approach in case of larage dataset management is data warehouses and data marts. Data warehouse is mainly responsible to store data that are sourced from operational systems whereas data mart is based on a data warehouse and facilitates analysis.

Analysis of large dataset requires more computational complexities. The major issue is to handle inconsistencies and uncertainty present in the datasets. In general, systematic modeling of the computational complexity is used. It may be difficult to establish a comprehensive mathematical system that is broadly applicable to Big Data. But a domain specific data analytics can be done easily by understanding the particular complexities. A series of such development could simulate big data analytics for different areas. Much research and survey has been carried out in this direction using machine learning techniques with the least memory requirements. The basic objective in these research is to minimize computational cost processing and complexities [20], [21], [22].

However, current big data analysis tools have poor performance in handling computational complexities, uncertainty, and inconsistencies. It leads to a great challenge to develop techniques and technologies that can deal computational complexity, uncertainty,and inconsistencies in a effective manner.

C. Scalability and Visualization of Data

The most important challenge for big data analysis techniques is its scalability and security. In the last decades researchers have paid attentions to accelerate data analysis and its speed up processors followed by Moore's Law. For the former, it is necessary to develop sampling, on-line, and multiresolution analysis techniques. Incremental techniques have good scalability property in the aspect of big data analysis. As the data size is scaling much faster than CPU speeds, there is a natural dramatic shift in processor technology being embedded with increasing number of cores [23]. This shift in processors leads to the development of parallel computing. Real time applications like navigation, social networks, finance, internet search, timeliness etc. requires parallel computing.

The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpretation. However, online marketplace like flipkart, amazon, e-bay have millions of users and billions of goods to sold each month. This generates a lot of data. To this end, some company uses a tool Tableau for big data visualization. It has capability to transform large and complex data into intuitive pictures. This help employees of a company to visualize search relevance, monitor latest customer feeback, and their sentiment analysis. However, current big data visualization tools mostly have poor performances in functionalities, scalability, and response in time.

We can observe that big data have produced many challenges for the developments of the hardware and software which leads to parallel computing, cloud computing, distributed computing, visualization process, scalability. To overcome this issue, we need to correlate more mathematical models to computer science.

D. Information Security

In big data analysis massive amount of data are correlated, analyzed, and mined for meaningful patterns. All organizations have different policies to safe guard their sensitive information. Preserving sensitive information is a major issue in big data analysis. There is a huge security risk associated with big data [24]. Therefore, information security is becoming a big data analytics problem. Security of big data can be enhanced by using the techniques of authentication, authorization, and encryption. Various security measures that big data applications face are scale of network, variety of different devices, real time security monitoring, and lack of intrusion system [25], [26]. The security challenge caused by big data has attracted the attention of information security. Therefore, attention has to be given to develop a multi level security policy model and prevention system.

Although much research has been carried out to secure big data [25] but it requires lot of improvement. The major challenge is to develop a multi-level security, privacy preserved data model for big data.

III. Open Research Issues in Big Data Analytics

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal processing. Effective integration of technologies and analysis will result in predicting the future drift of events. Main focus of this section is to discuss open research issues in big data analytics. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing. However it is not limited to these issues. More research issues related to health care big data can be found in Husing Kuo et al. paper [9].

A. IoT for Big Data Analytics

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile devices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. In a broader sense, just like the internet, Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Conversely, it is still mystifying to understand IoT well, including definitions, content and differences from other similar concepts. Several diversified technologies such as computational intelligence, and big-data can be incorporated together to improve the data management and knowledge discovery of large scale automation applications. Much research in this direction has been carried out by Mishra, Lin and Chang [27].

Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing. Therefore, it is essential to develop infrastructure to analyze the IoT data. An IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Understanding these streams of data generated from IoT devices and analysing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective. Key technologies that are associated with IoT are also discussed in many research papers [28]. Figure 2 depicts an overview of IoT big data and knowledge discovery process.

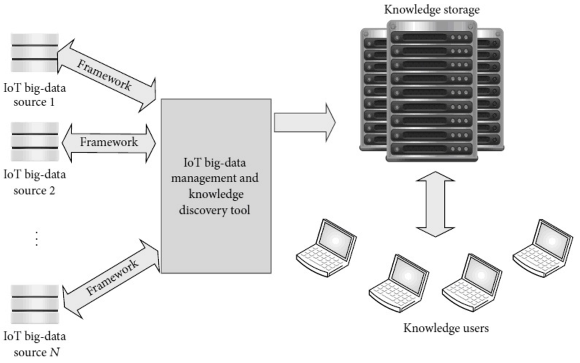


Fig. 2: IoT Big Data Knowledge Discovery

Knowledge exploration system have originated from theories of human information processing such as frames, rules, tagging, and semantic networks. In general, it consists of four segments such as knowledge acquisition, knowledge base, knowledge dissemination, and knowledge application. In knowledge acquisition phase, knowledge is discovered by using various traditional and computational intelligence techniques. The discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge. Knowledge dissemination is important for obtaining meaningful information from the knowledge base. Knowledge extraction is a process that searches documents, knowledge within documents as well as knowledge bases. The final phase is to apply discovered knowledge in various applications. It is the ultimate goal of knowledge discovery. The knowledge exploration system is necessarily iterative with the judgement of knowledge application. There are many issues, discussions, and researches in this area of knowledge exploration. It is beyond scope of this survey paper. For better visualization, knowledge exploration system is depicted in Figure 3.

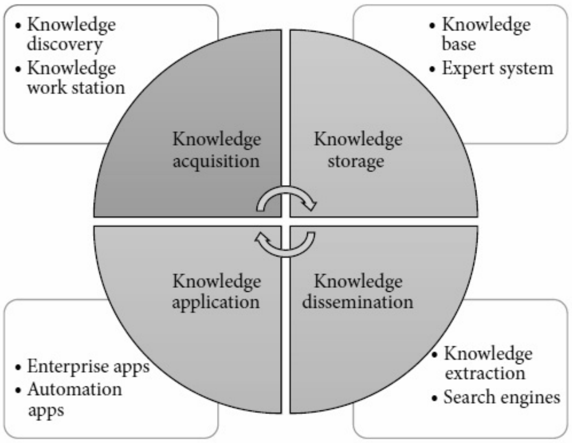


Fig. 3: IoT Knowledge Exploration System

B. Cloud Computing for Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on-demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big data and cloud computing are discussed in detail by many researchers which highlights the challenges in data management, data variety and velocity, data storage, data processing, and resource management [29], [30]. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools.

Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts to interactively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results. This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques.

Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the marketplace and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software as a service (SaaS) from a whole crew of companies such as NetSuite, Cloud9, Jobscience etc. Another advantage of cloud computing is cloud storage which provides a possible way for storing big data. The obvious one is the time and cost that are needed to upload and download big data in the cloud environment. Else, it becomes difficult to control the distribution of computation and the underlying hardware. But, the major issues are privacy concerns relating to the hosting of data on public servers, and the storage of data from human studies. All these issues will take big data and cloud computing to a high level of development.

C. Bio-inspired Computing for Big Data Analytics

Bio-inspired computing is a technique inspired ny nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications.

Huge amount of data are generated from variety of resources across the web since the digitization. Analyzing these data and categorizing into text, image and video etc will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio inspired computing etc whereas equilibrium of data can be done only by selecting right platform to analyze large and furnish cost effective results.

Bio-inspired computing techniques serve as a key role in intelligent data analysis and its application to big data. These algorithms help in performing data mining for large datasets due to its optimization application. The most advantage is its simplicity and their rapid concergence to optimal solution [31] while solving service provision problems. Some applications to this end using bio inspired computing was discussed in detail by Cheng et al [32]. From the discussions, we can observe that the bio-inspired computing models provide smarter interactions, inevitable data losses, and help is handling ambiguities. Hence, it is believed that in future bio-inspired computing may help in handling big data to a large extent.

D. Quantum Computing for Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously [33]. This exponential improvement in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building quantum computer could soon be possible. Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave quantumly. For example, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale quantum computers compared with classical computers. Hence it is a challenge for this generation to built a quantum computer and facilitate quantum computing to solve big data problems.

IV. Tools for big data processing

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analyzing big data with emphasis on three important emerging tools namely MapReduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, stream processing,and interactive analysis. Most batch processing tools are based on the Apache Hadoop infrastructure such as Mahout and Dryad. Stream data applications are mostly used for real time analytic. Some examples of large scale streaming platform are Strom and Splunk. The interactive analysis process allow users to directly interact in real time for their own analysis. For example Dremel and Apache Drill are the big data platforms that support interactive analysis. These tools help us in developing the big data projects. A fabulous list of big data tools and techniques is also discussed by much researchers [6], [34]. The typical work flow of big data project discussed by Huang et al is highlighted in this section [35] and is depicted in Figure 4.

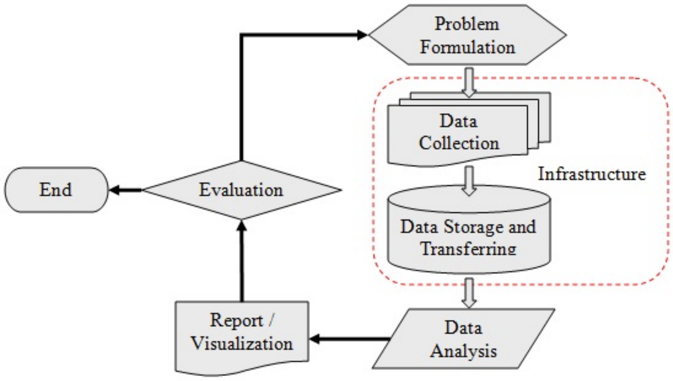


Fig. 4: Workflow of Big Data Project

A. Apache Hadoop and MapReduce

The most established software platform for big data analysis is Apache Hadoop and Mapreduce. It consists of hadoop kernel, mapreduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the subproblems in reduce step. Moreover, Hadoop and MapReduce works as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

B. Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionalty reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. The basic objective of Apache mahout is to provide a tool for elleviating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and facebook [36].

C. Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics.

It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala, or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing hadoop clusters. Figure 5 depicts the architecture diagram of Apache Spark. The various features of Apache Spark are listed below:

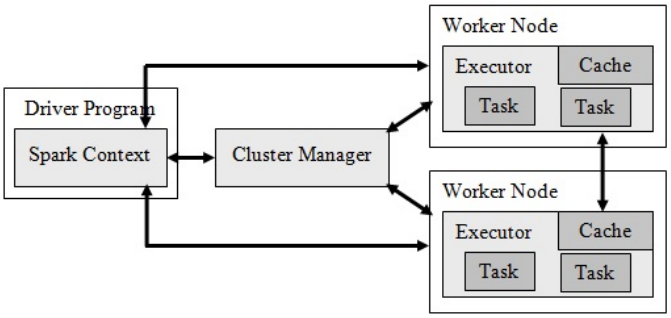


Fig. 5: Architecture of Apache Spark

- The prime focus of spark includes resilient distributed datasets (RDD), which store data in-memory and provide fault tolerance without replication. It supports iterative computation, improves speed and resource utilization.

- The foremost advantage is that in addition to MapReduce, it also supports streaming data, machine learning, and graph algorithms.

- Another advantage is that, a user can run the application program in different languages such as Java, R, Python, or Scala. This is possible as it comes with higher-level libraries for advanced analytics. These standard libraries increase developer productivity and can be seamlessly combined to create complex workflows.

- Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. It is possible because of the reduction in number of read or write operations to disk.

- It is written in scala programming language and runs on java virtual machine (JVM) environment. Additionally, it upports java, python and R for developing applications using Spark.

D. Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and an user use the resources of a computer cluster to run their program in a distributed way. Indeed, a dryad user use thousands of machines, each of them with multiple processors or cores. The major advantage is that users do not need to know anything about concurrent programming. A dryad application runs a computational directed graph that is composed of computational vertices and communication channels. Therefore, dryad provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invokinguser defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices [37].

E. Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The storm cluster is apparently similar to hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic difference is that map reduce job eventually finishes whereas a topology processes messages all the time, or until user terminate it. A storm cluster consists of two kinds of nodes such as master node and worker node. The master node and worker node implement two kinds of roles such as nimbus and supervisor respectively. The two roles have similar functions in accordance with jobtracker and tasktracker of map reduce framework. Nimbus is in charge of distributing code across the storm cluster, scheduling and assigning tasks to worker nodes, and monitoring the whole system. The supervisor complies tasks as assigned to them by nimbus. In addition, it start and terminate the process as necessary based on the instructions of nimbus. The whole computational technology is partitioned and distributed to a number of worker processes and each worker process implements a part of the topology.

F. Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also it has an objective to scale up on 10,000 servers or more and reaches the capability to process patabytes of data and trillions of records in seconds. Drill use HDFS for storage and map reduce to perform batch analysis.

G. Jaspersoft

The Jaspersoft package is an open source software that produce reports from database columns. It is a scalable big data analytical platform and has a capability of fast data visualization on popular storage platforms, including MangoDB, Cassandra, Redis etc. One important property of Jaspersoft is that it can quickly explore big data without extraction, transformation, and loading (ETL). In addition to this, it also have an ability to build powerful hypertext markup language (HTML) reports and dashboards interactively and directly from big data store without ETL requirement. These generated reports can be shared with anyone inside or outside user's organization.

H. Splunk

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their machine generated data through web interface. The results are exhibited in an intuitive way such as graphs, reports, and alerts. Splunk is different from other stream processing tools. Its peculiarities include indexing structured, unstructured machine generated data, real-time searching, reporting analytical results, and dashboards. The most important objective of Splunk is to provide metrices for many application, diagnose problems for system and information technology infrastructures, and intelligent support for business operations.

V. Suggestions for Future Work

The amount of data collected from various applications all over the world across a wide variety of fields today is expected to double every two years. It has no utility unless these are analyzed to get useful information. This necessitates the development of techniques which can be used to facilitate big data analysis. The development of powerful computers is a boon to implement these techniques leading to automated systems. The transformation of data into knowledge is by no means an easy task for high performance large-scale data processing, including exploiting parallelism of current and upcoming computer architectures for data mining. Moreover, these data may involve uncertainty in many different forms. Many different models like fuzzy sets, rough sets, soft sets, neural networks, their generalizations and hybrid models obtained by combining two or more of these models have been found to be fruitful in representing data. These models are also very much fruitful for analysis. More often than not, big data are reduced to include only the important characteristics necessary from a particular study point of view or depending upon the application area. So, reduction techniques have been developed. Often the data collected have missing values. These values need to be generated or the tuples having these missing values are eliminated from the data set before analysis. More importantly, these new challenges may comprise, sometimes even deteriorate, the performance, efficiency and scalability of the dedicated data intensive computing systems. The later approach sometimes leads to loss of information and hence not preferred. This brings up many research issues in the industry and research community in forms of capturing and accessing data effectively. In addition, fast processing while achieving high performance and high throughput, and storing it efficiently for future use is another issue. Further, programming for big data analysis is an important challenging issue. Expressing data access requirements of applications and designing programming language abstractions to exploit parallelism are an immediate need [38].

Additionally, machine learning concepts and tools are gaining popularity among researchers to facilitate meaningful results from these concepts. Research in the area of machine learning for big data has focused on data processing, algorithm implementation, and optimization. Many of the machine learning tools for big data are started recently needs drastic change to adopt it. We argue that while each of the tools has their advantages and limitations, more efficient tools can be developed for dealing with problems inherent to big data. The efficient tools to be developed must have provision to handle noisy and imbalance data, uncertainty and inconsistency, and missing values.

VI. Conclusion

In recent years data are generated at a dramatic pace. Analyzing these data is challenging for a general man. To this end in this paper, we survey the various research issues, challenges, and tools used to analyze these big data. From this survey, it is understood that every big data platform has its individual focus. Some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing. We belive that in future researchers will pay more attention to these techniques to solve problems of big data effectively and efficiently.

References

[1] M. K.Kakhani, S. Kakhani and S. R.Biradar, *Research issues in big data analytics*, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.

[2] A. Gandomi and M. Haider, *Beyond the hype: Big data concepts, methods, and analytics*, International Journal of Information Management, 35(2) (2015), pp.137-144.

[3] C. Lynch, *Big data: How do your data grow?*, Nature, 455 (2008), pp.28-29.

[4] X. Jin, B. W.Wah, X. Cheng and Y. Wang, *Significance and challenges of big data research*, Big Data Research, 2(2) (2015), pp.59-64.

[5] R. Kitchin, *Big Data, new epistemologies and paradigm shifts*, Big Data Society, 1(1) (2014), pp.1-12.

[6] C. L. Philip, Q. Chen and C. Y. Zhang, *Data-intensive applications, challenges, techniques and technologies: A survey on big data*, Information Sciences, 275 (2014), pp.314-347.

[7] K. Kambatla, G. Kollias, V. Kumar and A. Gram, *Trends in big data analytics*, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.

[8] S. Del. Rio, V. Lopez, J. M. Bentez and F. Herrera, *On the use of mapreduce for imbalanced big data using random forest*, Information Sciences, 285 (2014), pp.112-137.

[9] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki and D. K. Grunwell, *Health big data analytics: current perspectives, challenges and potential solutions*, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.

[10] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese, *A look at challenges and opportunities of big data analytics in healthcare*, IEEE International Conference on Big Data, 2013, pp.17-22.

[11] Z. Huang, *A fast clustering algorithm to cluster very large categorical data sets in data mining*, SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[12] T. K. Das and P. M. Kumar, *Big data analytics: A framework for unstructured data analysis*, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.

[13] T. K. Das, D. P. Acharjya and M. R. Patra, *Opinion mining about a product by analyzing public tweets in twitter*, International Conference on Computer Communication and Informatics, 2014.

[14] L. A. Zadeh, *Fuzzy sets*, Information and Control, 8 (1965), pp.338-353.

[15] Z. Pawlak, *Rough sets*, International Journal of Computer Information Science, 11 (1982), pp.341-356.

[16] D. Molodtsov, *Soft set theory first results*, Computers and Mathematics with Aplications, 37(4/5) (1999), pp.19-31.

[17] J. F.Peters, *Near sets. General theory about nearness of objects*, Applied Mathematical Sciences, 1(53) (2007), pp.2609-2629.

[18] R. Wille, *Formal concept analysis as mathematical theory of concept and concept hierarchies*, Lecture Notes in Artificial Intelligence, 3626 (2005), pp.1-33.

[19] I. T.Jolliffe, *Principal Component Analysis*, Springer, New York, 2002.

[20] O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, *Efficient machine learning for big data: A review*, Big Data Research, 2(3) (2015), pp.87-93.

[21] Changwon. Y, Luis. Ramirez and Juan. Liuzzi, *Big data analysis using modern statistical and machine learning methods in medicine*, International Neurourology Journal, 18 (2014), pp.50-57.

[22] P. Singh and B. Suri, *Quality assessment of data using statistical and machine learning methods*. L. C.Jain, H. S.Behera, J. K.Mandal and D. P.Mohapatra (eds.), Computational Intelligence in Data Mining, 2 (2014), pp. 89-97.

[23] A. Jacobs, *The pathologies of big data*, Communications of the ACM, 52(8) (2009), pp.36-44.

[24] H. Zhu, Z. Xu and Y. Huang, *Research on the security technology of big data information*, International Conference on Information Technology and Management Innovation, 2015, pp.1041-1044.

[25] Z. Hongjun, H. Wenning, H. Dengchao and M. Yuxing, *Survey of research on information security in big data*, Congresso da sociedada Brasileira de Computacao, 2014, pp.1-6.

[26] I. Merelli, H. Perez-sanchez, S. Gesing and D. D.Agostino, *Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives*, BioMed Research International, 2014, (2014), pp.1-13.

[27] N. Mishra, C. Lin and H. Chang, *A cognitive adopted framework for iot big data management and knowledge discovery prospective*, International Journal of Distributed Sensor Networks, 2015, (2015), pp. 1-13

[28] X. Y.Chen and Z. G.Jin, *Research on key technology and applications for internet of things*, Physics Procedia, 33, (2012), pp. 561-566.

[29] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, *Big data computing and clouds: Trends and future directions*, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.

[30] I. A. T. Hashem, I. Yaqoob, N. Badrul Anuar, S. Mokhtar, A. Gani and S. Ullah Khan, *The rise of big data on cloud computing: Review and open research issues*, Information Systems, 47 (2014), pp. 98-115.

[31] L. Wang and J. Shen, *Bioinspired cost-effective access to big data*, International Symposium for Next Generation Infrastructure, 2013, pp.1-7.

[32] C. Shi, Y. Shi, Q. Qin and R. Bai *Swarm intelligence in big data analytics*, H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li and X. Yao (eds.), Intelligent Data Engineering and Automated Learning, 2013, pp.417-426.

[33] M. A. Nielsen and I. L.Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, New York, USA 2000.

[34] M. Herland, T. M. Khoshgoftaar and R. Wald, *A review of data mining using big data in health informatics*, Journal of Big Data, 1(2) (2014), pp. 1-35.

[35] T. Huang, L. Lan, X. Fang, P. An, J. Min and F. Wang*Promises and challenges of big data computing in health sciences*, Big Data Research, 2(1) (2015), pp. 2-11.

[36] G. Ingersoll, *Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications*, White Paper, IBM Developer Works, (2009), pp. 1-18.

[37] H. Li, G. Fox and J. Qiu, *Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime*, Second International Conference on Cloud and Green Computing, 2012, pp.675-683.

[38] D. P. Acharjya, S. Dehuri and S. Sanyal *Computational Intelligence for Big Data Analysis*, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.

A Cloud-Based Platform for Democratizing and Socializing the Benchmarking Process

Fuad Bajaber
King Abdulaziz University
Jeddah, Saudi Arabia

Amin Shafaat
University of New South Wales
Sydney, Australia

Omar Batarfi
King Abdulaziz University
Jeddah, Saudi Arabia

Radwa Elshawi
Princess Nourah Bint Abdulrahman University
Riyadh, Saudi Arabia

Abdulrahman Altalhi
King Abdulaziz University
Jeddah, Saudi Arabia

Ahmed Barnawi
King Abdulaziz University
Jeddah, Saudi Arabia

Sherif Sakr
King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia
University of New South Wales, Sydney, Australia

Abstract—**Performances evaluation, benchmarking and reproducibility represent significant aspects for evaluating the practical impact of scientific research outcomes in the Computer Science field. In spite of all the benefits (e.g., increasing visibility, boosting impact, improving the research quality) which can be obtained from conducting comprehensive and extensive experimental evaluations or providing reproducible software artifacts and detailed description of experimental setup, the required effort for achieving these goals remains prohibitive. In this article, we present the design and the implementation details of the *Liquid Benchmarking* platform as a social and cloud-based platform for democratizing and socializing the software benchmarking processes. Particularly, the platform facilitates the process of sharing the experimental artifacts (computing resources, datasets, software implementations, benchmarking tasks) as services where the end users can easily design, mashup, execute the experiments and visualize the experimental results with zero installation or configuration efforts. Moreover, the social features of the platform enable the users to share and provide feedback on the results of the executed experiments in a form that can guarantee a transparent scientific crediting process. Finally, we present four benchmarking case studies that have been realized via the Liquid Benchmarking platform in the following domains: XML compression techniques, graph indexing and querying techniques, string similarity join algorithms and reverse K nearest neighbors algorithms.**

Keywords—***Cloud Computing; Benchmarking; Software-as-a-Service, Social Computing***

I. Introduction

Over the last decades, the scientific community has been witnessing a significant increase in the amount of research outlets. In general, one of the important characteristic of the Computer Science research field is that the research outcomes provides artifacts other than the research outlets, in particular, computer software. In principle, the Computer Science scientific community is continuously witnessing claims on performance improvement from the various researchers and publications which have called for the necessity of conducting comprehensive and reproducible experimental assessments and comparisons between *competing alternative* software of approaches, algorithms or entire systems with the objective of evaluating the significance or the practical impact of the reported research contributions. In practice, most of the research outlets usually report results of their experimental evaluation to assess/compare their introduced scientific contributions with the state-of-the-art, however, unfortunately, the accuracy and the quality of such experimental evaluations are usually constrained with various factors including the unavailability of sufficient time or manpower, the unavailability of adequate or standard testing scenarios or any other resource constraints. In addition, it is common that research outlets are usually concentrating on reporting the experimental results of the *sweet spots* of their contribution which can usually affect on the reflection of the actual picture on the real-world scenarios and suffer from file-drawer effect [37]. Furthermore, it is usually very challenging to assess and understand the performance characteristics of the design choices of a specific approach.

Practically, conducting a consistent, independent and comprehensive study of performance evaluation or benchmarking competing alternatives in a specific domain is mostly a resource and time consuming process. Therefore, it is common that the accuracy and the quality of reported experimental results can be limited and constrained with various conditions including the limited time, limited human power, shortage of computing resources and unavailability of publicly accessible software implementation of some contributions that have been reported in the research literature. Moreover, it is practically challenging to get an access to various configuration of computing environments/resources which can represent or cover the wide spectrum of various real-world use cases [31]. Hence, it is, unfortunately, common in many research areas to have little or no objective knowledge about the advantages and limitations of any group of competing research approaches/techniques that are sharing the focus on addressing a specific research problem.

In principle, the ability to repeat experiments is considered a hallmark of the scientific process which is used to confirm or refute hypotheses and previously obtained results [16]. In recent years, the importance of defining and performing comprehensive benchmarking and performance evaluation studies has been acknowledged by various research communities. Additionally, various scientific conferences, funding agencies and publishers have started to motivate the researchers to share the software artifacts and documentations that can facilitate the reproducibility of the experimental results which are reported in their research outlets. For instance, in the database research community, since 2008, the ACM SIGMOD conference, which is considered as the most prestigious conference of the community, has started to offer the chance to verify the reproducibility of the reported experimental results by providing the researchers with the opportunity to submit their software and other experimental artifacts (e.g., datasets) [27]. Moreover, since 2008, another prestigious conference of the commuity, the VLDB conference, initiated a new experimental and analysis paper track that motivates the researchers of community to submit manuscripts that document and report in-depth experimental evaluation and benchmarking studies[1]. Furthermore, various proposals [10] and scientific tutorials demonstrated in the main scientific venues of the database research community have been focusing on promoting the significant importance of reproducibility, performance evaluation and benchmarking studies in database research [18], [25]. As a result, some efforts such as Arizona Database Laboratory (AZDBLab) [41] has been proposed to support database researchers in performing a comprehensive and empirical study among various database management systems. Other research communities followed the same trend such as the Semantic Web[2,3,4], Semantic Web Service[5], Business Process[6], Information Retrieval[7] communities in addition to the general Executable Paper Grand Challenge[8]. In spite the fact that these initiatives for benchmarking efforts and research publications are important and useful, however, the main limitation of these efforts is that they report particular *snapshots* for the state-of-the-art that reflect the status at the time of their conduct. In practice, the state-of-the-art in any research domain is always *evolving* and *dynamic*, by default. For example, emerging or novel approaches or techniques that tackle the same research challenge of a formerly revealed snapshot publication can be proposed or the performance characteristics of formerly assessed approaches or techniques may develop and improve. Hence, this type of research publications can be outdated shortly after they have been released.

In practice, the recent advances in Web technologies (e.g. social web, cloud computing, software-as-service) have provided novel work environments that opened new opportunities to address the above mentioned challenges. As a result, recently, the scientific communities started to increasingly use personal/shared blogs and wikis (e.g. ACM SIGMOD blog[9], DBMS2[10], SemWebTec[11]) to share and discuss their findings. *PubZone*[12] has been designed as a service that provides the scientific community with a Wiki and discussion forum for publications. *crowdLabs*[13] and *myExperiment*[14] have been proposed as environments for sharing workflows that describe computational experiments, data analyses and visualizations. However, in practice, there is a still long way to go for achieving effective and collaborative innovations in the research practices. In particular, surprisingly, the Computer Science research communities have not been successful, so far, to make the best use of or effectively exploit the availability of the recent advances in Web technologies to establish platforms and form driving forces that can address the above mentioned challenges and implement functional and widely-used collaborative experimental evaluation and benchmarking platforms that can dynamically evolve and exploit the power of the crowd.

In this article, we present the design and the implementation details of the *Liquid Benchmarking* platform [35] as a novel research infrastructure that provides cloud-based, collaborative and social environment that attempts to tackle the above mentioned challenges and obstacles by facilitating the *democratization*, *socialization* and improving the quality of the performance evaluation and benchmarking processes in the Computer Science research domain. In particular, we summarize the main contributions of our presented platform as follows:

- The platform can significantly reduce the effort and time for executing performance evaluation experiments by facilitating the process of sharing the experimental artifacts (e.g., software implementations, benchmarking tasks, computing environments) and supporting its end users to easily design, mashup and execute the experiments with zero installation or configuration efforts.
- The platform supports for searching, comparing, analyzing and visualizing the results of previous experiments.
- The users of the platform can subscribe to get notifications about the results of any new running experiments for the domains/benchmarks of their own interests.
- The collaborative and social features of the platform enable turning the performance evaluation and benchmarking process into a *living* process where different users can run different experiments, share the results of their experiments with other users in addition to commenting on the results of the conducted experiments by themselves or by other users of the platform. Such features guarantee the utilization of the *wisdom of the crowd*, the *freshness* of the results, the establishment of a *transparent* process for scientific *crediting* and the development of scientific advances that trust and build on previous research contributions.

In addition, we present the implementation details of four

[1] http://www.vldb.org/pvldb/vol1.html
[2] http://challenge.semanticweb.org/
[3] http://2014.eswc-conferences.org/important-dates/call-challenges
[4] http://iswc2014.semanticweb.org/call-replication-benchmark-data-software-papers
[5] http://sws-challenge.org/wiki/index.php/Main\textunderscorePage
[6] http://processcollections.org/past/2013-2/matching-contest
[7] http://www2.informatik.hu-berlin.de/~wandelt/searchjoincompetition2013/
[8] http://www.executablepapers.com/
[9] http://wp.sigmod.org/
[10] http://www.dbms2.com/
[11] http://semwebtec.wordpress.com/
[12] http://www.pubzone.org
[13] http://www.crowdlabs.org/
[14] http://www.myexperiment.org/home

benchmarking case studies that have been fully realized and made available via the *Liquid Benchmarking* platform. The remainder of this paper is organized as follows. Section II motivates the practicality and significance of our platform by illustrating sample scenarios. Section III discusses some of the fundamental obstacles for conducting trustable and conclusive experimental evaluation or benchmarking research in the Computer Science filed. In Section IV, we describe the main entities and the conceptual model of the *Liquid Benchmarking* platform. The architecture and the implementation details of the platform is presented in Section V. Four Liquid Benchmarking case studies are presented in Section VI before we conclude the paper in Section VII.

II. Motivating Scenarios

In this section, we present two sample scenarios that motivate the practical importance of our presented *Liquid Benchmarking* platform as follows.

Scenario 1: $Alan$ is a graduate student in one of the world reputable research groups on data management systems. He and his advisor are researching on developing novel efficient techniques for querying graph-based biological databases. $Alan$ has been recommended by his advisor to perform a survey on the related literature and conduct an experimental evaluation for assessing the performance characteristics of the state-of-the-art. During this activity, $Alan$ got overwhelmed with a large number of literature which are reporting scientific proposals for techniques and approaches to tackle the problem of interest. As a result of an extensive research task, $Alan$ has been successful on getting the access to the software implementation for some of proposed approaches and techniques in the literatures while he exploited his technical software development skills to re-implement a set of the important approaches which were reported in the literature, according to their reported description, but that have no available software implementations. After a year of effort, $Alan$ prepared all the requirements to conduct a benchmarking study which assesses and compares between some of the proposed techniques for tackling his problem of interest. This experimental evaluation study supported Alan to gain useful insights for achieving his primary objective. Apparently, this is very time and effort consuming task (in addition to some other various obstacles which will be discussed in more details in Section III). In practice, the accuracy and the quality of the outcomes of such conducted experimental evaluation activity has been constrained with the amount of effort, time and attention which has been dedicated by two researchers: $Alan$ and his advisor, through this research activity. In general, it is common that graduate students in the various research areas of the Computer Science field usually go through a similar process at the initial stages of their research work. Unfortunately, prospect graduate students in the same domain, across the world, may not be able to leverage, extend or improve *Alan*'s effort unless there is an effective and workable solution or platform that allow $Alan$ to share the artifacts of his study and enable other students and researchers to collaborate on following up and contributing to this sort of experimental evaluation research. In addition, after one year, $Alan$ may not be able to reproduce his own results or explain them. In practice, constant time pressure and strict submission deadlines usually push the scholars to favor timely results over spending enough time on documenting the experiments and data traceability.

Scenario 2: $John$ is one of the active researchers in the domain of graph databases. During his research, $John$ got interested in benchmarking the state-of-the-art of the indexing and querying approaches for graph databases. As a result, *John* allocated about 24 weeks of his time in the following activities:

- Establishing a large corpus of graph databases that have various characteristics.
- Searching for the available software implementations of graph indexing and querying techniques in addition to implementing some of the techniques that have been presented in the literature but they have no available software implementation.
- Conducting extensive experiments to assess the performance characteristics of various proposed techniques, which were proposed in the literature, using the established collection of graph datasets and analyze their results.
- Documenting the results of the conducted experiments, sharing the artifacts of the study via a public web page and writing up a journal publication that disseminate the results and lessons of his benchmarking study.

Following its release, *John*'s benchmarking study has attracted a lot of interest from the research community of graph databases where some of the active researchers in this domain had communications with $John$ to inquiry about various aspects of the experimental study or seeking some advices in reproducing the results of some of the reported experiments. However, unfortunately, these communications remained offline in *John*'s mail box. After sometime, John has moved to a new position and his research interest shifted to another research area. Hence, he become less responsive to inquiries from researchers in the graph indexing and querying domain about his benchmarking study. In addition, the results of his benchmarking study has become out-of-date after the introduction of novel approaches and techniques that tackle the same problem and the improvement of formerly investigated approaches by John. In practice, effectively and cumulatively exploiting *John*'s effort calls for joint efforts from other active scholar in the graph indexing and querying domain in addition to an the availability of an adequate platform that can facilitate and support such efforts.

III. Benchmarking Challenges in Computer Science

In comparison to more traditional disciplines (e.g., natural sciences), computer science is considered a much younger discipline which is usuall said to have a somewhat sloppy relationship with the repeatability of published results [13]. In practice, each computer science scholar could share a story about of a failed attempt to reproduce the results of a some top-notch paper [13]. For example, Collberg et al [11] have reported that they have failed, in many cases failed, to systematically replicate artifacts from highly ranked research papers. In this section, we discuss some of the remarkable obstacles for conducting trustable and conclusive benchmarking studies in the Computer Science research field as follows.

- *Limited reproducibility of reported experimental results*: In an ideal world of Computer Science research, the authors of a research outlet document the details of their contributions in the manuscript and publicly provide the binaries/source codes of their software implementation with the other related software atrifacts (e.g., experimental datasets) to the other researchers so that they can be exploited for reproducing the reported results in their publication. This ideal process would provide several advantages. For example, other researchers in the same domain of the study would be able to *independently* asses the performance characteristics of the provided software implementation using other experimental setups (e.g., datasets, computing resources) in order to verify the reported claims and make sure that there is no hidden aspects which can affect the accuracy of the reported experimental results. In addition, other researchers can exploit this available software artifacts as a valuable starting point to evaluate and assess the significance of their own proposed contribution. One of the interesting examples for the value of such independent evaluation studies is the study of Sidirourgos et al. [39] where they have reported about an independent assessment of the published result by Abadi et al. in [4] which described an approach for implementing a vertically partitioned DBMS for Semantic Web data management. The outcomes of this independent assessment revealed many interesting aspects. For instance, in [4] Abadi et al. reported that the performance of binary tables is superior to that of the clustered property table for processing RDF queries while Sidirourgos et al. [39] reported that even in column-store database, the performance of binary tables is not always better than clustered property table and depends on the characteristics of the data set. In addition, the experiments of [4] reported that storing RDF data in column-store database is better than that of row-store database while [39] experiments have shown that the gain of performance in column-store database depends on the number of predicates in a data set. A main lesson from this example is that we cannot really be sure that published research results are accurate and comprehensive even if they were reported by the best scientists and went through the most rigorous peer review process. However, can have more confidence on these results if others can repeat the same experiments and obtain similar results [16]. However, it should be noted that such repetitions are considered part of the scientific process and they do not represent any mistrust for the scholars who published the original results. Instead, they represent part of the scientific process which aims of gaining more confidence in the original results or to provide more insights that can specify or delimit the range of their applicability.

  In practice, unfortunately, the research world is not usually following this ideal process. For instance, Sakr [32] has performed a benchmarking study for the state-of-the-art of XML compression techniques [1]. The results of this study have shown that many XML compression techniques which were presented in the literature have no available software implementations and thus it is hard or not straightforward to assess their performance characteristics. Collberg et al. [11] have also reported in their study for highly ranked research papers that when software was available, with a percentage of only 44% of the cases, it was difficult to have it running. Clearly, such limitation prevents the researchers from confirming the reproducibility of the reported figures in the original publications and hinders the chances of conducting comprehensive comparisons among the whole set of the proposed techniques for tackling the same research challenges. Recently, some groups have organized initiatives to establish open challenges in various research domains (e.g. Semantic Web Service Challenge, Semantic Web Challenge, Information Retrieval). In addition, recent editions of SIGMOD conference started to offer the opportunity for the researchers of the published manuscripts to evaluate their software using the experimental datasets to reproduce the reported experimental results. Unfortunately, so far, the repeatability reports of the SIGMOD conference have shown limited success on achieving this goal due to several reasons [7], [26], [27].

- *The dynamics and continuous evolution of the state-of-the-art*: In practice, conducting an independent, comprehensive and conclusive benchmarking study for the-state-of-the-art in any research area is a very useful but also a challenging task which involves considerable time, effort and resources. For example, it may require designing different scenarios, choosing different datasets and evaluating the performance characteristics using various metrics. Therefore, some journals (e.g., the *Elsevier Performance Evaluation* Journal[15]) focus their scope of interest around manuscripts that consider this type of experimental evaluation research. In 2008, the reputable VLDB scientific venue initiated a new experimental analysis research track that focuses on analyzing the advantages and drawbacks of various techniques which are tackling the same research challenge. Although this type of research publications are useful, they suffer from a main limitation that they reflect *snapshots* for the state-of-the-art at the time of their conduct. However, by default, the research contributions in any research area are always *evolving* and *dynamic*. For instance, novel techniques which are designed to address the same research challenge of a formerly published snapshot publication can be proposed or the performance characteristics of formerly evaluated techniques can develop and improve. Hence, these publications may go out-of-date after a relatively short period of their release. Assuming that the results of such benchmarking studies can be maintained on web pages, *continuous* evolving and maintenance of the reported results may require too much effort from the authors who may loose interest in re-executing the same task after sometime. Finally, it is not practically recommended in the current very dynamic environment to spend several years in conducting a set of benchmarking experiments in a certain research domain. In particular, nowadays, the development of such benchmarking studies should be fast, dynamic and reactive in order to be valuable.
- *Constraints on the availability of computing resources*:

[15] http://www.elsevier.com/locate/peva

In various domains, performing conclusive benchmarking study may require huge computing resources. In addition, conducting experimental evaluations may require experimenting with various configurations for the computing environments in order to reflect the various configurations of computing environments in real-world scenarios. In practice, the availability of such computing resources requirements for researchers who are aiming to conduct a benchmarking study in their home labs/environments can be limited which consequently can limit or prevent their capacity to conduct comprehensive and insightful benchmarking studies. For instance, Pavlo et al. [30] described a benchmarking study that compares between the performance and the development complexity of parallel databases and MapReduce in executing *large-scale* data analysis jobs. In practice, reproducing the results of the experiments which has been reported in this publication by other researchers is a very challenging task due to the high and demanding configurations of the the testing computing environment. In particular, the original experiments which have been reported in this publication were conducted using a computing cluster of about hundred machines. In general, conducting a *fair* and *apples-to-apples* comparison among any alternative software implementations would require executing the experiments using *exactly* similar computing environments and the same experimental artifacts. In addition, it is crucial that an experimental evaluation study test the performance characteristics of hardware components and subsystems in a realistic and meaningful way. Therefore, ideally, researchers should have the facility to access shared computing resources where they can compare/evaluate the various software, under study, consistently. The adequate configurations of such experimental computing environments should be also decided *collaboratively*.

- *Not enough standard benchmarks are available or widely-used*: A benchmark is a *standard* test or set of tests which is utilized to compare/evaluate different techniques that have a shared objective to address a certain research challenge. In practice, the unavailability of a standard benchmark in a specific research issue represents a major source of hardship for the researchers who want to comprare/evaluate their contribution in this domain and consequently leads to reporting about various adhoc experimental results in the various publication which documented research efforts that attempted to tackle this research challenge. In principle, a benchmark usually consists of a motivating scenario, a set of benchmarking tasks in addition to specifying a set of performance evaluation metrics. In principle, limited number of benchmarks usually succeed on gaining wide acceptance and achieving good success in their target research community. For instance, in the database research community, some benchmarks were successful on achieving such success including:
  - ○ The TPC group of benchmarks for evaluating the performance characteristics of transaction processing in relational database management systems [3].
  - ○ The oo7 benchmark [8] which has been presented as a standard benchmark for evaluating the performance characteristics of object-oriented database systems.
  - ○ The XML Benchmark Project (XMark) [38] which has been used as a mean to evaluate the performance characteristics of XML data management systems.

  However, on the other hand, there are still many other research aspects in the database research community which are in a significant need for defining standard benchmarks that fulfil the requirements of the researchers in assessing the impact of their contributions (e.g. graph databases, RDF databases, big data processing systems, NoSQL databases, scientific databases) [33], [34], [40]. In practice, for any benchmark to be successful, it needs to gain wide acceptance by its target community. Hence, the motivating scenario of the defined benchmark should be *simple*, the set of testing tasks and performance metrics should be *complete* and *generic* [12]. In addition, such standard benchmarks should satisfy other general and important qualities such as *portability*, *relevance*, *scalability* and *extensibility* [22]. In practice, it is challenging that a single benchmark can reflect the various usage scenarios and achieve all these quality goals. Therefore, it is common that many research domains require defining *microbenchmakrs* [5] that have deep focus in a specified detailed aspects. In principle, a well-designed benchmark in a certain domain is usually very useful to the active researchers in that domain as it constitutes the fundamental basis for evaluation and comparing their research contributions. Therefore, they become able to specify the advantages and disadvantages of their contribution which can effectively inspire their plans for the various directions of improvement. However, designing a successful benchmark is a quite challenging task which is usually not easily achievable by a single researcher or research group. Ideally, effectively tackling the challenge of establishing standard and successful benchmarks would require collaborative efforts from various groups of peer researchers within the target domain of the benchmark.

IV. Conceptual Model

The primary objective of the Liquid Benchmarking platform is to provide a cloud-based and social platform which can simplify and democratize the job of computer science scientific scholars in conducting solid experimental evaluations with high quality. In particular, the features of this platform is designed to provide scientific scholars with various mutual services including:

- Establishing repositories of related and competing software implementations where these implementations can be executed as software services that involve no installation or configuration requirements at the users side.
- Sharing testing computing environments.
- Collaboratively defining, discussing and evolving the specifications of standard benchmarks to assess the competing software implementations.
- Providing the end-users with an environment that supports easily creating and executing testing experiments and share their results.

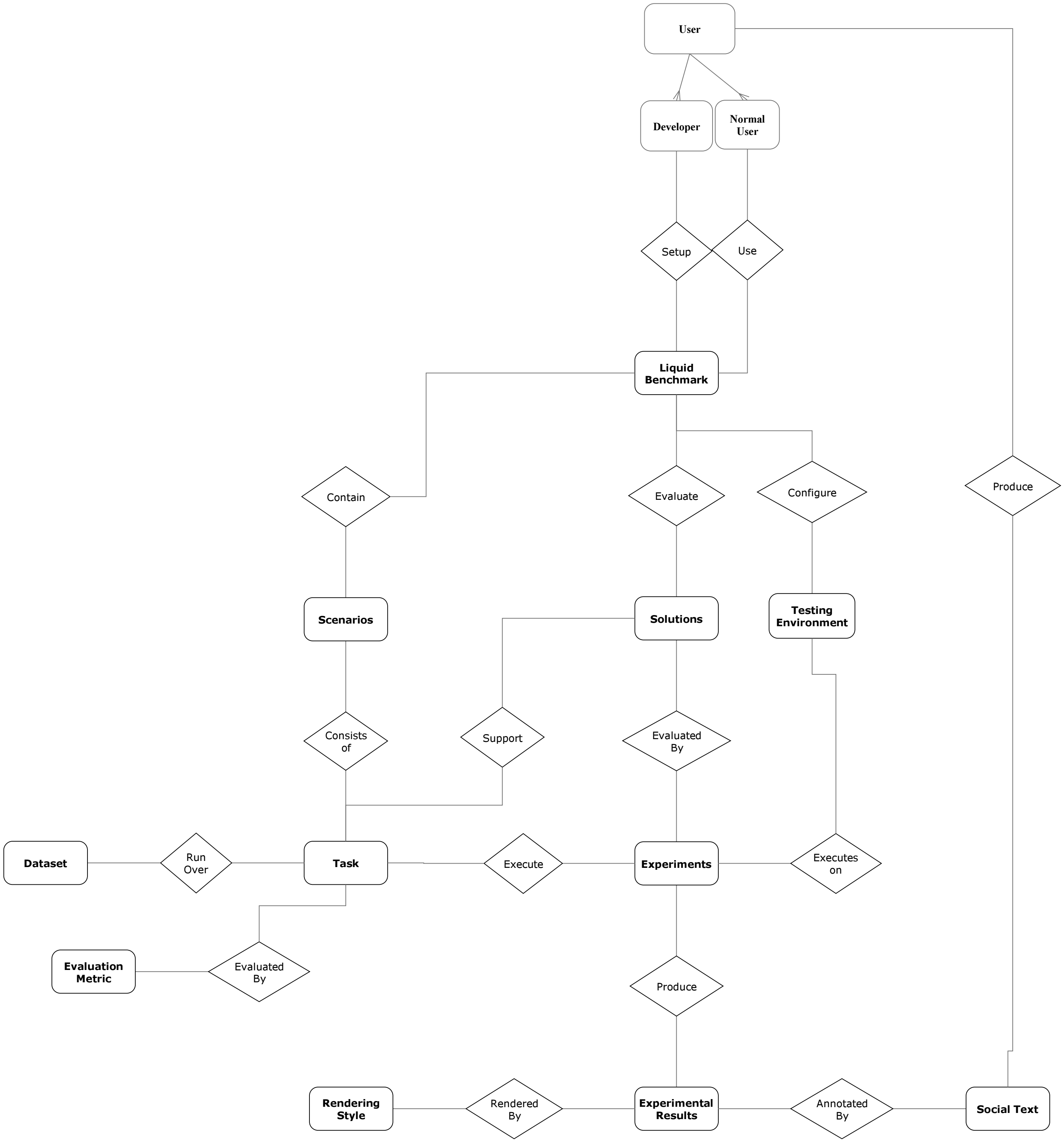


Fig. 1: Conceptual Model of Liquid Benchmarks

Figure 1 illustrates an overview of the conceptual model for the main entities of the Liquid Benchmarks. In this model, we differentiate among two types of users: *developer user* (benchmark developing committee) and *normal user*. Developer users represent the set of researchers who have the privilege to participate in the collaborative environment for defining the configurations of the different components of the benchmark (e.g. datasets, tasks, evaluated software) while normal users are only allowed to use the defined configurations of the benchmark to run their test experiments. However, normal users can be optionally allowed to do some configuration tasks such as: uploading their own datasets or defining their own tasks for running specially defined experiment in a *private* area which is separated from the public setup of the benchmarks. In particular, each liquid benchmark is configured by defining the following main components:

- **Scenarios**: In principle, each liquid benchmark consists of at least one scenario which models a use case that focuses on evaluating some aspects of the competing softwares in the target domain (e.g. MacroBenchmark or MicroBenchmark). In particular, each scenario is described by a *Service Schema* that defines the set of parameters (inputs and outputs) which need to be defined for interfacing with the services of the evaluated softwares.
- **Evaluated Solutions**: The set of competing software implementations (e.g. techniques, algorithms, systems) which are developed to tackle the specific problem of the liquid benchmark. In practice, each software implementation may have various *versions*. Each of these versions is treated as a separate (but linked) competing solution. Each solution need to register the set of its supported tasks in order to avoid the running of many failing tasks.
- **Task(s)**: Describes a set of operations which should be executed by the competing software implementations (e.g. update operations, queries, compressing operations). In practice, each operation usually assesses one or more target evaluation aspects which is in the scope of the benchmark specifications.
- **Metric(s)**: Represents the measures of evaluating the performance characteristics of the competing software implementations in performing the various defined tasks of the benchmark (e.g. execution time, response time, throughput). In particular, metrics represent the basis of comparing the competing software implementations.
- **Testing Environment(s)**: Represents a set of different configurations for computing environments (e.g. operating system, CPU, disk space, main memory) that reflect various real-world scenarios.

V. Platform Architecture and Implementation

In principle, the features and design decisions of the implementation of the Liquid Benchmarking platform[16] combine the facilities provided by different emerging Web technologies which are described as follows:

- *Software-as-a-Service*: The platform uses the RESTful architectural style as an effective software distribution technique [42] in which software implementations can be installed on the hosting computing environment and made available via an application programming interface to the end-users via the Internet. This technology requires zero downloading, installation or configuration effort at the side of the end user where all communication with software can be achieved using HTTP methods [29].
- *Cloud Computing*:
  Benchmarking in practical computer science requires more than just data and code, however, it also requires an appropriate and *shared* or *identical* computing environment in which to run experiments. The platform exploits cloud computing as an emerging effective technology for broad sharing of hardware resources and computing environments over the Internet [15]. In particular, virtualization is a key technology of the cloud computing paradigm which improves the manageability of hardware resources by flexibly allowing computing resources to be provisioned on demand (in the form of virtual machines) and hiding the complexity of resource sharing details from cloud users [36], [6]. In practice, conducting a fair and *apples-to-apples* comparison between any competing software implementations requires performing their experiments using *exactly* the same computing environment [31]. In addition, performing a comprehensive and insightful evaluation process that assess different performance characteristics of the evaluated software implementations may require using several virtual machines with variant and scaling (in terms of computing resources) configuration settings (e.g. main memory, disk storage, CPU speed) that reflect different real-world scenarios [31]. The Liquid Benchmarking platform utilizes the virtualization technology for maintaining the testing computing environments in cloud platforms in the form of pre-configured *virtual machines* (with different configurations) which are hosting the competing software implementations (in the form of web services) and are shared by the end-users of the benchmark.
- *Collaborative and Social Software*: The platform is enabled with different Web 2.0 and social Web capabilities (e.g. tagging, forums, user comments) that support human interaction and facilitates the establishment of online communities between groups of researchers who share the same interests (peers) where they can interact and work together in an effective and productive manner [14]. Most important, the platform supports sharing the performance evaluation and benchmarking artifacts (e.g., software implementations, datasets, virtual machines) in a *workable* environment.

Figure 2 illustrates the architecture of the Liquid Benchmarks platform which are equipped with several *components* that are described as follows:

- **Web-based User Interface**: This component provides the end user with a user-friendly interface where she/he can *mash up* the components (e.g., services, computing environments, tasks, metrics) of the experiment in a *drag and drop* style (Figure 3). In principle, according to the configuration of the components of the liquid benchmark, end users can design and run their *experiments* where each experiment is specified by: the *solution(s)* (software implementation(s)) to be evaluated, the *task* to be executed

[16] The implementation of the Liquid Becnhmarking platform is available onhttp://liquidbenchmark.net:8080/Liquid/

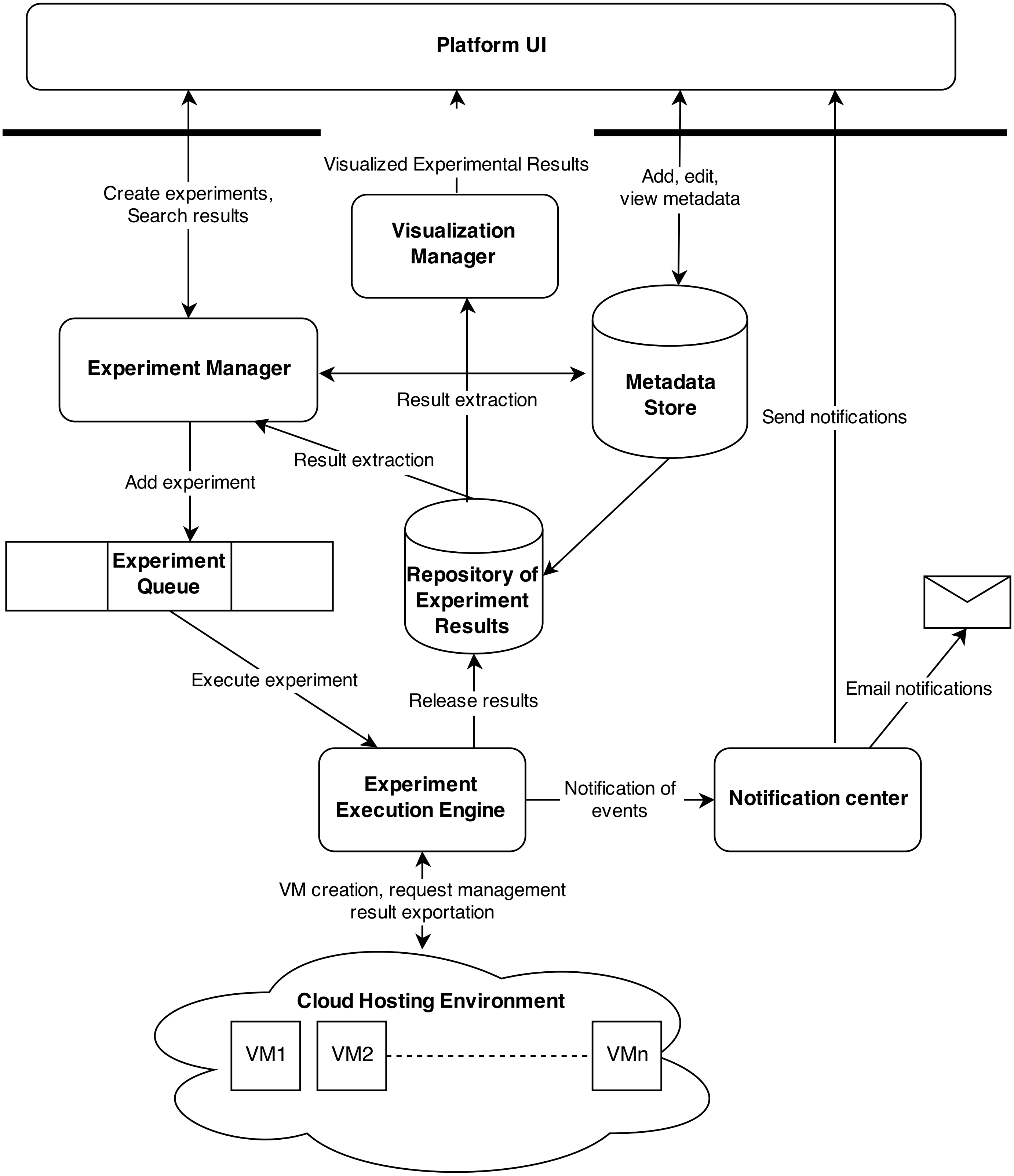
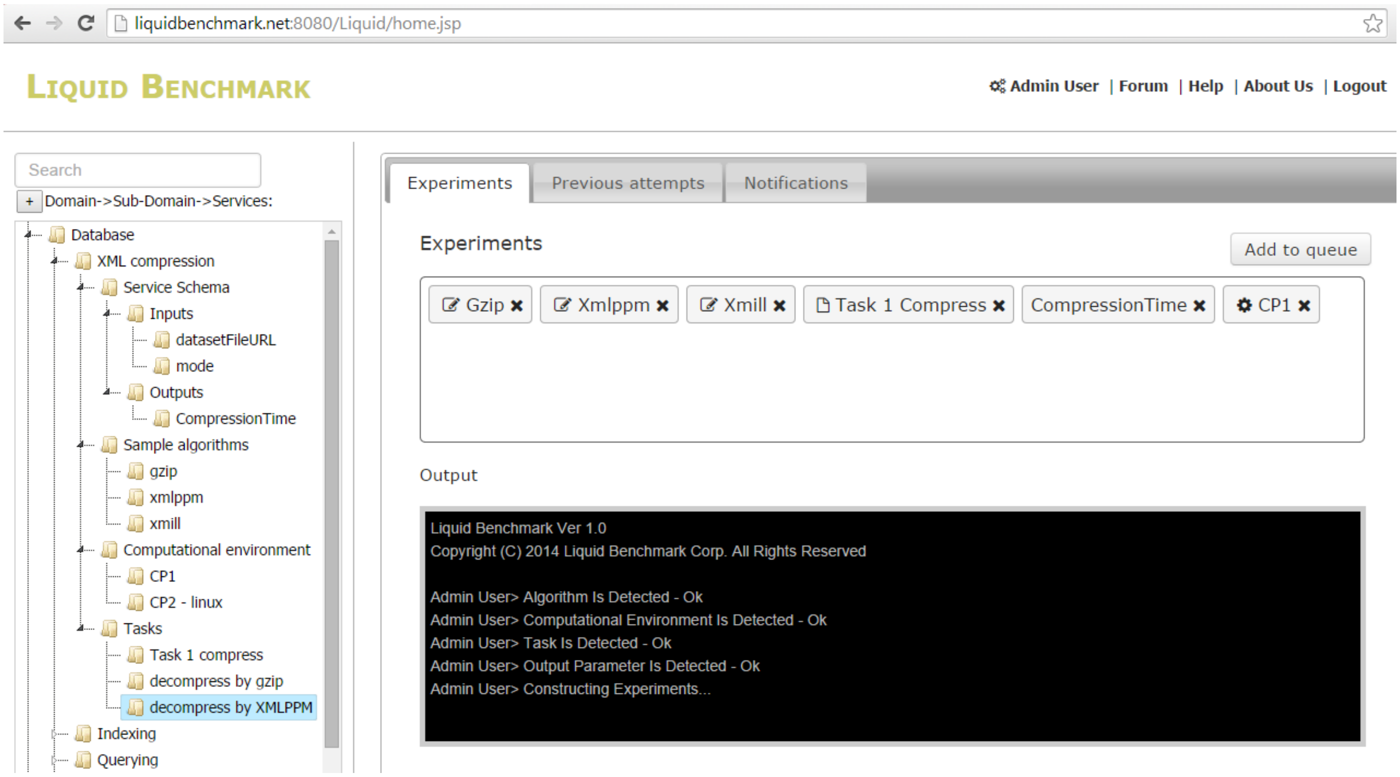


Fig. 2: Platfrom Architecture

with the associated instantiation of the parameters of the service schema, the selected metrics for evaluation and the testing *environment* which will be used for running the experiment. The platform user interface also provides



```
Liquid Benchmark Ver 1.0
Copyright (C) 2014 Liquid Benchmark Corp. All Rights Reserved

Admin User> Algorithm Is Detected - Ok
Admin User> Computational Environment Is Detected - Ok
Admin User> Task Is Detected - Ok
Admin User> Output Parameter Is Detected - Ok
Admin User> Constructing Experiments...
```

Fig. 3: Screenshot: Mashing Up an Experiment

the end-users with other facilities including managing user account, maintaining the metadata store, searching and commenting on the results of previous experiments, subscribing to the results of a benchmark in addition to analyzing and visualizing the experimental results.

- **Metadata Store**: This component stores the information about the various components of the benchmark (e.g., services, service schema, tasks, virtual machines).
- **Experiment Manager**: The experiment manager receives the specification of the user-defined experiment, which is configured by the Liquid Benchmark user interface, an registers this experiment for execution on the **Experiment Queue**. In principle, the experiment queue is used by the **Experiment Execution Engine** to ensure that the execution of one experiment in a testing environment is not going to influence the execution of another experiment in the same environment (an experiment can only start after the end of the current experiment, if exist, on the computing environment). Through the experiment life cycle, the **Experiment Execution Engine** sends a set of *notification events* to the **Notification Center** with the status of the experiment till its completion and storing its results in the **Repository of Experimental Results** for further analysis and visualization purposes. It should be noted that the **Experiment Execution Engine** is the component that is responsible for managing the life cycle of testing environments. In particular, it starts the virtual machine of a testing environment for running an experiment if it has been in a stopped mode or it stops the virtual machine if it has been idle for a while and has no pending experiments in the queue.
- **Repository of Experiment Results**: This is a central repository that stores the results of all experiments associated with their configuration parameters, *provenance* information (e.g. timestamp, user) and social information (e.g. comments, discussions). Clearly, end-users can search and view the contents of this repository to analyze, compare, visualize and comment on the results of the formerly running experiments without taking the time of re-running or creating them from scratch.
- **Visualization Manager**: This component is equipped with a set of *visualization styles* (e.g. line charts, column charts) for presenting and comparing the results (metrics) of the selected experiments by the end-users (Figure 4).

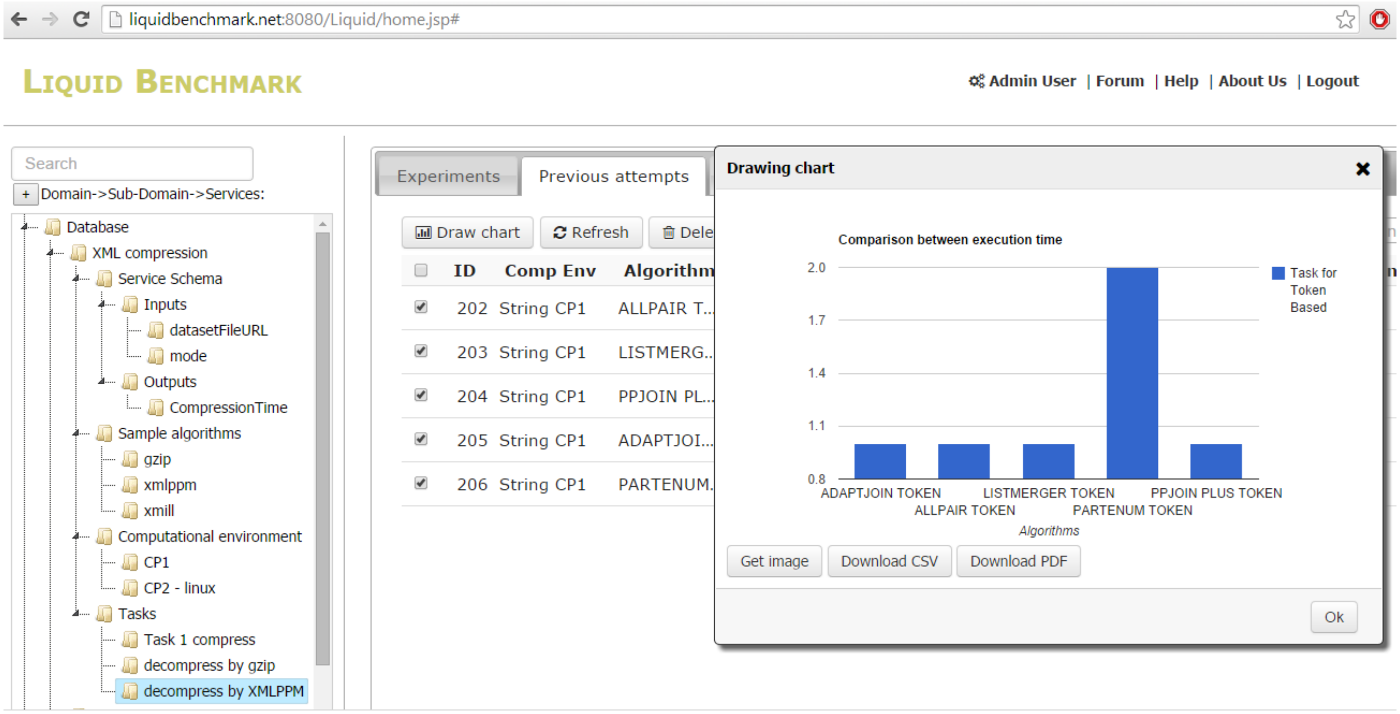
VI. Case Studies

In this section, we present four benchmarking case studies which have been realized using the Liquid Benchmarking platform[17] on the following domains:

- *XML compression*[18]: This case study is based on the benchmark of XML compressors (e.g., XMill [24], Gzip, Bzip, XMLPPM [9]) that has been presented in [32]. In particular, this case study provides services for the implementation of nine XML compression tools with benchmarking tasks over a large XML corpus that covers the different types and scales of XML documents. This case study evaluates the XML compressors by three

[17] The full documentation for using the platform is available on http://wiki.liquidbenchmark.net/

[18] The full documentation and screencast of this case study is available on http://wiki.liquidbenchmark.net/doku.php/casestudy-xmlcompression



□	ID	Comp Env	Algorithm
☑	202	String CP1	ALLPAIR T...
☑	203	String CP1	LISTMERG...
☑	204	String CP1	PPJOIN PL...
☑	205	String CP1	ADAPTJOI...
☑	206	String CP1	PARTENUM...

Fig. 4: Screenshot: Comparing and Visualizing Experimental Results

different metrics: compression time, decompression time and compression ratio.

- *Graph indexing and querying*[19]: This case study implements the *iGraph* framework [19], [20] for evaluating various graph indexing and querying techniques (e.g. Closure-Tree [21], gIndex [43], TreePi [45]). In particular, the case study provides the services of seven techniques and evaluates them on the basis of their indexing time, index size and query processing time using a real AIDS antiviral screen dataset (NCI/NIH) and synthetically generated datasets.
- *String Similarity Join*[20]: An implementation for the recent evaluation and comparison study which is presented by Jiang et al. [23]. The case study provides the implementation of twelve algorithms and provides six different experimental datasets. The evaluation of the benchmarked algorithms is based on two metrics: the running time and the size of candidate results.
- *Reverse K Nearest Neighbors (RkNN)*[21]: An implementation for the recent evaluation and comparison study which is presented by Yang et al. [44]. The case study provides the implementation of various Reverse k Nearest Neighbors Query Processing algorithms over various experimental datasets.

Each of our case studies is deployed in two cloud environments: the Amazon public cloud environment[22] with its various cloud services (e.g., Simple Storage Service (S3[23]), Elastic Compute Cloud (EC2[24])) in addition to our own private cloud environment which is managed by the *OpenStack* platform[25]. However, the platform can be easily adopted to run over other cloud environments (e.g., *CloudStack*[26], *Eucalyptus*[27]). In addition, each case study is configured using two different testing environments (virtual machines): The first environment is configured with high computing resources while the other environment is configured with limited computing resources in order to imitate the various real world scenarios. Furthermore, authenticated users of our platforms can access various services of the platform (e.g., searching the repository of results, creating and running experiments) via our provided RESTful interfaces and API-based SDK[28]. Tho social features of our platform has been implemented the open source social network platform, *elgg*[29].

VII. Conclusion

In principle, the field of practical computer science is suffering from the repeatability problem of the research results [11] which represents a keystone of the scientific process. It is common that the results of experiments tend to reside in some folders or repositories which have never been documented thoroughly. Several reasons are behind this problem but time pressure is the most prominent of them. In practice,

[19] The full documentation and screencast of this case study is available on http://wiki.liquidbenchmark.net/doku.php/casestudy-graph-indexing-querying

[20] The full documentation and screencast of this case study is available on http://wiki.liquidbenchmark.net/doku.php/casestudy-string-similarity-join

[21] The full documentation and screencast of this case study is available on http://wiki.liquidbenchmark.net/doku.php/reverse-k-nearest-neighbors

[22] http://aws.amazon.com/

[23] https://s3.amazonaws.com/

[24] http://aws.amazon.com/ec2/

[25] http://www.openstack.org/

[26] http://cloudstack.apache.org/

[27] https://www.eucalyptus.com/

[28] http://wiki.liquidbenchmark.net/doku.php/RESTful-interface

[29] http://elgg.org/

documentation of experiments and data traceability needs valuable work time, while publish or perish and strict conference deadlines call for timely results. In practice, the Web has dramatically enhanced the people's ability to share knowledge, ideas and contributions. We believe that the Computer Science research community should have the leadership in having such *scientific* collaborative environments that can significantly develop and improve the capacity of the scientific communities on deeply understanding the details of their research challenges, have careful, clean and insightful analysis for the state-of-the-art that can support them for developing new effective approaches, techniques and solutions.

In this article, we presented the design and implementation details of the *Liquid Benchmarking* platform that relies on the current advances in the Web technologies to provide *collaborative* Web-based platforms that democratize and socialize the key tasks of evaluating, comparing and analyzing the *continuous* scientific contributions in different domains of the Computer Science field. We believe that our platform can effectively exploit the increasing human power which are participating in the Computer Science research efforts and distributed over the world. In particular, we argue that our platform can empower the Computer Science research communities with many capabilities such as:

- Developing *focused* and centralized repositories for related software implementations [2] and their experimental results. These repositories can serve as a very positive step towards tackling the experimental *reproducibility* challenge in the Computer Science field.
- Facilitating the establishment of shared computing resources environments that can be exploited by different active contributors in the same domain who reside in different parts of the world.
- Providing *workable* environments to collaboratively establish standard benchmarks that can be widely utilized for achieving insightful evaluation for alternative research efforts. These environments can help researchers to optimize their time in assessing and improving the quality of their contribution. Having such environments will discourage authors from publishing paper with adhoc or poor experimental results.
- Facilitating *collaborative* maintenance of experimental studies to guarantee their *freshness*. This task can follow the same model of collaborative organization of international conferences or journals where each participating researchers or research groups in a specific community can play a volunteering managerial role for a specific period.
- Exploiting the *wisdom of the crowd* in providing feedbacks over the experimental results in a way that can provide useful insights for tackling further problems and improving the state-of-the-art.
- Creating a *transparent* platform for scientific *crediting* process based on collaborative community work.
- Establishing concrete foundations and feasible environments for providing *provenance* services [28] for scientific experimental results and time-analysis services for the evolution of research efforts.

Therefore, we hope that our platform can serve as the foundation for a fundamental rethinking of the experimental evaluation process in the Computer Science field. As a future work, we are planning to implement more case studies using our platform and make them available for the research community. In addition, we are planning to add more features of the social aspect of the platform with careful consideration to important details such as credit attribution and data anonymization [17].

ACKNOWLEDGEMENT

The Liquid Benchmark Project is supported by King Abdulaziz City for Science and Technology (KACST), NSTIP (MAARIFAH) project 11-INF1992-03.

REFERENCES

[1] Benchmark of XML compression tools. http://xmlcompbench.sourceforge.net/.

[2] SourceForge: A free repository of open source software. http://sourceforge.net/.

[3] Transaction Processing Performance Council. http://www.tpc.org/default.asp.

[4] Daniel J. Abadi, Adam Marcus, Samuel Madden, and Katherine J. Hollenbach. Scalable Semantic Web Data Management Using Vertical Partitioning. In *VLDB*, pages 411–422, 2007.

[5] Denilson Barbosa, Ioana Manolescu, and Jeffrey Xu Yu. Microbenchmark. In *Encyclopedia of Database Systems*, page 1737. 2009.

[6] David Bermbach, Liang Zhao, and Sherif Sakr. Towards Comprehensive Measurement of Consistency Guarantees for Cloud-Hosted Data Storage Services. In *TPCTC*, 2013.

[7] Philippe Bonnet, Stefan Manegold, Matias Bjørling, Wei Cao, Javier Gonzalez, Joel A. Granados, Nancy Hall, Stratos Idreos, Milena Ivanova, Ryan Johnson, David Koop, Tim Kraska, René Müller, Dan Olteanu, Paolo Papotti, Christine Reilly, Dimitris Tsirogiannis, Cong Yu, Juliana Freire, and Dennis Shasha. Repeatability and workability evaluation of SIGMOD 2011. *SIGMOD Record*, 40(2):45–48, 2011.

[8] Michael J. Carey, David J. DeWitt, and Jeffrey F. Naughton. The oo7 Benchmark. In *SIGMOD*, pages 12–21, 1993.

[9] James Cheney. Compressing XML with Multiplexed Hierarchical PPM Models. In *Proceedings of the Data Compression Conference (DCC)*, page 163, Washington, DC, USA, 2001. IEEE Computer Society.

[10] Fernando Seabra Chirigati, Matthias Troyer, Dennis Shasha, and Juliana Freire. A computational reproducibility benchmark. *IEEE Data Eng. Bull.*, 36(4):54–59, 2013.

[11] Christian Collberg, Todd Proebsting, Gina Moraila, Zuoming Shi, and Alex M Warren. Measuring Reproducibility in Computer Systems Research. Technical report, 2014.

[12] Alain Crolotte. Issues in Benchmark Metric Selection. In *TPCTC*, pages 146–152, 2009.

[13] Christian Dietrich and Daniel Lohmann. The dataref versuchung: Saving Time through Better Internal Repeatability. *Operating Systems Review*, 49(1):51–60, 2015.

[14] Peter Dolog, Markus Krötzsch, Sebastian Schaffert, and Denny Vrandecic. Social Web and Knowledge Management. In *Weaving Services and People on the World Wide Web*, pages 217–227, 2008.

[15] Thomas Erl, Ricardo Puttini, and Zaigham Mahmood. *Cloud Computing: Concepts, Technology & Architecture*. Prentice Hall, 2013.

[16] Dror G. Feitelson. From repeatability to reproducibility and corroboration. *Operating Systems Review*, 49(1):3–11, 2015.

[17] Juliana Freire, Philippe Bonnet, and Dennis Shasha. Exploring the Coming Repositories of Reproducible Experiments: Challenges and Opportunities. *PVLDB*, 4(12):1494–1497, 2011.

[18] Juliana Freire, Philippe Bonnet, and Dennis Shasha. Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD*, pages 593–596, 2012.

[19] Wook-Shin Han, Jinsoo Lee, Minh-Duc Pham, and Jeffrey Xu Yu. igraph: A framework for comparisons of disk-based graph indexing techniques. *PVLDB*, 3(1):449–459, 2010.

[20] Wook-Shin Han, Minh-Duc Pham, Jinsoo Lee, Romans Kasperovics, and Jeffrey Xu Yu. iGraph in action: performance analysis of disk-based graph indexing techniques. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1241–1242, 2011.

[21] Huahai He and Ambuj K. Singh. Closure-Tree: An Index Structure for Graph Queries. In *ICDE*, page 38, 2006.

[22] Karl Huppler. The Art of Building a Good Benchmark. In *TPCTC*, pages 18–30, 2009.

[23] Yu Jiang, Guoliang Li, Jianhua Feng, and Wen-Syan Li. String Similarity Joins: An Experimental Evaluation. *PVLDB*, 7(8):625–636, 2014.

[24] Hartmut Liefke and Dan Suciu. XMill: An efficient compressor for XML data. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 153–164. ACM, 2000.

[25] Stefan Manegold and Ioana Manolescu. Performance evaluation in database research: principles and experience. In *Proceedings of the 12th International Conference on Extending Database Technology (EDBT)*, page 1156, 2009.

[26] Stefan Manegold, Ioana Manolescu, Loredana Afanasiev, Jianlin Feng, Gang Gou, Marios Hadjieleftheriou, Stavros Harizopoulos, Panos Kalnis, Konstantinos Karanasos, Dominique Laurent, Mihai Lupu, Nicola Onose, Christopher Ré, Virginie Sans, Pierre Senellart, T. Wu, and Dennis Shasha. Repeatability & workability evaluation of SIGMOD 2009. *SIGMOD Record*, 38(3):40–43, 2009.

[27] Ioana Manolescu, Loredana Afanasiev, Andrei Arion, Jens Dittrich, Stefan Manegold, Neoklis Polyzotis, Karl Schnaitter, Pierre Senellart, Spyros Zoupanos, and Dennis Shasha. The repeatability experiment of sigmod 2008. *SIGMOD Record*, 37(1):39–45, 2008.

[28] Simon Miles, Paul T. Groth, Ewa Deelman, Karan Vahi, Gaurang Mehta, and Luc Moreau. Provenance: The Bridge Between Experiments and Data. *Computing in Science and Engineering*, 10(3):38–46, 2008.

[29] David Patterson and Armando Fox. *Engineering Software as a Service: An Agile Approach Using Cloud Computing*. Strawberry Canyon LLC, 2013.

[30] Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, and Michael Stonebraker. A comparison of approaches to large-scale data analysis. In *SIGMOD*, pages 165–178, 2009.

[31] S. Sakr and F. Casati. Liquid Benchmarks: Towards An Online Platform for Collaborative Assessment of Computer Science Research Results. In *TPCTC*, 2010.

[32] Sherif Sakr. XML compression techniques: A survey and comparison. *J. Comput. Syst. Sci.*, 75(5):303–322, 2009.

[33] Sherif Sakr. Cloud-hosted databases: technologies, challenges and opportunities. *Cluster Computing*, 17(2):487–502, 2014.

[34] Sherif Sakr, Anna Liu, and Ayman G. Fayoumi. The family of mapreduce and large-scale data processing systems. *ACM Comput. Surv.*, 46(1):11, 2013.

[35] Sherif Sakr, Amin Shafaat, Fuad Bajaber, Ahmed Barnawi, Omar Batarfi, and Abdulrahman H. Altalhi. Liquid Benchmarking: A Platform for Democratizing the Performance Evaluation Process. In *EDBT*, 2015.

[36] Sherif Sakr, Liang Zhao, Hiroshi Wada, and Anna Liu. CloudDB AutoAdmin: Towards a Truly Elastic Cloud-Based Data Store. In *ICWS*, 2011.

[37] J. D. Scargle. Publication bias: The File-Drawer problem in scientific inference. *Journal of Scientific Exploration*, 14(1):91–106, 2000.

[38] Albrecht Schmidt, Florian Waas, Martin L. Kersten, Michael J. Carey, Ioana Manolescu, and Ralph Busse. Xmark: A benchmark for XML data management. In *Proceedings of 28th International Conference on Very Large Data Bases (VLDB)*, pages 974–985, 2002.

[39] Lefteris Sidirourgos, Romulo Goncalves, Martin L. Kersten, Niels Nes, and Stefan Manegold. Column-store support for RDF data management: not all swans are white. *PVLDB*, 1(2):1553–1563, 2008.

[40] Michael Stonebraker. A New Direction for TPC? In *TPCTC*, pages 11–17, 2009.

[41] Young-Kyoon Suh, Richard T. Snodgrass, and Rui Zhang. Azdblab: A laboratory information system for large-scale empirical DBMS studies. *PVLDB*, 7(13):1641–1644, 2014.

[42] Erik Wilde and Cesare Pautasso, editors. *REST: From Research to Practice*. Springer, 2011.

[43] Xifeng Yan, Philip S. Yu, and Jiawei Han. Graph Indexing: A Frequent Structure-based Approach. In *SIGMOD*, pages 335–346, 2004.

[44] Shiyu Yang, Muhammad Aamir Cheema, Xuemin Lin, and Wei Wang. Reverse k Nearest Neighbors Query Processing: Experiments and Analysis. *PVLDB*, 8(5):605–616, 2015.

[45] Shijie Zhang, Meng Hu, and Jiong Yang. TreePi: A Novel Graph Indexing Method. In *ICDE*, pages 966–975, 2007.

Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews

Mahmoud Al-Ayyoub and Aya Nuseir
Jordan University of Science and Technology
Irbid, Jordan

Ghassan Kanaan and Riyad Al-Shalabi
Amman Arab University
Amman, Jordan

–

***Abstract*—Sentiment Analysis (SA) is one of hottest fields in data mining (DM) and natural language processing (NLP). The goal of SA is to extract the sentiment conveyed in a certain text based on its content. While most current works focus on the simple problem of determining whether the sentiment is positive or negative, Multi-Way Sentiment Analysis (MWSA) focuses on sentiments conveyed through a rating or scoring system (e.g., a 5-star scoring system). In such scoring systems, the sentiments conveyed in two reviews of close scores (such as 4 stars and 5 stars) can be very similar creating an added challenge compared to traditional SA. One intuitive way of handling this challenge is via a divide-and-conquer approach where the MWSA problem is divided into a set of sub-problems allowing the use of customized classifiers to differentiate between reviews of close scores. A hierarchical classification structure can be used with this approach where each node represents a different classification sub-problem and the decision from it may lead to the invocation of another classifier. In this work, we show how the use of this divide-and-conquer hierarchical structure of classifiers can generate better results than the use of existing flat classifiers for the MWSA problem. We focus on the Arabic language for many reasons such as the importance of this language and the scarcity of prior works and available tools for it. To the best of our knowledge, very few papers have been published on MWSA of Arabic reviews. One notable work is that of Ali and Atiya, in which the authors collected a large scale Arabic Book Reviews (LABR) dataset and made it publicly available. Unfortunately, the baseline experiments on this dataset had very low accuracy. We present two different hierarchical structures and compare their accuracies with the flat structure using different core classifiers. The comparison is based on standard accuracy measures such as precision and recall in addition to using the mean squared error (MSE) as a more accurate measure given the fact that not all misclassifications are the same. The results show that, in general, hierarchical classifiers give significant improvements (of more than 50% in certain cases) over flat classifiers.**

***Keywords*—multi-way sentiment analysis, hierarchical classifiers, support vector machine, decision tree, naive bayes, k-nearest neighbor, mean squared error**

I. Introduction

In the last decade, the number of Internet users has increased significantly. This increase can be seen as a result of the technologies that facilitated the widespread of the Internet, along with the various services provided through the Internet. These services includes social networking (Facebook, Twitter, etc.), publications (news, books, etc.) and other day-to-day services. The exposure of people to these online services allowed them to express their feelings and emotions regarding the provided services or in reaction to some subject in their lives. Furthermore, organizations of various types utilized the Internet to allow them to collect people's opinions about almost all the subjects the concern them through easing the process of getting feedback or by collecting what people are feeling from the various public websites. After the collection of the raw unstructured data containing these expressions, some processing must be performed to analyze the people sentiments. As a result, the interdisciplinary Sentiment Analysis field has emerged.

Sentiment Analysis (SA), also known as Opinion Mining (OM), refers to the use of natural language processing, text analysis and computational linguistics to identify and extract the sentiment orientation of textual materials.[1] The extraction of a sentiment can be made either on a whole document (document-level SA), on each paragraph (paragraph-level SA), or on each sentence (sentence-level SA) [37]. The considered sentiment orientations are usually assumed to simply be positive and negative only; making SA a binary classification problem. Some researchers, however, add more classes for neutral or conflicted sentiments. Note that this is different from the more general problem of emotion analysis, where the authors are interested in identifying more complex emotions such as joy, fear, etc. [11], [20].

The reason behind the immense interest in SA is because obtaining truthful information about the opinions of the stakeholders is a crucial point in any decision making process [45]. The authors of [37], [21] list several examples such as a company's use of SA tools to obtain a true indicator of its customers' satisfaction with its products and services. It can plan ahead according to such feedback to guarantee wider acceptance and larger market share. Another example of SA application is as a quicker and more accurate alternative of public polls. Instead of relying on public polls with all of their problems and expenses, government can measure the public's opinion by simply crawling through what is written on social networks and evaluate it using SA tools. Finally, recommender systems can benefit greatly from efficient and effective SA tools.

Most of the available SA tools can be categorized into one of two approaches: the corpus-based (supervised) approach and the lexicon-based (unsupervised) approach [38]. The corpus-based approach simply views SA as being a special case of text classification in which the classes are simply the sentiment

[1] http://en.wikipedia.org/wiki/Sentiment_analysis

orientations. A large dataset of manually annotated examples is used to train the classifier and testing techniques such as cross validation are used to evaluate the performance of the classifier. On the other hand, the lexicon-based approach, as its name implies, utilizes a lexicon composed of terms along with their sentiment values. To determine the sentiment value of certain text, the lexicon-based approach searches through the lexicon for the sentiment values of the terms composing the text and combines them. Combining these two approaches resulted in a hybrid approach known as the weakly-supervised approach [25].

Each approach has its pros and cons. Compared to the effort required by the lexicon-based approach, the corpus-based approach is considered very expensive due to the need for a large annotated dataset. In [46], with the variation of the topics, domains and time-periods, the corpus-based approach has the advantage of higher accuracy in SA. The focus of this work is on the corpus-based approach.

Many of the current works on SA consider the simple binary (or ternary) setting. However, there are few works that consider sentiment orientations based on some scoring or rating systems, in what is known as the Multi-Way Sentiment Analysis (MWSA) problem. People might shy away from working on this problem due to the challenges associated with it despite its apparent importance and strong association with recommender systems.

Having more classes is not the only additional challenge imposed by MWSA. The obvious difficulty of MWSA does not only come from considering more sentiment orientations, it is the relationship between these sentiment orientations that makes things difficult. To fully understand the effect of this issue, consider the settings of this work. This work focuses on the problem of automatically determining the rating of an Arabic review on a scale from 1 to 5. In a 5 star rating system, both 4 and 5 stars are considered positive sentiment orientations, while 1 and 2 stars are considered negative ones, and 3 stars as a neutral orientation. One of the challenges of this settings is the difficulty in distinguishing between the two positive ratings. It is much easier to tell whether a review is positive or not compared to telling whether it is a strong positive or a weak one. Distinguishing between the two negative ratings is equally hard. So, it is natural to decompose the MWSA problem into a set of sub-problems and employ a hierarchical classification structure in which an optimized classifier is devised to address each sub-problem separately. For example, one classifiers can be trained to distinguish between positive and negative reviews while another classifier is trained independently to distinguish between strong and weak positive reviews. Similarly a third classifier is trained independently to distinguish between strong and weak negative reviews. This gives the intuition that such an inherently hierarchical problem cannot be effectively addressed using a flat approach. The hierarchical classification approach has already been shown to be very useful for MWSA of English reviews [22]. Our goal is to investigate the effectiveness of this approach for Arabic reviews.

Most studies on SA in general and specifically on MWSA have been conducted on the English language with few considering other languages. This might be due to having large datasets publicly available for the English language. In this project, we consider the Arabic language for applying MWSA. Arabic is the language of 22 countries with more than 400 million inhabitants. Natural Language Processing for the Arabic language is considered challenging due to the special characteristics of the language such as: orthography, the existence of short vowels, the complex morphology compared to English, the widespread of synonyms and the lack of publicly and freely accessible corpora [29], [35], [10].

To the best of our knowledge, very few papers have been published on the MWSA problem. One notable work is that of Ali and Atiya [21], in which the authors collected a large scale Arabic Book Reviews (LABR) dataset and made it publicly available. Unfortunately, the baseline experiments on this dataset had very low accuracy. Motivated by the intuition that employing a flat classifier to handle an inherently hierarchical problem such as MWSA is one of the main reasons behind such poor accuracy, we propose to use hierarchical classification. We present two different hierarchical structures and compare their accuracy with the flat structure using different core classifiers. The comparison is based on standard accuracy measures such as precision and recall in addition to using the mean squared error (MSE) as a more accurate measure given the fact that not all misclassifications are the same.

The rest of this paper is structured as follows. Section II presents the related works. In Section III, we present our system model and evaluate it in Section IV. Finally, we conclude in Section V.

II. Background and Related Works

The problem at hand is the MWSA of Arabic reviews using hierarchical classification. The following coverage of the literature focuses on similar works on the English language before discussing the existing works on SA and MWSA of Arabic text.

The word hierarchical classification has appeared in several contexts such as: hierarchical labeling (in which the labels are structured into a hierarchy of classes and subclasses), hierarchical classifier (in which the classifiers themselves are structured in a hierarchy), ensemble methods, One-versus-All (OVA), One-versus-One (OVO), etc. Each one of these concepts is investigated by many researchers to study its applicability and effect on the classification processes. The focus of this work is on hierarchical classifiers. Using the divide and conquer mentality, a hierarchical classifier breaks the classification problem into several sub-problems and attacks each sub-problem according to a tree or a Directed Acyclic Graph (DAG) hierarchy [34].

Most of the available researches are based on flat classification. However, the exploding number of online data makes the use of flat classification methods more difficult giving rise to the concept of hierarchical classification. The hierarchical classification is based on divide-and-conquer principle, where the problem can be divided into sub problems and easily can be solved.

The most commonly used structure in hierarchical classification is the tree-like structure, however, many researchers have proposed using Directed Acyclic Graphs (DAG). One difference between the different structures proposed in the literature is whether a node can have more than one parent or not. Another difference is whether the class is given in leaf nodes only or in leaf as well as internal nodes [55], [52].

In [32], the authors developed two types of hierarchies for emotion classification using SVM classifier. The dataset consisted of six emotional classes (happiness, sadness, fear, anger, disgust, and surprise) in addition to the no-emotion class which contain instances conveying no feelings or emotions. The first type of hierarchy is a two-level classification hierarchy where the first level tests whether the instance is emotional or non-emotional, then the second level takes instances that are classified as emotional and classify them into one of the six emotional classes. The second type is a three-level classification where the first level is the same as first level of the two-level type, the second level classifies the emotional instances into positive (happiness) or negative, and the last level classifies negative instances into one of the five classes (sadness, fear, anger, disgust, and surprise).

The authors of [22] proposed a hierarchical classifier tree (MCST) consisting of standard binary classifiers (linear SVM since it is considered as a very efficient text classifier) to perform the MWSA of English text. To construct MCST, Kruskal's algorithm is used along with a similarity measure between every pair of classes as follows. First, the algorithm determined the representative feature vector for each class (using two methods: centroid and sample selection). Then, it evaluates the distance between every pair of classes using Euclidean distance and Tanimoto Coefficient. According to [22], the major benefit of using MCST comparing with other hierarchical classifiers (OVA, OVO, and DAGSVM) is that it handles the overfitting problem in a better way.

In a followup work [23], the authors proposed a probabilistic approach that combines information from lexicons with a Naive Bays classifier. They compared their approach with a Naive Bayes classifier coupled with feature selection in addition to [22]'s approach. They used different datasets from different domains such as: movies, kitchen appliances, music, and post office. In addition to these works, the interested readers are referred to [38], [44], [31], [33], [51] for further information about English MWSA. In the following, we shift our attention to the works specifically geared towards the Arabic language.

Similar to the SA work for the English language, Arabic SA tools can be generally categorized into the following three approaches: the corpus-based approach (supervised), the lexicon-based approach (unsupervised) and a hybrid of the two called the weakly- or semi-supervised approach. Based on our study of the literature, the first paper on Arabic SA appeared in 2006, but it was not until 2010 that we started to see a surge of interest in Arabic SA manifested in an increasing number of published papers. Below, we discuss most of the influential papers in the field providing a comprehensive and up-to-date coverage. In 2006, Ahmad, Cheng and Almas [9] attempted automatic SA on financial news on Arabic and Chinese introducing the local grammar approach that was developed on an English archive and used it on Arabic and Chinese with almost the same results. In [18] Ahmad and Almas extended this work to be on English, Arabic and Urdu. They concluded that, considering the F-measure, Arabic text polarity identification was a bit better than for English text. An observation from the considered languages is that the number of positive sentiments were significantly more than the negative ones in Arabic, English and Urdu. Another work on business-related SA is the work of Elhawary and Elfeky [27] in which a MapReduce implementation was employed to improve the performance of the existing SA systems.

In 2010, Farra [30] followed the same general approach as Ahmad et al. [9], [18] by proposing another SA tool based on a grammatical approach. The approach also took advantage of a precompiled lexicon. The authors compared between the performance of their system on sentence-level and document-level. Other lexicon-based works include [12], [7].

Many papers [5], [36] appeared in the literature to compare the two most common approach for SA: the corpus-based approach and the lexicon-based approach. Moreover, El-Halees [25] studied the two approaches and proposed a hybrid approach that incorporates a third approach based on Maximum Entropy (ME). In another hybrid approach, Abdulla et al. [6] proposed to use an annotated dataset to automatically expand manually-created lexicons leading to significant improvements in the accuracy of the lexicon-based approach.

With the growing interest in Arabic SA, the need for standardized and publicly available datasets to serve as benchmarks became imminent. The works on the OCA and AWATIF datasets is considered pioneering in this aspect. In 2011, Rushdi-Saleh et al. [49], [48] published two papers explaining their Opinion corpus for Arabic (OCA), which consists of 500 movie reviews divided equally among the positive/negative classes. In addition to the process of dataset collection and annotation, the authors performed some experiments on their dataset including studying the effect of applying machine translation on it, which would generate an English version of OCA (EVOCA). On the other hand, the AWATIF dataset [2] was generated with many issues from linguistics point of view in mind such as whether the annotators are aware of certain linguistic features of subjectivity and sentiment analysis or not and how such knowledge would affect their decision. Other works on dataset collection was recently conducted [8], [14], where the authors provided more information about the collected comments than typical SA datasets which focus only on the sentiment orientation of the comment. The dataset of [8], [14] included information about the dialect used, the domain, the gender of the author, etc. Finally, the LABR dataset used in this project was collected by Aly and Atiya [21] in 2013. Since this is the dataset used in this project, more details about it will be provided in the following section. The same group presented another dataset named the Arabic Sentiment Tweets Dataset (ASTD) [42] consisting of more than 10,000 tweets (most of which are non-subjective tweets). Another dataset of large-scale nature was collected by ElSahar and El-Beltagy [28] consisting of more than 33,000 reviews in different domains such as hotels, movies, etc.

One interesting aspect of [3], [1], [37], [2], [4] is that it adds another dimension to the SA problem by bringing subjectivity analysis into the picture. Before thinking of any commercialization of any SA tool, one has to deal with determining whether a text is subjective or objective. Another useful aspect of this line of work is that it brings a lot of interesting and useful ideas from the traditional field of linguistics. Taking subjectivity analysis into account, the authors of [47] used a simple two-level hierarchical classification system in which the first level distinguish polar vs. objective sentiments and the second level take the polar reviews and decide if they convey positive or negative sentiments.

Other works aimed at solving issues related to SA in general such as how to handle short text documents (e.g., tweets) [50], the different Arabic dialects [50], [13], [24], the imbalance in the dataset [41], the credibility of the comments [19] and how to identify the opinion holder [26], etc.

As mentioned before, the most relevant work to ours are those of Aly and Atiya [21] and Al Shboul el al. [15]. In [21], the authors constructed the Large scale Arabic Book Reviews (LABR) dataset which is, as the authors claim, one of the largest Arabic Sentiment Analysis corpora available recently. The dataset was collected from social network site www.goodreads.com and was subjected to data preparation process. After that, the authors conducted their experiments in order to investigate two main tasks. The first one is to decide whether a review is positive (with rate 4 or 5) or negative (with rate 1 or 2). The second task is to determine the rate of review on a scale from 1 to 5. By employing different features and classifiers (Multinomial Naive Bayes, Bernoulli Naive Bayes, and Support Vector Machines(SVM)) they found that the best accuracy for task one was 91% using SVM, and for task two the greatest accuracy was 50% using also SVM classifier. In another work on the same dataset, the authors of [15] experimented with other classifiers without being able to give significantly better results than the baseline provided by [21]. They also explored the imbalance issue of the LABR dataset arguing that it negatively affects the accuracy of the classifiers.

Other works benefited from the LABR dataset. One example of is the Human Annotated Arabic Dataset (HAAD) of Al-Smadi et al. [17], which consists of more than 1,500 reviews selected from the LABR dataset and annotated for Aspect-Based SA (ABSA) according to the SemEval2014 Task4 guidelines.[2]. In [17], the authors presented several baseline experiments, which they later improved in [43]. The same group followed the same approach of ABSA in another study on the effect of news on the users of social media [16].

III. Methodology and Experiment Setting

This section outlines the methodology and materials used in our work, including the description of the used dataset, the data mining tools employed, and the accuracy measurements used to evaluate the proposed approach.

A. Dataset

The dataset used in this study is the Large Scale Arabic Book Reviews (LABR dataset) which consists of 63,257 book reviews written in MSA as well as colloquial Arabic. The reviews were collected from goodreads.com during 2013. Each book review has a rating (1 to 5) along with the text of the review [21]. The distribution of reviews across the different ratings is discussed in Section III-D. The collected dataset underwent a filtering step to remove newline characters, HTML tags, hyperlinks, repeated dots, non-Arabic characters and some special unicode characters such as the heart symbol and special quotation symbols.

B. Preprocessing and Mining Tools

The tool that is used is Weka 3.7.10. It opens the source software and combines a large set of Machine Learning algorithms, tools for data preparation and preprocessing, and data visualization. Weka allows users to use its algorithms by invoking them using Java code or by applying them directly using its GUI. Our system, which is implemented using the Java programming language, imports Weka library to make benefit from its Machine Learning algorithms.

In our work we use four different classifiers: Support Vector Machine (SVM), Naive Bayes (NB), KNN and Decision Tree (DT). In this work we construct different hierarchical classifiers tree since weka doesn't offer the ability to use such hierarchical classifiers, then we compare the performance of this four classifiers when we use it during hierarchical classifiers tree with the performance of same four classifiers during flat classification problem.

In order to get the ability of using these documents in weka first it needs to be converted to arff file, and this is done by a program written in java code with weka TextDirectoryToArff converter. The next step is to extract features from these documents, the most common approach for features extraction is the Bag-Of-Word. Using StringToWordsVector weka filter we can represent each document as vector, this filter offers the ability to use different techniques help in features extraction and reduction, but before using these techniques to reduce large features vector, we need to tokenize the text to get meaningful words, the WordTokenizer used as tokenizer, which splits or extracts words from the text using standard delimiters. To reduce the number of features, we remove stop words such as pronouns, prepositions, and names of days the week, etc. As for the stemmer, in this our work we did not use any type of stemmers since some of the reviews were written in delicate form of language. The generated dataset divided into two dataset, the first one is the training set which involves about 66% of the original dataset, and the second one is the test dataset which has the rest percentage of the original dataset.

C. Core Classifiers

During this section we will introduce the classifiers that we used in the flat classification problem, as we mentioned before in chapter 2 about the classification methods as SVM, Naive Bayes, Decision Trees, and KNN classifiers. These four classifiers used in our implemented approaches trained under set of rules that will be expressed in the following:

Support Vector Machine (SVM). We used Sequential Minimal Optimization (SMO) algorithm for training SVM classifier, we made for this type of classifiers three experiments according to the kernel type, kernel ideas emerged when data cant be linearly separated, based on finding similarity between two points. In this research we have made our choice to utilize two widely used kernels, Polynomial Kernel (PK) of degree p and Radial Basis Function (RBF) Kernel with σ as the width of the radial basis function. We experimented with both kernel types and different values of p and σ as suggested by [54],

[2] http://alt.qcri.org/semeval2014/task4/

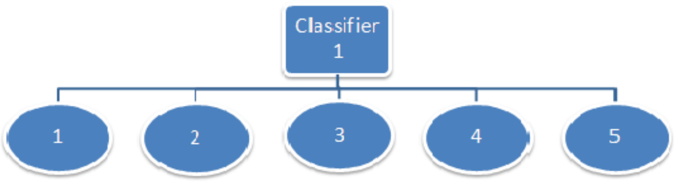


Fig. 1: The flat classifier.

[40]; however, we only report the settings that produced the best results.

Decision Tree (DT). For this classifier, we used J48 algorithm, using this algorithm we made several experiments by using different values for confidence factor parameter, which controls the size of tree. The default value for this parameter is 0.25 in weka, however, better results are obtained by using other values such as 0.2.

K**-Nearest Neighbor (KNN).** For this classifier, we used the IBk algorithm. We made several experiments using different numbers of K. The default value for K is 1. We tried other values besides the default one and report the best results. It is worth mentioning that we kept the default distance function, which is the Euclidean distance.

Naive Bayes (NB). For this one, we used the NaiveBayes algorithm. One experiment was made, using default settings that weka provides.

D. Classification Structures

In addition to the flat classification structure, in this work we construct two different hierarchical classification structures. In this section, we explain these structures and discuss the intuition behind them. For the hierarchical structures, the top down approach was followed in constructing the hierarchies and every node represents a binary core classifier (i.e., an instance of one of the four classifiers discussed in the previous subsection). The following provides more details about the structures considered in this work.

The first structure we discuss is the simplest one. It is the flat structure. It is a one-level structure with a single node containing a core classifier. This classifier is trained on the entire training set as is in order for it to distinguish the five classes under consideration in one shot. Figure 1 shows a graphical depiction of this structure.

For this structure, the LABR dataset is used as is. The distribution of the reviews across the five classes under consideration is as follows. Class 1 contains 2,939 reviews, class 2 contains 5,285 reviews, class 3 contains 12,201 reviews, class 4 contains 19,054 reviews and finally class 5 contains 23,778 reviews. This is a very unbalanced dataset.

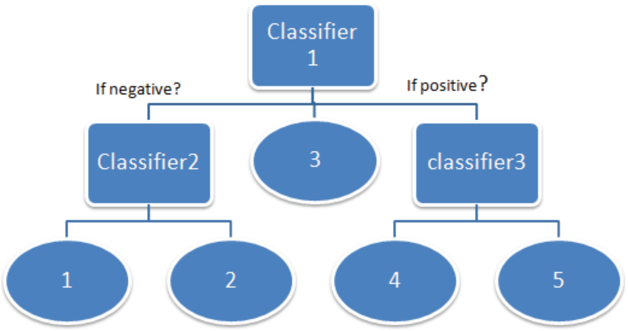


Fig. 2: The 2-level hierarchical classifier.

The first non-flat hierarchical structure is created under two levels as shown in Figure 2. As the figure shows, the first level checks whether an instance is negative, neutral, or positive category. For the LABR dataset, the number of negative reviews (with ratings 1 or 2) is 8,224, while the number of positive reviews (with ratings 4 or 5) is 42,832. The remaining 12,201 reviews are neutral. This level is still suffering from the same imbalance of the original dataset. The classifiers in the next level are customized to determine whether a positive review is weak or strong or to determine whether a negative review is weak or strong. Each one of these classifiers deal with a more balanced datasets compared with the first level classifiers.

Inspired by a mixture of OVA hierarchical structures as well as the high imbalance in the dataset towards positive classes, we devise another hierarchical structures consisting of four levels as shown in Figure 3. As the figure shows, each level has a single core classifier responsible for making a single binary decision. The top level starts by determining whether a review belongs to the majority class (class label 5) or not (i.e., it belongs to one of the class labels 1 through 4). If not, then the decision is moved to the second level classifier which is responsible for determining whether a review belongs to class label 4 or not. This continues until we reach the bottom level classifier which is responsible for determining whether a review belongs to class label 1 or 2. One intuition behind building such a structure is to place the easiest OVA decisions (i.e., the one associated with the majority class) as near the top as possible.

E. Evaluation Measures

In order to evaluate the performance of the different classifiers produced by the different combination of structures with core classifiers, we report different metrics from each experiment. Five metrics are used: accuracy, micro-average precision, micro-average recall, F-measure and Mean Square Error (MSE). To explain these measures, it is generally assumed for a binary classification problem such as ours that there is a “positive” class and a “negative” one. *Precision* calculates the ratio of the true positives to the total number of positives predicted by the classifier. The higher the precision,

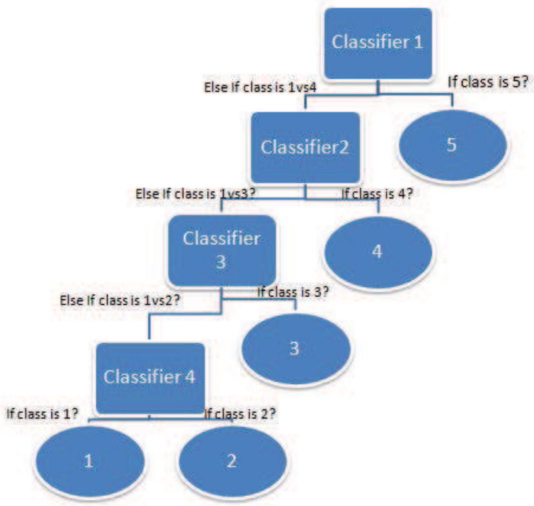


Fig. 3: The 4-level hierarchical classifier.

the more accurate the predication of the positive class. On the other hand, *recall* divides the true positives by the total actual positives belongs to that class. A high recall means high number of comments from the same class is labeled to its exact class. F-measure is weighted average of precision and recall. As for the *accuracy*, it simply reports the ratio of the correctly classified documents regardless of their class. The following are the formulas for these measures [56]:

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+FP+TN+FN} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{Recall} &= \frac{TP}{TP+FN} \\
\text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{MSE} &= \frac{1}{n}\sum_{i=1}^{n}(y'_i - y_i)^2
\end{aligned}$$

where TP, FP, TN and FN are the numbers of true positives, false positives, true negatives and false negatives, respectively, and Ci is the number of instances belong to the class i. y' is predicted value y is the true value. True positives and negatives are the correctly classified comments whereas false positives are the number of comments that are incorrectly classified as positive and false negative are the number of comments that are incorrectly classified as negative.

Note that in previous paragraph, we mentioned MSE as an evaluation metric without discussing it. This is because the other four metrics are standard metrics that are assumed to be reported in any work like ours. However, these metrics alone fail to paint a correct picture of the accuracies of the considered classifiers. According to [53], these four measures are not efficient to use with hierarchical classification problems, since they fail to account for the the relationship between categories. Instead, they tackle each one in isolation from the other categories. Simply put, an error of classifying a strong positive review (with a class label 5) as a weak positive review (with a class label 4) is not as serve as classifying it as a strong negative review (with a class label 1). MSE compensate for such issues by relying on the absolute distance between the actual class and the predicted class making it more suitable for a problem like MWSA.

IV. Results and Analysis

In this section, the results of our experiments are described in details. The goal is to study and comparing the performance of the different classification structures using different core classifiers. For the classification process, we use four core classification algorithms.

- The first one is the Sequential Minimal Optimization (SMO) algorithm for training SVM classifier using Radial Basis Function (RBF) kernel with $\sigma = 0.10$. We perform extensive experiments for the different available choices for each kernel and its parameters based on the suggestions and observations of previous similar works in the literature [54], [40]. However, we only report here the settings that gave the best results.
- The second one is the Naive Bayes (NB) classifier used with default settings as provided by the Weka tool.
- The third one is the decision tree algorithm. Specifically, we use the Java implementation of DT provided by the Weka tool, which is known as J48. We note here the we experimented with many values for the decision factor parameter and the best results are obtained when it is set to 0.2.
- The last one is a variant of the KNN algorithm known as the IBK algorithm. After several experiments, we reached a conclusion that using $K = 1$ returned the best results, which is in accordance with previous observations in the literature that argue that increasing the value of K will cause a degradation in accuracy as it decreases the distinct boundaries between categories [39].

As for the testing option, we use the holdout method since the dataset is large enough to allow such a choice. The dataset is split into a training set consisting of two thirds of the original dataset and a testing set consisting of the remaining third of the dataset. Specifically, the testing dataset consists of 21,508 instances distributed among the different classes in a way that preserves the percentages in the original dataset as follows. Class 1 has 981 instances, class 2 has 1,814 instances, class 3 has 4,181 instances, class 4 has 6,451 instances, and finally class 5 has 8,081 instances.

Table I reports the accuracy measures for the flat classification structure with different core classifiers. This table serves two important purposes. It allows us to related our results with existing approaches in the literature on the same dataset (which are all flat approaches), which helps us in arguing about our choices for the evaluation metrics. The second purpose is to allow us to compare the hierarchical structures with the flat structure to determine under which setting resorting to hierarchical classification pays off.

TABLE I: Accuracy of the flat classifier.

	Accuracy	Precision	Recall	F1	MSE
SVM	45.7%	45%	45%	45%	1.6
DT	40.2%	38.7%	40%	39%	1.74
NB	38.2%	36%	38%	37%	1.87
KNN	38.6%	36%	38%	37%	2.05

TABLE II: Accuracy of the 2-level hierarchical classifier.

	Accuracy	Precision	Recall	F1	MSE
SVM	45.2%	54%	45%	49%	0.86
DT	43.9%	54%	43%	48%	0.84
NB	39.9%	42.5%	44.7%	43.5%	2.04
KNN	46.2%	54%	46%	50%	0.9

As was reported in previous works on this dataset ([21], [15]), SVM gives the most accurate results. Note that the table shows non-negligible difference between SVM and DT in terms of the standard accuracy measures. However, the difference in terms of MSE is not as high which means that the mistakes DT is making are probably marginal mistakes. The same thing is observed when comparing NB and KNN. The standard accuracy measures suggests that KNN is slightly better than NB (which is unexpected for a text classification problem); however, MSE suggests otherwise, which is in accordance to what is known in the literature. These observations strengthen our argument that MSE is a more accurate measure of performance than the other four standard measures.

Table II reports the accuracy measures for the 2-level hierarchical classification structure with different core classifiers. The table shows seemingly conflicting results in terms of standard accuracy measures versus MSE. For the standard measures, SVM is negatively affected by the imposition of the first hierarchical strictures, whereas the other classifiers (especially, KNN) are positively affected. As argued before, these observations might be misleading. SVM might be making more mistakes with the use of the first hierarchical structure. But is this necessarily a bad thing? To answer this questions, we need to inspect SVM's mistakes. We do this by computing MSE. The tables shows that imposing the first hierarchical structures cut MSE in half. This is a significant improvement enjoyed by other classifiers. This means that even if the classifiers are not improving significantly in terms of their fine-grained decisions, they are improving in the sense that their mistakes are less severe. The only exception is the NB classifier, which seems like a single-shot classifier that is actually hurt by the decomposition of the original large problem into a set of smaller sub-problems in hierarchical classification.

TABLE III: Accuracy of the 4-level hierarchical classifier.

	Accuracy	Precision	Recall	F1	MSE
SVM	47.4%	65%	47%	55%	1.16
DT	47.6%	66%	47%	55%	1.54
NB	48.9%	70%	48%	57%	2.71
KNN	57.8%	70%	57%	63%	0.96

TABLE IV: Improvements in accuracy over the flat classifier.

	2-level	4-level
SVM	-01.2%	+3.7%
DT	+9.2%	+18.2%
NB	+4.6%	+28.1%
KNN	+19.7%	+49.7%

Table III reports the accuracy measures for the 4-level hierarchical classification structure with different core classifiers. Compared with the results of the two previously discussed classification structures, the results of this structure also seem to have conflicting observations in terms of standard accuracy measures versus MSE. For the standard measures, it looks like this structure is the best one with the accuracy of certain classifiers surpassing the best known results for this dataset. However, inspecting MSE shows that the results of this structure lies in between the results of the two previously discussed structures. Another observation of this table and the ones before it is that NB enjoys both improvement in terms of accuracy and degradation in terms of MSE with the increase in the number of levels in the hierarchical structures. So, it looks like increasing the number of levels allows NB to make less mistakes; however, the mistakes it makes are becoming more severe. On the other hand, the classifier that benefited the most from increasing the number of levels is KNN whose accuracy witnesses significant jumps with every increase in the number of levels.

To appreciate the effect of using a hierarchical classification approach, we report the improvements of each hierarchical structure over the flat structure in terms of both accuracy and MSE in Tables IV and V, respectively. These improvements are computed by taking the difference between the new values and the old values and dividing it by the old values.

The tables show that improvements of 50% or more are obtained in both measures for different settings. The tables show that the best improvement in terms of accuracy is about 50% and it is enjoyed by KNN when imposing the second hierarchical structure. As for the MSE, several settings shows improvements of more than 50% with the best improvements enjoyed by KNN. These results make hierarchical classification a very appealing approach to address a problem like MWSA of Arabic reviews.

TABLE V: Improvements in MSE over the flat classifier.

	2-level	4-level
SVM	+45.9%	+27.4%
DT	+51.8%	+11.5%
NB	-8.8%	-45%
KNN	+55.7%	+53%

V. Conclusion

In this work, we addressed the Multi-Way Sentiment Analysis (MWSA) problem for Arabic reviews. This important problem is yet to find sufficient interest within the research community of Arabic text processing and mining. Among the very limited existing works are a couple of papers following simple flat baseline approaches on a publicly available dataset (LABR). Motivated by the intuition that employing a flat classifier to handle an inherently hierarchical problem such as MWSA is one of the main reasons behind such poor accuracy, we proposed to use hierarchical classification. We presented two different hierarchical structures and compared their accuracy with the flat structure using different core classifiers. The comparison is based on standard accuracy measures such as precision and recall in addition to using the mean squared error (MSE) as a more accurate measure given the fact that not all misclassifications are the same. The results showed that, in general, hierarchical classifiers gave significant improvements (of more than 50% in certain cases) over flat classifiers.

References

[1] Muhammad Abdul-Mageed and Mona T Diab. Subjectivity and sentiment annotation of modern standard arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*, 2011.

[2] Muhammad Abdul-Mageed and Mona T Diab. Awatif: A multi-genre corpus for modern standard arabic subjectivity and sentiment analysis. In *LREC*, pages 3907–3914, 2012.

[3] Muhammad Abdul-Mageed, Mona T Diab, and Mohammed Korayem. Subjectivity and sentiment analysis of modern standard arabic. In *ACL*, 2011.

[4] Muhammad Abdul-Mageed et al. SAMAR: A system for subjectivity and sentiment analysis of arabic social media. In *WASSA*, 2012.

[5] Nawaf Abdulla et al. Arabic sentiment analysis: Corpus-based and lexicon-based. In *AEECT*. IEEE, 2013.

[6] Nawaf Abdulla et al. Automatic lexicon construction for arabic sentiment analysis. In *FiCloud*, 2014.

[7] Nawaf Abdulla et al. Towards improving the lexicon-based approach for arabic sentiment analysis. *International Journal of Information Technology and Web Engineering (IJITWE)*, 9(3):55–70, 2014.

[8] Nawaf A Abdulla, Mahmoud Al-Ayyoub, and Mohammed Naji Al-Kabi. An extended analytical study of arabic sentiments. *International Journal of Big Data Intelligence*, 1(1):103–113, 2014.

[9] Khurshid Ahmad, David Cheng, and Yousif Almas. Multi-lingual sentiment analysis of financial news streams. In *Grid in Finance*, 2006.

[10] Nizar Ahmed et al. Scalable multi-label arabic text classification. In *ICICS*. IEEE, 2015.

[11] Mohammad Al-A'abed and Mahmoud Al-Ayyoub. A lexicon-based approach for emotion analysis of arabic social media content. In *The International Computer Sciences and Informatics Conference (ICSIC)*, 2016.

[12] Mahmoud Al-Ayyoub, Safa Bani Essa, and Izzat Alsmadi. Lexicon-based sentiment analysis of arabic tweets. *International Journal of Social Network Mining (IJSNM)*, 2(2):101–114, 2015.

[13] Mohammed Al-Kabi et al. An opinion analysis tool for colloquial and standard arabic. In *ICICS*, 2013.

[14] Mohammed Al-Kabi et al. A prototype for a standard arabic sentiment analysis corpus. *The International Arab Journal of Information Technology*, 13(1A):163–170, 2016.

[15] Bashar Al Shboul, Mahmoud Al-Ayyoub, and Yaser Jararweh. Multi-way sentiment classification of arabic reviews. In *Information and Communication Systems (ICICS), 2015 6th International Conference on*, pages 206–211. IEEE, 2015.

[16] Mohammad Al-Smadi et al. Using aspect-based sentiment analysis to evaluate arabic news affect on readers. In *The 8th IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*, 2015.

[17] Mohammad Al-Smadi, Omar Qawasmeh, Bashar Talafha, and Muhannad Quwaider. Human annotated arabic dataset of book reviews for aspect based sentiment analysis. In *Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference on*, pages 726–730. IEEE, 2015.

[18] Yousif Almas and Khurshid Ahmad. A note on extracting sentiments in financial news in english, arabic & urdu. In *CAASL*, 2007.

[19] Izzat Alsmadi, Mohammed Naji Al-Kabi, Heider Wahsheh, and Bassima Bassam. Video spam and public opinion in current middle eastern conflicts. *International Journal of Social Network Mining*, 1(3):318–333, 2013.

[20] Kholoud Alsmearat et al. Emotion analysis of arabic articles and its impact on identifying the author's gender. In *ACS/IEEE AICCSA*, 2015.

[21] Mohamed A Aly and Amir F Atiya. Labr: A large scale arabic book reviews dataset. In *ACL (2)*, pages 494–498, 2013.

[22] Adrian Bickerstaffe and Ingrid Zukerman. A hierarchical classifier applied to multi-way sentiment detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 62–70. Association for Computational Linguistics, 2010.

[23] Minh Duc Cao and Ingrid Zukerman. Experimental evaluation of a lexicon- and corpus-based ensemble for multi-way sentiment analysis. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 52–60, Dunedin, New Zealand, December 2012.

[24] Samhaa R El-Beltagy and Ahmed Ali. Open issues in the sentiment analysis of arabic social media: A case study. In *IIT*, 2013.

[25] Alaa El-Halees. Arabic opinion mining using combined classification approach. In *ACIT*, 2011.

[26] Mohamed Elarnaoty, Samir AbdelRahman, and Aly Fahmy. A machine learning approach for opinion holder extraction in arabic language. *arXiv*, 2012.

[27] Mohamed Elhawary and Mohamed Elfeky. Mining arabic business reviews. In *ICDMW*. IEEE, 2010.

[28] Hady ElSahar and Samhaa R El-Beltagy. Building large arabic multi-domain resources for sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, pages 23–34. Springer, 2015.

[29] Mosab Faqeeh et al. Cross-lingual short-text document classification for facebook comments. In *FiCloud*, 2014.

[30] Noura Farra, Elie Challita, Rawad Abou Assi, and Hazem Hajj. Sentence-level and document-level sentiment mining for arabic texts. In *ICDMW*, 2010.

[31] Michael Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics, 2004.

[32] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical approach to emotion recognition and classification in texts. In *Advances in Artificial Intelligence*, pages 40–50. Springer, 2010.

[33] Andrew B Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52. Association for Computational Linguistics, 2006.

[34] Joshua Congfu He, Wee Kheng Leow, and Tet Sen Howe. Hierarchical classifiers for detection of fractures in x-ray images. In *Computer Analysis of Images and Patterns*, pages 962–969. Springer, 2007.

[35] Ismail Hmeidi et al. Automatic arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, 41(1):114–124, 2015.

[36] Maher M. Itani et al. Classifying sentiment in arabic social networks: Naive search versus naive bayes. In *ACTEA*, 2012.

[37] Mohammed Korayem, David Crandall, and Muhammad Abdul-Mageed. Subjectivity and sentiment analysis of arabic: A survey. In AboulElla Hassanien, Abdel-BadeehM. Salem, Rabie Ramadan, and Tai-hoon Kim, editors, *Advanced Machine Learning Technologies and Applications*, volume 322 of *Communications in Computer and Information Science*, pages 128–139. Springer Berlin Heidelberg, 2012.

[38] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[39] Ashis Kumar Mandal and Rikta Sen. Supervised learning methods for bangla web document categorization. *International Journal of Artificial Intelligence and Applications (IJAIA)*, 5(5), Sept 2014.

[40] D Ben Ayed Mezghani, S Zribi Boujelbene, and N Ellouze. Evaluation of svm kernels and conventional machine learning algorithms for speaker identification. *International Journal of Hybrid Information Technology*, 3(3):23–34, 2010.

[41] Asmaa Mountassir, Houda Benbrahim, and Ilham Berrada. An empirical study to address the problem of unbalanced data sets in sentiment classification. In *SMC*. IEEE, 2012.

[42] Mahmoud Nabil, Mohamed Aly, and Amir F Atiya. Astd: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, 2015.

[43] Islam Obaidat et al. Enhancing the determination of aspect categories and their polarities in arabic reviews using lexicon-based approaches. In *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2015.

[44] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[45] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2), 2008.

[46] Jonathon Read and John Carroll. Weakly supervised techniques for domain-independent sentiment classification. In *TSA*. ACM, 2009.

[47] Eshrag Refaee and Verena Rieser. Subjectivity and sentiment analysis of arabic twitter feeds with limited resources. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 16, 2014.

[48] Mohammed Rushdi-Saleh, M Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. Oca: Opinion corpus for arabic. *JASIST*, 62(10):2045–2054, 2011.

[49] Mohammed Rushdi-Saleh, Maria Teresa Martín-Valdivia, L Alfonso Ureña-López, and José M Perea-Ortega. Bilingual experiments with an arabic-english corpus for opinion mining. In *RANLP*, 2011.

[50] Amira Shoukry and Ahmed Rafea. Sentence-level arabic sentiment analysis. In *Collaboration Technologies and Systems (CTS), 2012 International Conference on*, pages 546–550. IEEE, 2012.

[51] Benjamin Snyder and Regina Barzilay. Multiple aspect ranking using the good grief algorithm. In *HLT-NAACL*, pages 300–307, 2007.

[52] Aixin Sun and Ee-Peng Lim. Hierarchical text classification and evaluation. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 521–528. IEEE, 2001.

[53] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. Performance measurement framework for hierarchical text classification. *Journal of the American Society for Information Science and Technology*, 54(11):1014–1028, 2003.

[54] Shrawan Kumar Trivedi, Shubhamoy Dey, and Prabandh Shikhar. Effect of various kernels and feature selection methods on svm performance for detecting email spams. *International Journal of Computer Applications*, 66(21), 2013.

[55] Mohammed Abdul Wajeed and T Adilakshmi. Text classification using machine learning. *Journal of Theoretical and Applied Information Technology*, 7(2):119–123, 2009.

[56] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

Weighted G^1-Multi-Degree Reduction of Bézier Curves

Abedallah Rababah
Department of Mathematics,
Jordan University of Science and Technology
Irbid 22110 Jordan

Salisu Ibrahim
Department of Mathematics,
Jordan University of Science and Technology
Irbid 22110 Jordan

***Abstract*—In this paper, weighted G^1-multi-degree reduction of Bézier curves is considered. The degree reduction of a given Bézier curve of degree n is used to write it as a Bézier curve of degree $m, m < n$. Exact degree reduction is not possible, and, therefore, approximation methods are used. The weight function $w(t) = 2t(1-t),\ t \in [0,1]$ is used with the L_2-norm in multi-degree reduction with G^1-continuity at the end points of the curve. Since we consider boundary conditions this weight function improves approximation in the middle of the curve. Numerical results and comparisons show that the proposed method produces fewer errors and outperform all existing methods.**

Keywords*—*Bézier curves; multiple degree reduction; G*^{1}*-continuity; geometric continuity

I. Introduction

Degree reduction of Bézier curves and surfaces is an important issue in Computer Aided Geometric Design (CAGD) that is tackled by many researchers. It facilitates data exchange, compression, transfer, and comparison. In degree reduction, we approximate a Bézier curve of degree n by a Bézier curve of degree $m, m < n$; moreover, the boundary conditions have to be satisfied and gives minimum error. The struggles of finding a solution are disturbed by the requirement of solving a non-linear problem, in which numerical methods have to be used. In 2000, J. Peter and U. Reif proved in [5] that degree reduction of Bézier curves in the L_2 norm equals best Euclidean approximation of Bézier points. These results are generalized to the constrained case by Ahn et. al. in [1], and discrete cases have been studied in [2], [8].
The existing methods to find degree reduction have many issues including accumulate round-off errors, stability issues, complexity, accuracy, losing conjugacy, requiring the search direction to be set to the steepest descent direction frequently, experiencing ill-conditioned systems, leading to a singularity, and the most challenging difficulty is in applying the methods (difficulty and indirect). A. Rababah and S. Mann presented in [10] a method to find the G^2-degree reduction for Bézier Curves based on exploiting additional parameters as in [7]. These results are expressive to researchers as well as to industrial practitioners. Their examples show that the C^2 method fails to reproduce the inner loop of the heart, while their C^1/G^2 method reproduces the loop and provides a better approximation elsewhere along the curve.

In all existing degree reducing methods, the conditions and free parameters were applied at the end points. So, there is a need to better approximate those parts close to the centre of the curve. In this paper, we introduce a weight to take care of the centre of the curve, it is appropriate to consider degree reduction with the weight function $w[t] = 2t(1-t)$, $t \in [0,1]$. The examples show that the proposed methods provide better approximation at the centre of the curves with minimum error and also reproduced these loops correctly better than existing methods.

II. Preliminaries

A Bézier curve $P_n(t)$ of degree n is defined algebraically as follows:

$$P_n(t) = \sum_{i=0}^{n} p_i B_i^n(t), \qquad 0 \le t \le 1, \tag{1}$$

where

$$B_i^n(t) = \binom{n}{i}(1-t)^{n-i}t^i, \quad i = 0, 1, \ldots, n,$$

are the Bernstein polynomials of degree n, and $p_0, p_1, \ldots, p_n$ are called the Bézier control points or the Bézier points, for more see [4].

The first derivative of the Bézier curve is given by:

$$\frac{d}{dt}P_n(t) = n\sum_{i=0}^{n-1} \Delta p_i B_i^{n-1}(t),$$

where

$$\Delta p_i = p_{i+1} - p_i,\ i = 0, 1, \ldots, n-1.$$

The multiplication of two Bernstein polynomials with the weight function $w(t) = 2t(1-t)$ is given by

$$B_i^m(t)B_j^n(t)2t(1-t) = \frac{2\binom{m}{i}\binom{n}{j}}{\binom{m+n+2}{i+j+1}} B_{i+j+1}^{m+n+2}(t). \tag{2}$$

We define the Gram matrix $G_{m,n}$ as $(m+1) \times (n+1)$-matrix with weight function as follows:

$$\begin{aligned} g_{ij} &= \int_0^1 B_i^m(t)B_j^n(t)2t(1-t)dt \\ &= \frac{2\binom{m}{i}\binom{n}{j}}{(m+n+3)\binom{m+n+2}{i+j+1}}, i = 0, \ldots, m,\ j = 0, \ldots, n. \end{aligned} \tag{3}$$

The matrix $G_{m,m}$ with weight function is real, symmetric, and positive definite like the case in [10].

III. Degree Reduction of Bézier curves

Degree reduction is approximating a given Bézier curve of degree n by a Bézier curve of degree $m, m < n$. It is approximative process in nature and exact degree reduction is not possible. In this paper, our aim is to find a Bézier curve $R_m(t)$ of degree m with control points $\{r_i\}_{i=0}^m$ that approximates a given Bézier curve $P_n(t)$ of degree n with control points $\{p_i\}_{i=0}^n$, where $m < n$. The Bézier curve R_m has to satisfy the following two conditions:

1) P_n and R_m are G^1-continuous at the end points, and
2) the weighted L_2-error between P_n and R_m is minimum.

We can write the two Bézier curves $P_n(t)$ and $R_m(t)$ in matrix form as

$$P_n(t) = \sum_{i=0}^{n} p_i B_i^n(t) =: B_n P, \quad 0 \le t \le 1, \tag{4}$$

and similarly

$$R_m(t) = \sum_{i=0}^{m} r_i B_i^m(t) =: B_m R, \quad 0 \le t \le 1.$$

In the following sections we investigate the case of G^1-continuity with weighted degree reduction of Bézier curves.

IV. Weighted G^1-Degree Reduction

$P_n(t)$ and $R_m(t)$ are G^1-continuous at $t = 0, 1$ if they satisfy the following conditions

$$R_m(i) = P_n(s(i)), \quad i = 0, 1. \tag{5}$$

$$R'_m(i) = s'(i) P'_n(s(i)), \quad s'(i) > 0, \quad i = 0, 1. \tag{6}$$

This means that the two curves P_n and R_m have to have common end points

$$r_0 = p_0, \quad r_m = p_n,$$

and the direction of the tangent at the two end points of P_n and R_m should coincide, but they need not to be of equal length. To simplify the problem and have a linear system, the authors in [10] used $s'(i) = \delta_i$, for $i = 0, 1$. We analogously use these substitutions for the case of weighted degree reduction to get

$$R'_m(i) = \delta_i P'_n(i), \quad i = 0, 1. \tag{7}$$

Using $s'(i) = \delta_i$, $i = 0, 1$, we can solve (5) and (7) for the two control points at either ends of the curve to get

$$r_0 = p_0,$$

$$r_1 = p_0 + \frac{n}{m} \Delta p_0 \delta_0,$$

$$r_{m-1} = p_n - \frac{n}{m} \Delta p_{n-1} \delta_1,$$

$$r_m = p_n.$$

The points r_0, r_1, r_{m-1} and r_m are determined by G^1-continuity conditions at the boundaries; accordingly, the elements of R_m can be decomposed into two parts stated as follows. The boundaries part $R_m^c = [r_0, r_1, r_{m-1}, r_m]^t$ and the interior part with interior points $R_m^f = R_m \backslash R_m^c = [r_2, \ldots, r_{m-2}]^t$. Similarly, B_m is decomposed in the same way into B_m^c and B_m^f.

The weighted distance between P_n and R_m is measured using weighted L_2-norm; therefore, the error term becomes:

$$\begin{aligned} \varepsilon &= \int_0^1 ||B_n P_n - B_m R_m||^2 \; 2t(1-t) dt \\ &= \int_0^1 ||B_n P_n - B_m^c R_m^c - B_m^f R_m^f||^2 \; 2t(1-t) dt. \end{aligned} \tag{8}$$

Differentiating ε with respect to the unknown control points R_m^f we get

$$\frac{\partial \varepsilon}{\partial R^f} = 2 \int_0^1 ||B_n P_n - B_m^c R_m^c - B_m^f R_m^f|| \; B_m^f \; 2t(1-t) \; dt.$$

Evaluating the integral and equating to zero gives

$$\frac{\partial \varepsilon}{\partial R^f} = G_{m,n}^p P_n - G_{m,m}^c R_m^c - G_{m,m}^f R_m^f = 0, \tag{9}$$

where

$$\begin{aligned} G_{m,n}^p &:= G_{m,n}(2, \ldots, m-2; 0, 1, \ldots, n), \\ G_{m,m}^c &:= G_{m,m}(2, \ldots, m-2; 0, 1, m-1, m), \\ G_{m,m}^f &:= G_{m,m}(2, \ldots, m-2; 2, \ldots, m-2), \end{aligned}$$

and $G_{m,n}(\ldots;\ldots)$ is the sub-matrix of $G_{m,n}$ formed by the indicated rows and columns.

Differentiating ε with respect to δ_i and equating to zero gives

$$\frac{\partial \varepsilon}{\partial \delta_0} = \left(G_{m,n}^1 P_n - G_{m,m}^{1;c} R_m^c - G_{m,m}^{1;f} R_m^f \right) \cdot \Delta p_0 = 0, \tag{10}$$

$$\frac{\partial \varepsilon}{\partial \delta_1} = \left(G_{m,n}^{m-1} P_n - G_{m,m}^{m-1;c} R_m^c - G_{m,m}^{m-1;f} R_m^f \right) \cdot \Delta p_{n-1} = 0, \tag{11}$$

where for $q = 1, m-1$:

$$\begin{aligned} G_{m,n}^q &:= G_{m,n}(q; 0, 1, \ldots, n), \\ G_{m,m}^{q;c} &:= G_{m,m}(q; 0, 1, m-1, m), \\ G_{m,m}^{q;f} &:= G_{m,m}(q; 2, \ldots, m-2). \end{aligned} \tag{12}$$

Note that (9) are point valued equations while (10) and (11) are scalar valued equations, expanding (9) into its $x, y, z, \ldots$ coordinates and joining them together with (10) and (11) yields $d(m-3)+2$ equations in $d(m-3)+2$ unknowns, see Rababah-Mann [10].

In the planar case, the control points of the Bézier curve are expanded into their x and y components. Therefore, the variables of our system of equations are $r_k^x, r_k^y, k = 2, \ldots, m-2$, δ_0 and δ_1. To express the system in a clear form, we have to decompose each of r_1 and r_{m-1} into a constant part and a part involving δ_0 and δ_1, respectively. Let v_1 and v_{m-1} be the constant parts of r_1 and r_{m-1} respectively. Hence

$$v_1 \quad = p_0, \quad v_{m-1} = p_n.$$

The following vectors are defined to express the linear system in explicit form:

$$\begin{array}{rcl} P_n^C & = & [p_0^x, ..., p_n^x, p_0^y, ..., p_n^y]^t, \\ R_m^F & = & [r_2^x, \ldots, r_{m-2}^x, r_2^y, \ldots, r_{m-2}^y, \delta_0, \delta_1]^t, \\ R_m^C & = & [r_0^x, \ v_1^x, \ v_{m-1}^x, \ r_m^x, \ \ r_0^y, \ v_1^y, \ v_{m-1}^y, \ r_m^y]^t. \end{array}$$

Let $\oplus$ be the direct sum. Define the matrices

$$\begin{array}{rcl} G_{m,n}^{p+} & = & G_{m,n}^p \oplus G_{m,n}^p, \\ G_{m,m}^{c+} & = & G_{m,m}^c \oplus G_{m,m}^c, \\ G_{m,m}^{f+} & = & G_{m,m}^f \oplus G_{m,m}^f. \end{array} \tag{13}$$

The Gram matrix $G_{m,m}^{f+}$ has the same properties of the matrix $G_{m,m}^f$.
Write $G := G_{m,m}$ and define

$$C = \left[\begin{array}{ll} \Delta p_0 \Delta p_0 G(1,1) & \Delta p_0 \Delta p_{n-1} G(1, m-1) \\ \Delta p_0 \Delta p_{n-1} G(m-1, 1) & \Delta p_{n-1} \Delta p_{n-1} G(m-1, m-1) \end{array} \right],$$

$$= \left[\begin{array}{cc} \Delta p_0 & 0 \\ 0 & \Delta p_{n-1} \end{array} \right] \left[\begin{array}{cc} G(1,1) & G(1, m-1) \\ G(m-1,1) & G(m-1, m-1) \end{array} \right] \times \left[\begin{array}{cc} \Delta p_0 & 0 \\ 0 & \Delta p_{n-1} \end{array} \right].$$

Further define $L_{m,n}, L_{m,m}^c, L_{m,m}^f$ as

$$L_{m,n} = \left[\begin{array}{cc} G_{m,n}^1 \Delta p_0^x & G_{m,n}^1 \Delta p_0^y \\ G_{m,n}^{m-1} \Delta p_{n-1}^x & G_{m,n}^{m-1} \Delta p_{n-1}^y \end{array} \right],$$

$$L_{m,m}^c = \left[\begin{array}{cc} G_{m,m}^{1;c} \Delta p_0^x & G_{m,m}^{1;c} \Delta p_0^y \\ G_{m,m}^{m-1;c} \Delta p_{n-1}^x & G_{m,m}^{m-1;c} \Delta p_{n-1}^y \end{array} \right],$$

$$L_{m,m}^f = \left[\begin{array}{cc} G_{m,m}^{1;f} \Delta p_0^x & G_{m,m}^{1;f} \Delta p_0^y \\ G_{m,m}^{m-1;f} \Delta p_{n-1}^x & G_{m,m}^{m-1;f} \Delta p_{n-1}^y \end{array} \right],$$

Further define $L_{m,n}, L_{m,n}^f$ as

$$L_{m,n} = \left[\begin{array}{cc} G_{m,n}^2 \Delta p_0^x & G_{m,n}^2 \Delta p_0^y \\ G_{m,n}^{m-2} \Delta p_{n-1}^x & G_{m,n}^{m-2} \Delta p_{n-1}^y \end{array} \right],$$

$$L_{m,n}^f = \left[\begin{array}{cc} G_{m,n}^{c;2} \Delta p_0^x & G_{m,n}^{c;2} \Delta p_0^y \\ G_{m,n}^{c;m-2} \Delta p_{n-1}^x & G_{m,n}^{c;m-2} \Delta p_{n-1}^y \end{array} \right],$$

where $G_{m,n}^q, G_{m,m}^{q;c}$, and $G_{m,n}^{q;f}$ are defined in (12). The matrices C, $L_{m,n}$, $L_{m,m}^c$, and $L_{m,m}^f$ are obtained from (10) and (11), (the derivatives with respect to the δ_is).

The coordinate form of the expansion of (9) becomes

$$G_{m,m}^F R_m^F = G_{m,n}^{PC} P_n^C - G_{m,m}^C R_m^C, \tag{14}$$

where

$$G_{m,n}^{PC} = \left[\begin{array}{c} G_{m,n}^{p+} \\ L_{m,n} \end{array} \right],$$

$$G_{m,m}^C = \left[\begin{array}{c} G_{m,m}^{c+} \\ L_{m,m}^c \end{array} \right],$$

$$G_{m,m}^F = \left[\begin{array}{ll} G_{m,m}^{f+} & \frac{n}{m} (L_{m,m}^f)^t \\ L_{m,m}^f & \frac{n}{m} C \end{array} \right].$$

From (14) and because $G_{m,m}^F$ is invertible, we can find our unknowns as

$$R_m^F = (G_{m,m}^F)^{-1} \Big(G_{m,n}^{PC} P_n^C - G_{m,m}^C R_m^C \Big). \tag{15}$$

V. Applications

In this section, some examples are given to illustrate the effectiveness of the proposed method of weighted G^1-degree reduction. Comparisons with other existing methods are also presented in this section.

Example 1: Given the Bézier curve (spiral) $P_n(t)$ of degree 19 with the control points, see Fig. 11 in [10]:
$\mathbf{P}_0 = (37, 38)$, $\mathbf{P}_1 = (43, 37)$, $\mathbf{P}_2 = (39, 27)$, $\mathbf{P}_3 = (29, 26)$, $\mathbf{P}_4 = (23, 36)$,
$\mathbf{P}_5 = (26, 50)$, $\mathbf{P}_6 = (45, 56)$, $\mathbf{P}_7 = (58, 47)$, $\mathbf{P}_8 = (58, 29)$, $\mathbf{P}_9 = (46, 14)$,
$\mathbf{P}_{10} = (26, 6)$, $\mathbf{P}_{11} = (5, 15)$, $\mathbf{P}_{12} = (0, 40)$, $\mathbf{P}_{13} = (3, 58)$, $\mathbf{P}_{14} = (24, 68)$,
$\mathbf{P}_{15} = (50, 75)$, $\mathbf{P}_{16} = (79, 69)$, $\mathbf{P}_{17} = (79, 36)$, $\mathbf{P}_{18} = (65, 12)$, $\mathbf{P}_{19} = (50, 0)$,

It is reduced to Bézier curve $R_m(t)$ of degree 8. Fig. 1 depicts the original curve in solid-blue and weighted G^1-degree reduction in dashed-red curve. Fig. 2 shows the curves with control polygons; original curve (dashed-Black); weighted G^1 (dashed-Green). Fig. 3 shows the error plots for weighted G^1-degree reduction in Example 1.

Example 2: Given the Bézier curve $P_n(t)$ of degree 10 with the control points, see [6]:
$\mathbf{P}_0 = (0, 1.2)$, $\mathbf{P}_1 = (0.04, 0.6)$, $\mathbf{P}_2 = (0.15473790322581, 0.507)$,
$\mathbf{P}_3 = (0.32207661290323, 0.878)$, $\mathbf{P}_4 = (0.30897177419355, 0.086)$,
$\mathbf{P}_5 = (0.51864919354839, 0)$, $\mathbf{P}_6 = (0.62449596774194, 0.8)$, $\mathbf{P}_7 = (0.89, 0.874)$,
$\mathbf{P}_8 = (0.92, 0.6)$, $\mathbf{P}_9 = (0.92, 0.3)$, $\mathbf{P}_{10} = (0.75352822580645, 0)$.
This curve (blue) is reduced to Bézier curve (red) $R_m(t)$ of degree 6 using Weighted G^1 method. The corresponding degree reduced Bézier curve plot is depicted in Fig. 4 and the error plot is depicted in Fig. 5.

Example 3: Given the Bézier curve $P_n(t)$ of degree 13 with double loop control points, see [10]:
$\mathbf{P}_0 = (4, 9)$, $\mathbf{P}_1 = (23, 2)$, $\mathbf{P}_2 = (49, 19)$, $\mathbf{P}_3 = (67, 20)$, $\mathbf{P}_4 = (52, 48)$,
$\mathbf{P}_5 = (0, 23)$, $\mathbf{P}_6 = (26, 0)$, $\mathbf{P}_7 = (71, 4)$, $\mathbf{P}_8 = (71, 37)$, $\mathbf{P}_9 = (30, 54)$,
$\mathbf{P}_{10} = (4, 25)$, $\mathbf{P}_{11} = (24, 5)$, $\mathbf{P}_{12} = (41, 0)$, $\mathbf{P}_{13} = (62, 1)$,
This curve (solid-Blue) is reduced to Bézier curve of degree 8

(dashed-Red) using weighted G^1 method, see Fig. 6. Comparing this plot of double loop with the example from [10] shows that our method produces better approximations and makes the loops that other methods did not.

Example 4: This example focuses on a "heart" data set , given a Bézier curve $P_n(t)$ of degree 13 with control points; see [10].
$\mathbf{P}_0 = (22, 10)$, $\mathbf{P}_1 = (37, 5)$, $\mathbf{P}_2 = (86, 18)$, $\mathbf{P}_3 = (81, 23)$, $\mathbf{P}_4 = (69, 56)$,
$\mathbf{P}_5 = (14, 26)$, $\mathbf{P}_6 = (40, 3)$, $\mathbf{P}_7 = (85, 7)$, $\mathbf{P}_8 = (85, 40)$, $\mathbf{P}_9 = (44, 57)$,
$\mathbf{P}_{10} = (18, 29)$, $\mathbf{P}_{11} = (38, 9)$, $\mathbf{P}_{12} = (55, 3)$, $\mathbf{P}_{13} = (77, 5)$.
The heart (solid-Blue) is reduced to Bézier curve of degree 8 (dashed-Red) using weighted G^1-degree reduction. The corresponding degree reduced Bézier curves and the example of heart in [10] are depicted in Fig. 7. Again the plot of double loop example from [10] shows that our method produces better approximations and makes the loops what other methods did not.

The examples show that considering a weight with geometric degree reduction is of great benefit. The results are better than the equivalent methods of C^1/G^k-methods considered by [10].

VI. Conclusion

In this paper, we have presented a method of weighted G^1-degree reduction. The weighted G^1-degree reduction is better than the G^1-degree reduction method in [10]. Referring to the examples in Fig. 1, Fig 3, Fig. 4, Fig 6, and Fig 7, the weighted G^1-degree reduction is the best approximation and provides less error than the linear G^1-degree reduction method and the linear C^1/G^2 method in [10]. The examples in Fig 3, Fig. 4, Fig 6, and Fig 7 show the effectiveness of our proposed weighted G^1-degree reduction method. Our weighted G^1-degree reduction reproduced these loops correctly and is better than existing methods.

Acknowledgment

The authors would like to thank the reviewers for invaluable comments that improve the manuscript.

References

[1] Y. Ahn, B.G. Lee, Y. Park, and J. Yoo, *Constrained polynomial degree reduction in the L_2-norm equals best weighted Euclidean approximation of Bézier coefficients*, Computer Aided Geometric Design 21: 181-191, 2004.

[2] R. Ait-Haddou, *Polynomial degree reduction in the discrete L_2-norm equals best Euclidean approximation of h-Bézier coefficients*, BIT Numer Math. http://dx.doi.org/10.1007/s10543-015-0558-9, 2015.

[3] R.T. Farouki and V.T. Rajan, *Algorithms for polynomials in Bernstein form*, Computer Aided Geometric Design 5: 1-26, 1988.

[4] K. Höllig, J. Hörner, *Approximation and Modeling with B-Splines*, SIAM, Titles in Applied Mathematics 132, 2013.

[5] J. Peters and U. Reif, *Least squares approximation of Bézier coefficients provides best degree reduction in the L_2-norm*, Journal of Approximation Theory 104: 90-97, 2000.

[6] L. Lu and G. Wang, *Optimal multi-degree reduction of Bézier curves with G^2-continuity*, Computer Aided Geometric Design 23: 673-683, 2006.

[7] A. Rababah, *Taylor theorem for planer curves*, Proceedings of the American Mathematical Society, Vol. 119, No.3: 803-810, 1993.

[8] A. Rababah, *Distances with rational triangular Bézier surfaces*, Applied Mathematics and Computation 160, 379-386, 2005.

[9] A. Rababah, B.G. Lee, and J. Yoo, *Multiple degree reduction and elevation of Bézier curves using Jacobi-Bernstein basis transformations*, Numerical Functional Analysis and Optimization 28(9-10): 1179-1196, 2007.

[10] A. Rababah and S. Mann, *Linear Methods for G^1, G^2, and G^3-Multi-Degree Reduction of Bézier Curves*, Computer-Aided Design 45(2):405414, 2013.

[11] P. Woźny and S. Lewanowicz, *Multi-degree reduction of Bézier curves with constraints, using dual Bernstein basis polynomials*, Computer Aided Geometric Design. 26: 566-579, 2009.

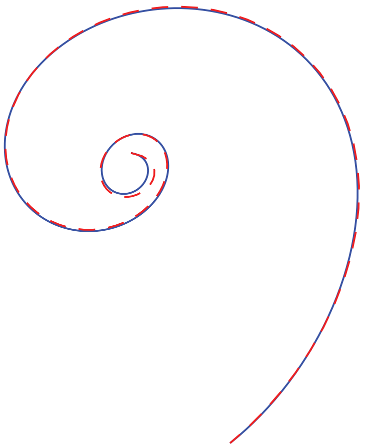


Fig. 1. Spiral Curve: Original curve degree 19 (Solid-Blue); reduced to degree 8 with weighted G^1-degree reduction (dashed-Red).

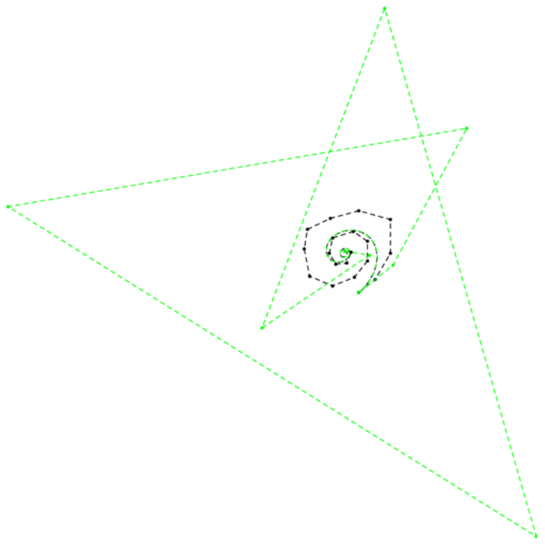


Fig. 2. Curves with control polygons; original curve (dashed-Black); weighted G^1 (dashed-Green).

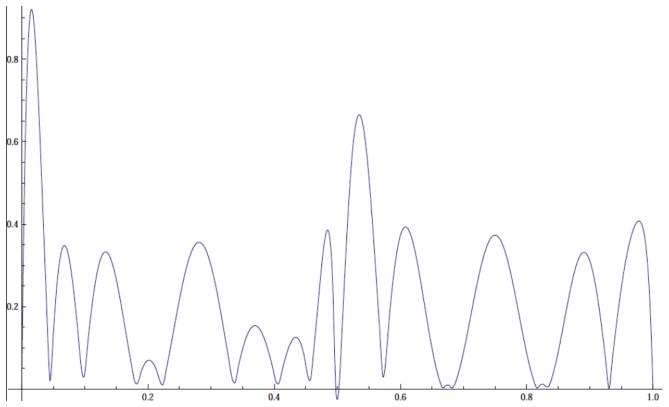


Fig. 3. Error plots for weighted G^1-degree reduction in Example 1.

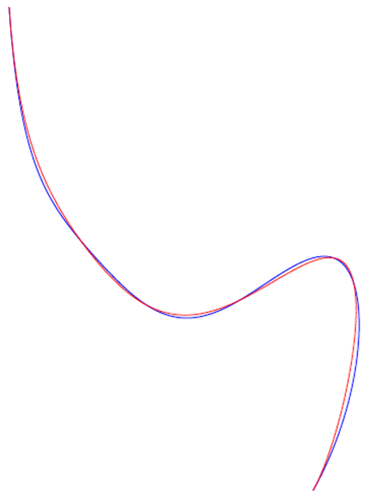


Fig. 4. Curve of degree 10 (Blue) reduced to degree 6 with weighted G^1 method (Red).

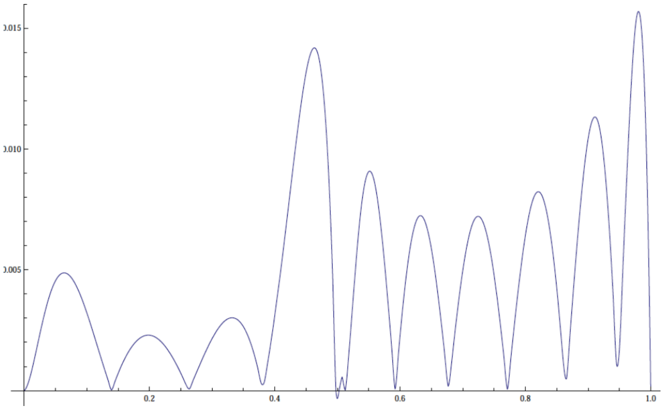


Fig. 5. Error plot for weighted G^1-degree reduction in Example 2.

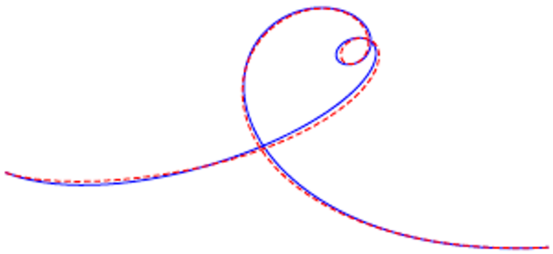


Fig. 6. Original curve in (solid-Blue); weighted G^1-degree 13 reduce to 8 (dashed-Red).

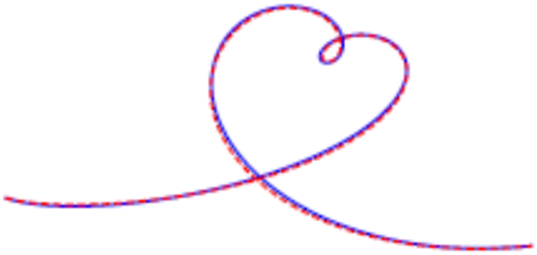


Fig. 7. Original curve in (solid-Blue); weighted G^1-degree 13 reduce to 8 (dashed-Red).

Abnormalities Detection in Digital Mammography Using Template Matching

Ahmed M. Farrag
Faculty of Computers and Information, Helwan University, Egypt

***Abstract*—Breast cancer affects 1 in 8 women in the United States. Early detection and diagnosis is key to recovery. Computer-Aided Detection (CAD) of breast cancer helps decrease morbidity and mortality rates. In this study we apply Template Matching as a method for breast cancer detection to a novel data set comprised of mammograms annotated according to ground truth. Performance is evaluated in terms of Area Under the Receiver Operator Characteristic Curve (Area Under ROC) and Free-response ROC.**

***Index Terms*—Classification, Detection, Image Processing, Digital Mammography, Breast Cancer, Computer Aided Detection (CAD), Template Matching.**

I. Introduction

Breast cancer is a disease that causes cells of the breast tissue to behave abnormally and grow out of control. They start invading the breast tissue and spreading (metastasizing) to other organs of the body. Breast tumors or Masses can be benign or malignant. Benign masses are characterized by their oval well-defined boundary, while malignant masses have a more speculated boundary.

Breast cancer is the most common cancer among women in the U.S. (excluding skin cancer) [1]. In Egypt, 38.8% of cancer in women is breast cancer (ranking first for women) [2]. Early detection of breast cancer is key to higher survival rates. Women above 45 with average risk of getting the disease are advised to be screened twice every year. However, radiologists' inexperience, fatigue, inattention and haste lead to false-negatives [3].

Computer-Aided Detection (CAD) systems are systems that process digital or digitized images and mark suspicious areas that the radiologist should pay attention to. Breast cancer CAD helps detect otherwise missed breast carcinoma. In a study by [4], the use of CAD resulted in 19.5% more malignant cases being diagnosed. However, [5]'s study showed no change in detection accuracy and recall rates with and without CAD. This can be explained by the difference in performance. "Highly performing" CAD was shown to improve radiologists' performance, while "poor performing" CAD negatively affected their performance [6]. This shows how important designing a highly performing CAD system is.

A. Data Collection

Our data comprises of only digital mammograms. Our radiologist consultant provides, reviews and annotates each image. She marks the boundaries of the tumor and assigns its category. Markings are done according to ground truth, i.e. all subjects were biopsied prior to the radiologist's marking. The tumor is categorized according to the "Breast Imaging Reporting and Data System" (BIRADS) scoring system (Table I). A special in-house tool was prepared specifically for this purpose. Both lesions and images were classified by the radiologist according to their BIRAD score.

Category 0	mammographic assessment is incomplete
Category 1	negative
Category 2	benign finding(s)
Category 3	probably benign finding(s)
Category 4	suspicious abnormality
Category 5	highly suggestive of malignancy

TABLE I: "Breast Imaging Reporting and Data System" (BIRADS) scoring.

B. Related Work

Template Matching is categorized as a model-based method of segmentation, meaning that it involves training and learning. Here, the training is to choose the best method parameters that yield the best result according to some assessment criteria (See section III). As early as 1989, Lai et al. [7] used template matching for detecting circumscribed masses in mammograms. Ng and Bischof in 1992 [8] used the same method. The problem with that approach was the disregard to different masses sizes. This problem is addressed in this paper by multiscale templates described in section II.

The previously mentioned methods use correlation as a similarity measure. A different similarity measure is mutual information. It has been used by Tourassi et al. in 2003 [9].

Oliver et al. in 2006[10] and in 2008[11] proposed a probablistic template to match against. The template is learned from different training tumors.

Brake et al. in 1999[12] proposed a multi-scale approach which we use in our study. Multiple templates of different sizes are used and the correlation to each template is calculated.

The rest of the paper is organized as follows. In section II we detail our approach. In section III, we describe our assessment methods, namely the Area Under the ROC (AUC) and the FROC. In Section IV we show our results. In Section V we conclude the paper and discuss our recommendations for future work.

II. Design

In this section, we introduce the method used in our study. The method belongs to the pixel-based algorithms. Pixel-based algorithms take as input features the plain grey-level of each pixel and its surrounding pixels up to L levels. In contrast, region-based algorithms use some image processing techniques to first segment the image into sub-regions which are then used for feature extraction. Classification for pixel-based algorithms is done per pixel, while in region-based algorithms it is done for the entire region.

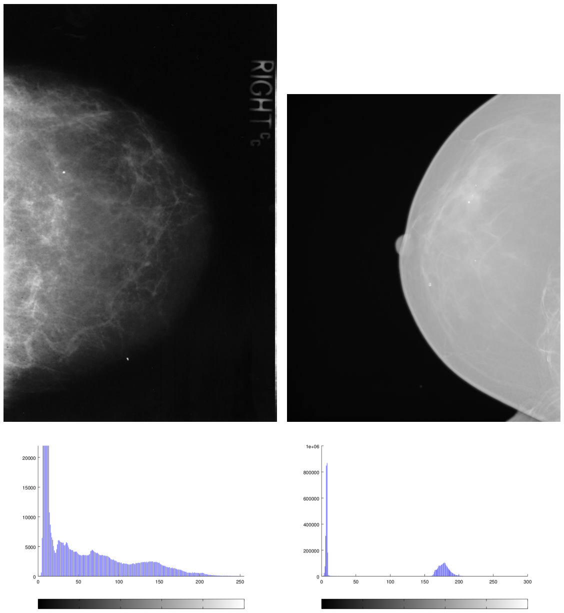


Fig. 1: A figure that illustrates the difference between analogue and digital mammography. To the left is an analogue image and below it is its histogram. The right is the same but with a digital image. It's evident how the gray levels span a much broader range in an analogue image.

Before we start taking grey-level values as our features, preprocessing of all images is carried out. Images are captured using different imaging machines having different sensors with different sensitivity to radiation. Thus, images' grey-level values need to be normalized first so that comparing their values would make sense. In addition to normalization, the breast needs to be segmented away from the background. In our study we used digital mammograms only; those mammograms have no artifacts and the background is almost all zeros making a whole literature on segmenting the breast in analogue mammograms irrelevant. In our study, simple methods like K-means or Otsu's thresholding[13] were sufficient.

A. Template Matching

Template Matching is carried out by sliding a window of size $Z \times Z$ containing the desired template over the pixels of the image. Each pixel of the image gets a score according to how similar it and its surrounding pixels are to the template. This similarity is calculated in terms of correlation, which measures the covariance between the subimage and the template normalized by their variances. Examples of templates that have been used for breast cancer detection in the literature are the spherical template (Figure 2(a))[12] and the hyperbolic secant template (Figure 2(b))[14]

The spherical template is defined as

$$T\left(x,y\right) = \begin{cases} R^2 - x^2 - y^2, & \text{if } x^2 + y^2 < R^2 \\ 0, & \text{if } x^2 + y^2 \geq R^2 \end{cases},$$

where R is the template radius. The hyperbolic secant template is defined as

$$T\left(x,y\right) = sech(x+y).$$

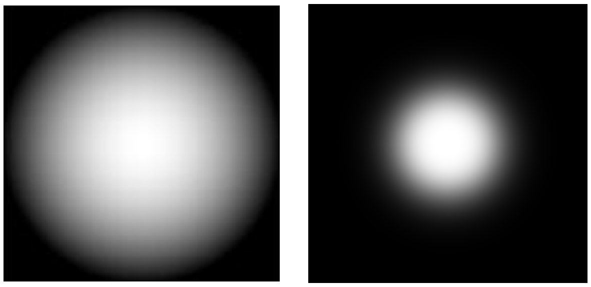
Different template sizes have been tested starting at radius $R = 4\% \ldots 12\%$ of the image height. The multiresolution analysis conducted by Brake et al.[12] has been adopted. The scores for the same pixel produced by correlation of different template sizes have been combined. The final score of the pixel is the maximum of those combined scores. The combination of template sizes $(6\%, 8\%, 10\%)$ yielded best results in terms of an AUC of 0.8656 (AUC is described in III).

III. Assessment

A. AUC

Different classification methods produce scores for every pixel (or region in case of a region-based approach as discussed in II). A certain threshold has to be selected above which the pixel is labeled Class A and below which it is labeled Class B. At different thresholds, some pixels are True Positives (correctly labeled as positives), and some are False Positives. Plotting the True Positive Fraction (TPF) against the False Positive Fraction (FPF) as the decision threshold is varied is called Receiver Operating Characteristic (ROC) curve [15].

The Area Under the ROC Curve (AUC) is a single number that can describe the performance of a classification method. It conveys how separable and distinguishable the two classes have become after using the method. An AUC of one indicates total separability. The AUC of a score image (probability images) is



(a) Spherical Template (b) Hyperbolic Secant Template

Fig. 2: A rendering of different templates

defined as

$$\widehat{AUC} = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} I_{(x_i<y_j)},$$

$$I_{(c)} = \begin{cases} 1, & \text{if c is true} \\ 0, & \text{if c is false} \end{cases}$$

where x is the set of scores of the normal region, y is the set of scores of the malignant region(s), and n_1 and n_2 are the respective number of pixels. It should be noted that the AUC assessment is a pixel-based assessment.

B. FROC

Free Response Operating Characteristic (FROC), on the other hand, assesses the performance in a region-based manner. It is similar to ROC analysis except that in the x-axis there is the number of false positives (regions) per image. Many definitions, thus, exist for what constitutes a detected region. An example of such criterion is a 50% overlap between the detected region and the radiologist's annotation. The common practice in the literature is to assess CADs in terms of the Free-response Receiver Operating Characteristics (FROC) [16].

IV. Results and Discussion

We applied the method introduced in Section II to a data set consisting of 50 malignant cases and 100 normals. The assessment method is the average AUC (defined in section III), where the average is taken over images. The average AUC for multiscale template matching with scales 6% 8% and 10% is 0.866.

Figure 4 compares the AUC of the scores of template matching to that of using plain gray-level as a score. The first row shows the probability image(scores image). The second shows the regions detected at a certain threshold.

Figure 5 shows a comparison between template matching and two other in-house developed methods. Different criteria for defining what a detection is exist in the literature. A comparison between 4 of them and our proposed criterion is shown.

One problem with using FROC plots to evaluate pixel-based approaches like template matching is that the resulting curve is non-monotonic. This is caused by the fact that at a given threshold, there are different groups of pixels that merge to for a single marker; but as the threshold is increased, regions would grow and merge with each other until eventually all regions merge into a single big region that occupies the entire breast. This can be seen in Figure 3 by moving from right to left.

V. Conclusion and Future Work

In this paper we showed some of the relevant literature. We introduced our novel data set and its collection method. Template Matching with a spherical template has been tested and evaluated using both AUC and FROC as evaluation criteria.

Although Template Matching alone shows no superior results, when it's combined with the scores of other in-house developed methods in an ensemble or a Multiple Classifier System (MCS) the performance was boosted. The MCS we experimented was a simple one. The final score was just a weighted average of Template Matching's score and the two other methods.

For the future we intend to further investigate MCS. In addition, Deep Learning has shown good results with similar problems. We think applying Deep Learning techniques in the literature should yield better performance.

References

[1] American Cancer Society, *Breast Cancer Facts & Figures 2015-2016*. Atlanta: American Cancer Society, 2015. [Online]. Available: http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-046381.pdf

[2] A. S. Ibrahim, H. M. Khaled, N. N. Mikhail, H. Baraka, and H. Kamel, "Cancer Incidence in Egypt: Results of the National Population-Based Cancer Registry Program," *Journal of Cancer Epidemiology*, vol. 2014, pp. 1–18, 2014. [Online]. Available: http://www.hindawi.com/journals/jce/2014/437971/

[3] R. M. Kamal, N. M. Abdel Razek, M. A. Hassan, and M. A. Shaalan, "Missed breast carcinoma; why and how to avoid?" *J Egypt Natl Canc Inst*, vol. 19, no. 3, pp. 178–194, 2007.

[4] T. W. Freer and M. J. Ulissey, "Screening Mammography with Computer-Aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center," *Radiology*, vol. 220, no. 3, pp. 781–786, 2001. [Online]. Available: http://dx.doi.org/10.1148/radiol.2203001282

[5] D. Gur, J. H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. Hardesty, W. R. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," *J Natl Cancer Inst*, vol. 96, no. 3, pp. 185–190, 2004. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=14759985

[6] B. Zheng, M. Ganott, C. Britton, and C. Hakim, "Soft-Copy Mammographic Readings with Different Computer-assisted Detection Cuing Environments: Preliminary Findings1," *Radiology*, 2001. [Online]. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed{&}cmd=Retrieve{&}dopt=AbstractPlus{&}list{_}uids=17856925475310200975related:j-h8GByL0PcJ$\delimiter"026E30F$npapers2://publication/uuid/3D711C9C-BE1F-4858-8FB7-FEF3B5B8222C

[7] S. M. Lai, X. Li, and W. F. Bischof, "On techniques for detecting circumscribed masses in mammograms," *IEEE Transactions on Medical Imaging*, vol. 8, no. 4, pp. 377–386, 1989.

[8] S. L. Ng and W. F. Bischof, "Automated Detection and Classification of Breast-Tumors," *Computers and Biomedical Research*, vol. 25, no. 3, pp. 218–237, 1992. [Online]. Available: ⟨GotoISI⟩://A1992HX26400003

[9] G. D. Tourassi, R. Vargas-Voracek, D. M. Catarious, and C. E. Floyd, "Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information." *Medical physics*, vol. 30, no. 8, pp. 2123–2130, 2003.

[10] A. Oliver, J. Freixenet, E. R. E. Denton, R. Zwiggelaar, and C. Science, "Mammographic mass eigendetection," *Medical Image Understanding and Analysis*, pp. 71–75, 2006.

[11] A. Oliver, J. Freixenet, R. Mart, J. Pont, P. Elsa, E. R. E. Denton, and R. Zwiggelaar, "A Novel Breast Tissue Density Classification Methodology," *IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE*, vol. 12, no. 1, pp. 55–65, 2008.

[12] G. M. te Brake and N. Karssemeijer, "Single and multiscale detection of masses in digital mammograms." *IEEE transactions on medical imaging*, vol. 18, no. 7, pp. 628–639, 1999.

[13] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 1, pp. 62–66, 1979. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?{&}amp;arnumber=4310076

[14] S. Morrison and L. Linnett, "A Model Based Approach to Object Detection in Digital Mammography," in *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, 1999, pp. 182–186 vol.2.

[15] A. P. Bradley, "The Use Of The Area Under The ROC Curve In The Evaluation Of Machine Learning Algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[16] D. Chakraborty and K. Berbaum, "Observer studies involving detection and localization: Modeling, analysis, and validation," *Medical Physics*, vol. 31, no. 8, pp. 2313–2330, 2004.

[Online]. Available: http://link.aip.org/getpdf/servlet/GetPDFServlet?filetype=pdf{&}id=MPHYA6000031000008002313000001

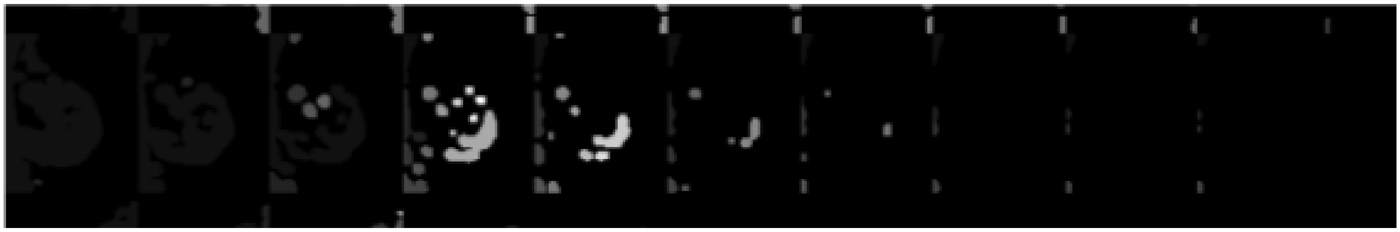


Fig. 3: The detected regions. By moving from right to left (increasing the threshold), different regions grow and merge to form a new bigger region.

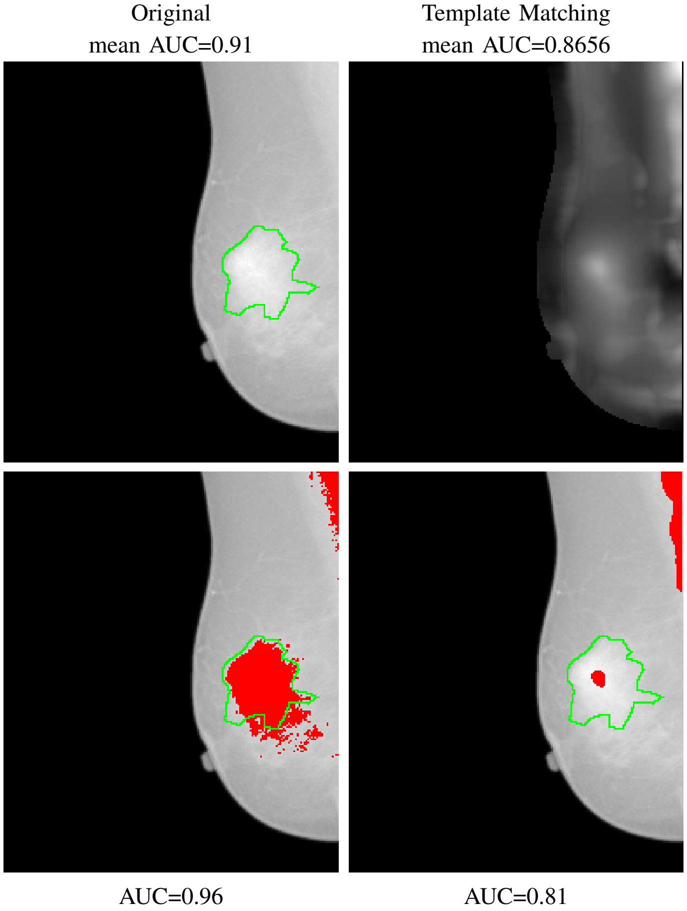


Fig. 4: Upper left is an image with radiologist's marking; the other image in the same row is the scores given by Template Matching method. The second row is the detected regions by setting a threshold for the score image. The AUC is indicated at the bottom of the row.

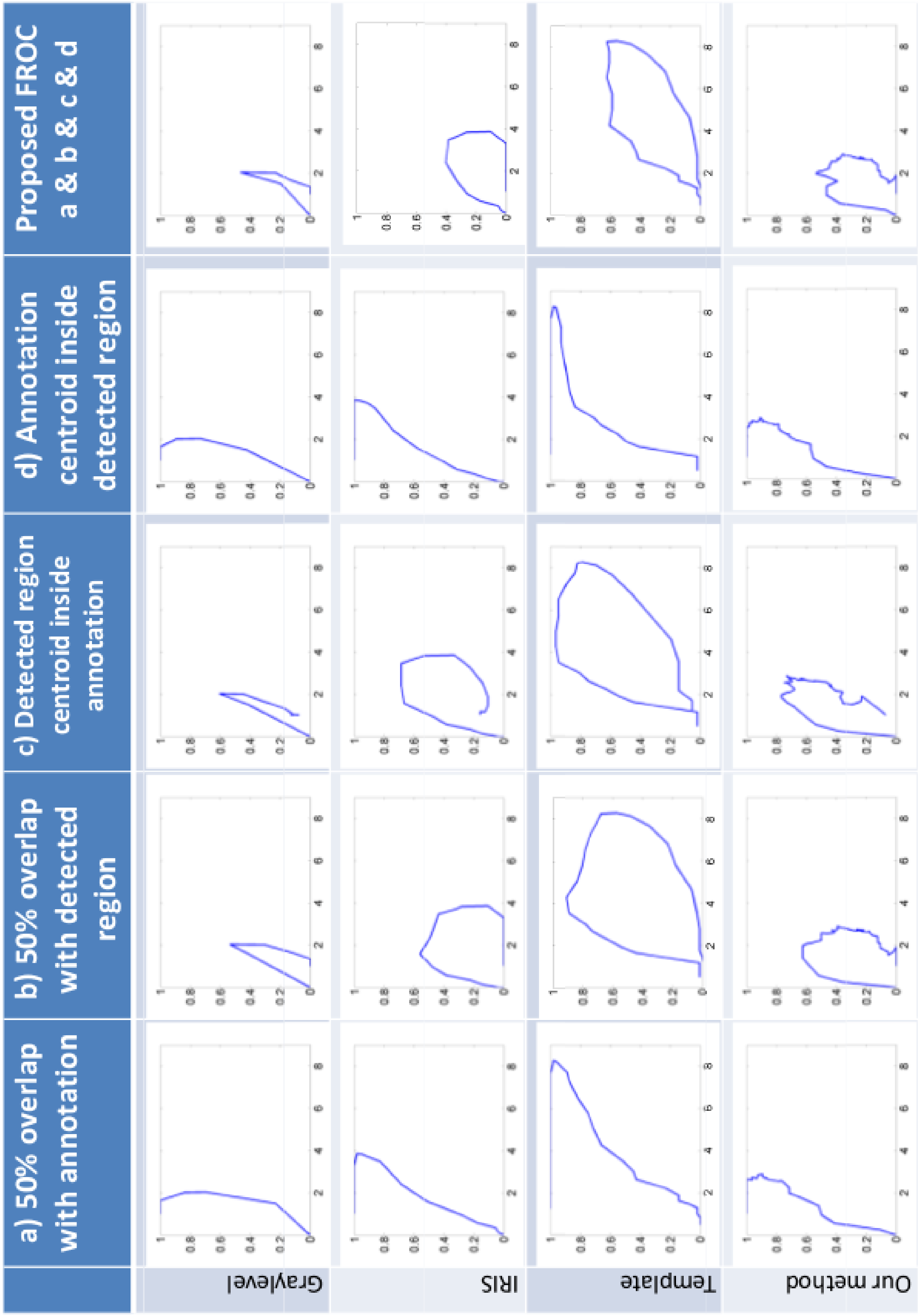


Fig. 5: Conventional FROC results using four criterion for the four detection methods.

An Anti-Pattern-based Runtime Business Process Compliance Monitoring Framework

Ahmed Barnawi
King Abdulaziz University
Jeddah, Saudi Arabia

Ahmed Awad
Cairo University
Cairo, Egypt

Amal Elgammal
Cairo University
Cairo, Egypt

Radwa Elshawi
Princess Nourah Bint Abdulrahman University
Riyadh, Saudi Arabia

Abduallah Almalaise
King Abdulaziz University
Jeddah, Saudi Arabia

Sherif Sakr
King Saud bin Abdulaziz University for Health Sciences
Riyadh, Saudi Arabia
University of New South Wales, Sydney, Australia

Abstract—**Today′s dynamically changing business and compliance environment demand enterprises to continuously ensure their compliance with various laws, regulations and standards. Several business studies have concluded that compliance management is one of the main challenges companies face nowadays. Runtime compliance monitoring is of utmost importance for compliance assurance as during the prior design-time compliance checking phase, only a subset of the imposed compliance requirements can be statically checked due to the absence of required variable instantiation and contextual information. Furthermore, the fact that a BP model has been statically checked for compliance during design-time does not guarantee that the corresponding running BP instances are usually compliant due to human and machine errors. In this paper, we present a *generic proactive runtime* BP compliance monitoring framework, *BP-MaaS*. The framework incorporates a wide range of expressive high-level graphical *compliance patterns* for the abstract specification of runtime constraints. Compliance monitoring is achieved using *anti-patterns*, a novel mechanism which is agnostic towards any underlying monitoring execution technology. As a proof-of-concept, complex event processing (CEP) technology is adopted as one of the possible realizations of the monitoring engine of the framework. An integrated tool-suite has been developed as an instantiation artifact of BP-MaaS, and the validation of the approach is undertaken in several directions, which includes internal validity and case study conducts considering two real-life case studies from the banking domain.**

Keywords—***Business Process Management; Business Process Monitoring; Business Process Compliance***

I. Introduction

The global regulatory environment has grown in complexity and scope since the financial crisis in 2008. This is causing significant problems for organizations in almost all industrial sectors, as the complexities of hard and soft regulations are little understood or appreciated [6]. For example, banking regulations such as anti-Money Laundering Directives are generally complex and far-reaching, with a raft of major banks found to be not in compliance in 2012. Standard Chartered Bank, London, for example, was fined a total of $459 million in 2012[1]. Worse still, HSBC Holdings Plc. had paid a record of $1.92 billion. These incidents were preceded by scandals and business failures such as Enron, and WorldCom back in 2001. Subsequently, much attention has been paid to compliance management from both the academic and the industrial communities.

Research on compliance management has focused on the checking and enforcement of compliance in one of the BP lifecycle phases; i.e. design-time verification (e.g., [4, 22]), runtime monitoring (e.g., [34, 57]) and offline monitoring (e.g., [1]). While each of these phases has its strengths and limitations, we consider *proactive* runtime compliance monitoring as a vital component. With *proactive*, we mean violations are detected as soon as they occur, and appropriate recovery action(s) are taken to mitigate/minimize their impacts. Only a subset of compliance rules can be checked during design-time due to the lack of necessary contextual information; e.g., time span constraints between tasks can not be checked at design-time. Besides, due to human and machine errors, a violation might still occur in a running instance resulting from a statically compliant BP model. Offline monitoring, on the other hand, is more beneficial for statistical analysis and diagnostic purposes, as it is usually performed on a large number of completed executions, following a retrospective (after-the-fact) approach.

Driven by this emergent *business need*, runtime compliance monitoring has recently been an active research topic (e.g., [38, 40, 64]); however despite the efforts, important aspects of compliance monitoring have been overlooked. For instance, the technology used to enable monitoring, the nature of the process execution environment, the type of events generated from these environments and their granularity have not been well discussed in literature. Moreover, such approaches are introduced as a component of a larger system that is usually tightly interwoven with pre-existing components.

In this paper, we described the design and the implementation details of an end-to-end *runtime* business process compliance monitoring as a service framework, *BP-MaaS*, where we aim at providing an independent and comprehensive platform as a compliance monitoring service [2]. In particular, we summarize the main contributions and strengths of our

[1] http://www.accuity.com/industry-updates/free-resources/trends-in-aml-compliance-infographic/

framework as follows:

- The framework supports a rich and wide set of compliance patterns that represent the abstract specification of monitoring requirements and cover the four structural facets of BPs; i.e. control-flow, data, employed resources and timing constraints.
- The monitoring engine of our framework is based on the concept of *anti-patterns*, a novel compliance evaluation mechanism which is agnostic towards the used technology for implementing the compliance monitoring engine.
- In order to ease the task of process designers, our framework is equipped with a user-friendly modeling environment that relies on using graphical notations for modeling the compliance rules.
- The framework is equipped with a Compliance Management Knowledge base (CMKB) that incorporates and integrates a set of ontologies to capture and homogenize the different perspectives of the compliance and business sphere.
- We present the details of a proof-of-concept implementation as an instantiation artifact for the monitoring engine of our framework, which adopts complex event processing (CEP) technology. In addition, our modeling environment has been equipped with a plugin that automatically maps the modeled compliance rules into the concrete utilized queries/scripts for executing the runtime monitoring process.
- We demonstrate the applicability and utility of the proposed framework by employing two real-world case studies that deal with business processes of companies operating in the banking domain. The findings have shown that the proposed framework supports a major subset of real-life compliance requirements addressing the heavily regulated banking domain.

In the following, Section II provides the necessary background for the paper. Section III introduces two case studies which are used for serving the purpose of our examples in the rest of the paper. In addition, these case studies are used for evaluating the applicability of the proposed approach that validates its usefulness and utility, and highlight its limitations. Section IV gives an overview of the proposed framework. Section V details the anti-patterns that back the monitoring engine of our framework. The details of the PoC implementation of our framework is presented in Section VI. Case studies-based evaluation of the proposed framework is presented in Section . Related work is highlighted in Section VIII, with a comparative analysis that aligns and appraises our approach against prominent related work efforts. Finally, conclusions and future work directions are drawn in Section IX.

II. Background

This section introduces the main concepts and techniques that form the groundwork for our approach. In addition, we introduce a case study that forms the basis of our discussion and examples in the rest of the paper.

A. Reference Life Cycle Models

As the objective of this work is to enable runtime monitoring of process compliance, we depend on events generated by the execution environment. Events represent the evolution of the running process instances. We assume that these events reflect a change in state of either the whole process instance or one of its task instances. The set of states and transitions among them is assumed to be predefined by a process/task lifecycle. This is important as the events generated by the execution environment will form the input of raw events to the monitoring component upon which violation will be detected. We build on top of the work of Russell et al. [50] and Lerner et al. [30] where we combine both works and project the relevant parts for compliance monitoring.

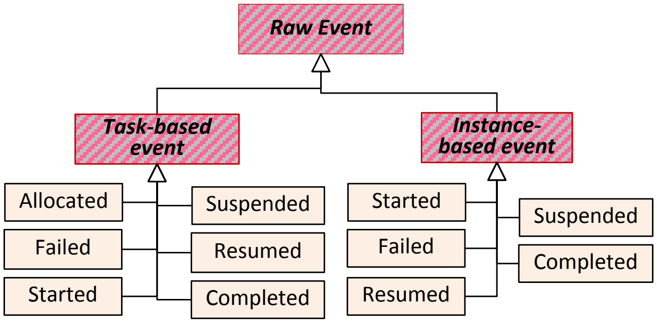


Fig. 1: Classification of raw events

Figure 1 summarizes the different types of events that are expected to be generated by the execution environment. Definition 2.1 formalizes what a *raw* event means in the context of this paper.

Definition 2.1 (Raw Event): Let PM be the set of all process models, PI be the set of all process instances and TI be the set of all task instances, and R be the set of all resources. A raw event is a tuple (*type, task_instance, process_instance, timestamp, data, resource*), where:

- type $\in \{started, failed, allocated, resumed, suspended, completed\}$ to indicate the actual type of the event,
- $task\_instance \in TI \cup \{\perp\}$ is a reference to the task instance for which the event occurred. When this property is not applicable, e.g. this is a process level event, the property has the value $\perp$,
- $process\_instance \in PI$ is a reference to the process instance within which the event occurred,
- $timestamp \in \mathbb{N}$ indicates the time stamp where the event occurred.
- $data$ is the data payload of the event. This basically holds states of data elements processed within a case but can be extended by execution environment related data. When this property is not applicable or irrelevant in context, the property has the value $\perp$.
- $resource \in R \cup \{\perp\}$ is a reference to the human resource performing a task. When this property is not applicable, e.g., this is an automated step, this property has the value $\perp$.

The set RE defines the different raw events that can be received on the stream. Instead of representing the occurrence of start event of task A within a certain process instance i as $e = (started, A, i, t, d, r)$ where t is the event time stamp, d is the data payload and r is the human resource, we write it as $e = started(A, i, t, d, r)$. Also, we might have a shorthand as $e = started(A, i)$ when the other pieces of information are not relevant.

B. Compliance Patterns

In general, pattern-based modeling of compliance rules is well accepted in the community and several studies have provided a comprehensive set of patterns that cover the different aspects as control flow, data flow, resource allocation and timing as summarized by Ly et al. [32]. In this work, we build on top of those patterns. In particular, Figure 2 summarizes the set of compliance patterns which are supported by our framework. We use *pattern* and *rule* interchangeably where a rule is an instantiation of a pattern.

In practice, any pattern can optionally be limited to a scope in which the rule is required to hold. The scope represents a time window that is bounded by case or task instance-related events. In principle, the default scope is the whole process instance execution. Also, the pattern can be refined by a condition where the rule is required to hold only when this condition is true. The condition may refer to process execution data that are reflected in the event data payload. With each pattern, two actions are defined. The *Violation Action* describes the action taken by the monitoring component when the violation occurs whereas the *Prediction Action* describes the action taken when there is a possibility of violation. The nature of the action depends on how the monitoring component is integrated with the execution environment. For instance, the simplest action that can be taken is to alert administrators.

Definition 2.2 (Atomic Compliance Rule): Let PM be the set of all process models to be monitored. A compliance rule is a tuple $(pattern, model, antecedent, consequent,$ $condition, scope\ start, scope\ end, multiplicity, WA, timespan,$ $alerttimespan, isBefore, violationaction, predictiveaction)$ where:

- $pattern \in \{Exists, Absence, Sequence, Next, Precedes,$ $One\ to\ one\ precedes, Response, One\ to\ one\ response,$ $SoD, BoD, Performed\ by\ role, Performed\ by\ resource\}$ defines the pattern from which the rule is instantiated,
- $model \in PM$ is a reference to the process model against which the rule has to be monitored,
- $antecedent \in \{ex, not(ex)|ex \in RE\}$ where $not(ex)$ means that event ex has not been observed,
- $consequent \in \{ex, not(ex)|ex \in RE\}$ where $not(ex)$ means that event ex has not been observed,
- condition is the data condition that is to be examined at the occurrence of the rule's $antecedent$
- $scope\ start \in RE$ defines the delimiting start event of the rule's scope,
- $scope\ end \in RE$ defines the delimiting end event of the rule's scope,
- $multiplicity$ is a constraint on the number of occurrences of the rule's $antecedent$,
- $WA \subset RE$ is the set of events that must not occur between the $antecedent$ and $consequent$,
- $time\ span$ is the time window in/out of which the $consequent$ event must be observed,
- $alert\ time\ span$ is the time window after which there is a possibility of violation if the $consequent$ was not observed,
- $isBefore \in \{0, 1\}$ is a Boolean value indicating wether the $consequent$ event must be observed *before* or *after* the end of the $time\ span$
- $violation\ action \in \{alert, suspend\}$ defines the action to be taken upon the occurrence of a violation,
- $predictive\ action \in \{alert, suspend\}$ defines the action to be taken when there is a possibility of a violation.

When any property does not apply to a rule pattern, it is represented as $\perp$. We define CR as the set of all compliance rules registered with the monitoring component.

As per Definition 2.2, there might be a *time span* window that puts further constraints on the observation of the $consequent$ event with respect to the $antecedent$ event. This is also further controlled by the *isBfore* property. So, if *isBore = 1*, the pattern requires that the $consequent$ event to be observed *before* time span elapses otherwise there is a violation. Whereas, if *isBefore = 0*, then $consequent$ has to be observed *after* the time span elapses.

Composite patterns are used to logically connect other patterns by Boolean operators AND, OR, NOT, etc. This is used to define complex rules that can not be expressed merely by atomic patterns, which is especially helpful when sub-ideal level of compliance is also needed [32].

C. Complex Event Processing (CEP)

In traditional systems, data are static in the system while the queries used are changing. For example in traditional database systems, the data are stored in tables and users can write different queries that access those tables to process data and get the results. However, when using complex event processing (CEP) technology, the roles of data and queries are reversed; where the queries will be static and data or events will be dynamic based on the input event streams from different sources. These queries will be checked against incoming streams of events to verify that queries are answered correctly.

In the context of business process compliance, monitoring queries (rules) are commonly represented in the form of event-condition-action (ECA) rules [23]. Thus, the first trigger to evaluate a query is to find the matching input event on the stream. In general, there are two types of events, *raw* (*low-level*) events and *business-level* (*composite*) events [31]. Composite events might result from the evaluation of one or more query whereas raw events are generated from the different execution environment. In the context of process compliance monitoring, *raw* events are generated based on a change of the state of running process instances (cases) and their activities. In practice, an event stream contains a set of associated events, that are chronologically ordered. These events are generated from different sources. This stream of events could be homogeneous in which all events must be from the same type, or heterogeneous in which all events may be of different types [23].

III. Case Studies

In this section, we introduce two case studies addressing the banking domain. The *first* case study has been conducted in the Governance, Risk and Compliance Technology Centre

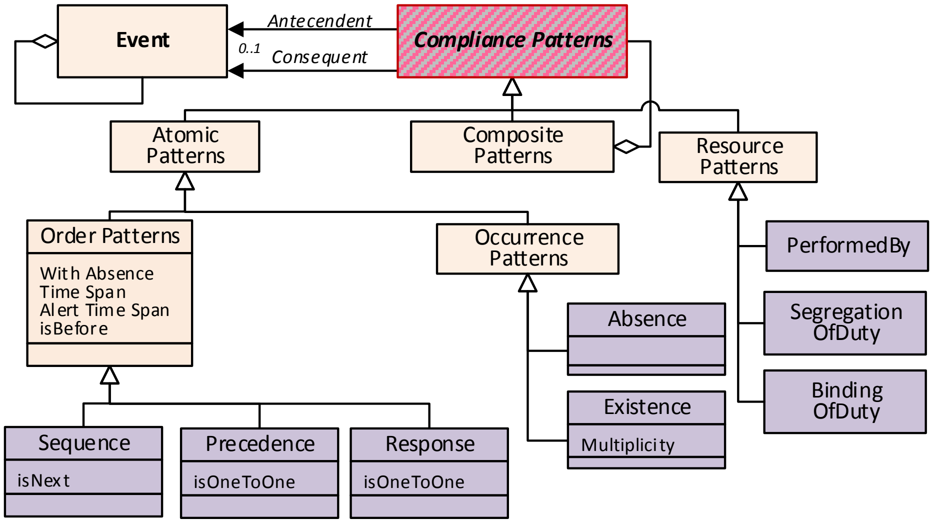


Fig. 2: Compliance Patterns

(GRCTC)[2] focusing on Anti-money Laundering (AML). The *second* case has been conducted within the scope of the EU-funded research project COMPAS[3], targeting the loan processing business scenario. This case study was performed by Thales Services, France[4], as one of COMPAS industrial partners; a company heavily operating in the banking domain. Taking into account the demands for strong regulation compliance schemes, such as Sarbanes-Oxley (SOX), BASEL-III, ISO 27000 and sometimes contradictory needs of the different stakeholders, such business environments raise several interesting compliance requirements. Anti-money laundering is a pressing concern to any organization operating in the financial industry, as it is tightly adjunct to terrorism and proliferation financing. Despite the fact that it is not possible to precisely quantify the amount of money laundered every year, in [48], it has been shown that billions of US dollars certainly are. On-going work in GRCTC in collaboration with respectable Irish financial organizations focuses on developing an end-to-end AML business process encoded in the BPMN v2.0 standard, which is established based on best practices and the Financial Action Task Force (FATF) 40 recommendations [55]. The U.S. Patriot act of 2001 [44] was considered as the main source of compliance requirements targeting anti-money laundering, which constitutes a large number of compliance requirements structured into twelve sections. The interpretations of embedded requirements and encoding them in the Semantics of Business Vocabulary and Business Rule (SBVR) standard [45] as a structured natural language is ongoing work in GRCTC [20]. Since the interpretation of compliance sources mandatory requires legal expertise [58, 60], GRCTC has hired three legal experts, who work very closely with compliance experts to interpret the AML directives and encode them in SBVR. We use the second case study (Loan Approval scenario) as a running scenario to exemplify the next discussion. The conducts and the overall findings of both case studies are discussed later in Section VII.

A simplified model of the loan origination and approval process is depicted in Figure 3 using BPMN (Business Process Model and Notation). The process flow can be described as follows: Once a customer loan request is received, the credit broker checks customer′s banking privileges status. If privileges are not suspended, the credit broker accesses the customer information and checks if all loan conditions are satisfied. Next, a loan threshold is calculated, and if the threshold amount is less than 1M Euros, the post-processing clerk checks the credit worthiness of the customer by outsourcing to a credit bureau service. Next, the post-processing clerk initializes the loan form and approves the loan. If the threshold amount is greater than 1M Euros, the supervisor is responsible for performing the same activities instead of the post-processing clerk. Next, the manager evaluates the loan risk, after which she normally signs the loan form and sends the form to the customer to sign. A legal waiting time of 7 days is provided to the customer to send back the signed contract. If a timeout occurs, which means that seven days have passed and the Customer has not sent the signed contract, the relevant loan approval application is closed by the system and the process terminates.

Table I shows an excerpt of the compliance requirements imposed on this loan approval scenario. This table is populated after applying the refined methodology described in [58, 60]. The first and second columns of the table allocate a unique reference and an organization-specific interpretation of the requirement, respectively. The third column represents the pattern-based representation of interpreted compliance requirements.

IV. BP-MaaS: Framework Overview

Figure 4 provides an overview of the generic business process compliance management BP-MaaS framework. In principle, the framework has two major components: Compliance management component (right hand side of Figure 4) and Business process management component (left hand side of Figure 4). The framework is generic, however, in the

[2] GRCTC: http://www.grctc.com/

[3] COMPAS: http://www.compas-ict.eu

[4] Thales Group: https://www.thalesgroup.com/

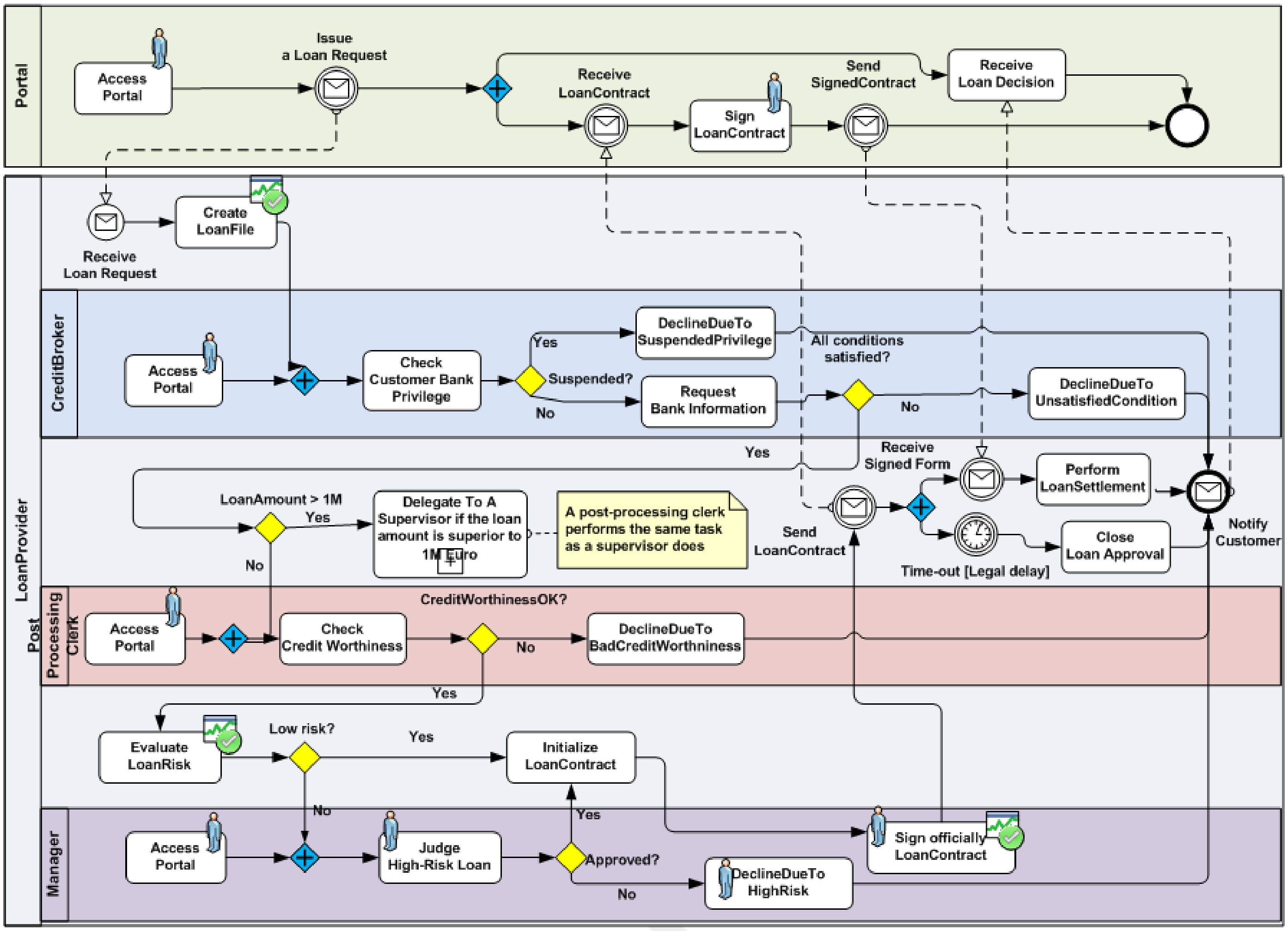


Fig. 3: A simplified BPMN model of the loan origination and approval process

next discussion we are considering Business Process Modelling and Notation standard [25] as a possible instantiation of the proposed framework. As shown in the figure, the BP-MaaS exhibits two basic abstract roles: *business expert* and *compliance expert*. The *business expert* is responsible for the definition, design, development and management of service-enabled business processes, while taking compliance requirements into consideration. The *compliance expert* is responsible for the refinement, interpretation and specification of compliance requirements emerging from various internal and external compliance sources in close collaboration with the business expert.

On the top of the monitoring components is the Compliance Management Knowledge base (CMKB) that incorporates and integrates a set of ontologies to capture the different perspectives of the compliance and business spheres [21]. In practice, compliance requirements usually originate from various sources, including laws and regulations, standards, public and internal policies, partner agreements, etc., and organizations are continuously required to comply with increasing number of diverse and evolving laws and regulations. Furthermore, compliance and business concepts may be treated differently by different stakeholders with different backgrounds. This ambiguity results in inconsistency, which makes it infeasible to share and re-use business and compliance specifics. All these problems make it infeasible for automated compliance monitoring and analysis. In principle, the need to manage regulatory and compliance data, especially in heavily-regulated domains, exceeds the abilities of current information systems. The majority of compliance management approaches in the literature (cf. Section VIII) assumes this ontological alignment, due to the complexity of this problem. Therefore, our framework is equipped with a uniform conceptualization of the process and compliance space which provides various advantages including:

- Enabling the sharing and re-use of compliance and business knowledge
- Eliminating any ambiguity that may result in unforeseen inconsistencies
- Significantly facilitating the communication between stakeholders with different backgrounds, e.g., compliance and business experts
- Ensuring the ontological alignment between business and compliance concepts

TABLE I: Compliance Requirements

Compliance Requirements	Control	Pattern Representation
1) Access to sensitive data is restricted	Only Credit Broker, Post-processing Clerk and Supervisor roles can access the Credit Bureau service and the Customer Information File service.	R1.1: PerformedByRole(Verify Banking Privilege,Role=Credit Broker) R1.2: PerformedByRole(Acquire Bank Information,Role=Credit Broker) R1.3: PerformedByRole(Check Credit Worthiness,Role$\in$ {Post Processing Clerk ,Supervisor}) R1.4: PerformedByRole(Prepare Package Price,Role $\in$ {Post Processing Clerk , Supervisor})
2) Only credit worthy customers receive a loan.	The Credit Broker checks the Customer Bank Privilege and rejects the loan request if the Credit Bureau Service indicates that the customer's banking privileges have been suspended.	R2.1: PerformedByRole(Verify Banking Privilege,Role = CreditBroker) R2.2: Response(Verify Banking Privilege, Decline By Suspeneded Banking Privilege, condition=(Suspeneded = Yes))
3) Duties in Loan Origination are segregated.	The Credit Broker checks the banking privileges and the Post-processing Clerk or the Clerk Supervisor checks the credit worthiness.	R3.1: Response(VerifyBankingPrivilege,CheckCreditWorthiness) R3.2: Absence(CheckCreditWorthiness,scop_end = VerifyBankingPrivilege) R3.3:SoD(VerifyBankingPrivilege,CheckCreditWorthiness)
4) High loan request have to be processed by supervisors.	If the loan requests credit is below 1 million EURO, the Post-processing Clerk of Credit Operations checks the credit worthiness, if it is higher than 1 million EURO the Clerk Supervisor checks the credit worthiness of the customer.	R4.1: PerformedByRole(Check Credit Worthiness,Role = Post Processing Clerk, condition = Threshold $< 1M$ Euro) R4.2: PerformedByRole(Check Credit Worthiness,Role = Supervisor, condition = Threshold $\geq 1M$ Euro)
5) Duties in Loan Origination are adequately segregated.	As a final control, the branch office Manager has to check high-risk loan requests whether it is profitable and risks are acceptable and makes the final approval (or denial) of the request. Only the Office Manager is able to perform the final approval.	R5.1: PerformedByRole(Evaluate Loan Risk,Role= Manager) R5.2: Exists(Evaluate Loan Risk) R5.3: PerformedByRole(Sign Loan Contract,Role= Manager) R5.4: PerformedByRole(Decline By High Risk,Role= Manager) R5.5: Response(Evaluate Load Risk, Decline By High Risk,condition=(LowRisk = No)) R5.6:Response(Evaluate Load Risk, Sign Loan Contract,condition=(LowRisk = Yes) R5.7: Exists (Sign Loan Contract) XOR Exists(Decline By High Risk)
6) Customers receive a certain period of time for reflection.	The Credit Broker can start a loan approved by the customer, only if five work days or more have elapsed since the loan approval form was sent.	R6.1: Response(Send Loan Contract,Perform Loan Settlement , time span = after 5 Working days) R6.2: PerformedBy(Perform Loan Settlement, Role= Credit Broker)
7) Customers personal data is handled confidentially.	The customer receives an email notification when his data is collected from the Credit Bureau service.	R7.1:Response(Acquire Bank Information,Send Notification To Customer,WA=All other) R7.2: Response(Check Credit Worithness,Send Notification To Customer,WA=All other)

- Improving the level of automation provided by the framework

To achieve these goals, three main ontologies are employed: (i) *BP ontology*: an ontology that captures the semantics of the adopted BP language, e.g. BPMN, BPEL, (ii) *Domain Ontology*: an ontology that represents the concepts and relationships that exist in the domain of interest, e.g., medical, transport, aerospace, etc. and (iii) *Regulatory Ontology*: an ontology that formalizes the requirements, controls and rules of compliance imperatives. Ontologies in the CMKB may be represented formally in the Ontology Web Language (OWL2.0) [63]. We build upon the BPMNO ontology [18] as the Business Process ontology. BPMNO provides a rich ontological representation of the BPMN v2.0 standard, which we consider as one of the possible instantiation of the framework. If another business process language is adopted, e.g., Business Process Execution Language (BPEL) [43], an ontology should be incorporated to capture its semantics. Moreover, we utilize the Financial Industry Business Ontology (FIBO) [15, 12] as the domain ontology since the case study we use in this paper (and introduced in Section III) concerns with the financial/banking domain. FIBO is an adopted OMG standard, a collaborative initiative led by industry members of the Enterprise Data Management Council (EDMC) in collaboration with the Object Management Group (OMG). Similarly, if another domain is considered, relevant domain ontology should be incorporated.

Business process management component (left hand side of Figure 4) starts by the business expert defining new BPMN model or re-use/update an existing one. In a green scenario, if the BPMN process is built from scratch, concepts and relations from FIBO can be directly used to guide its design and development. However, if BP models already exist (which is the common case), concepts from FIBO can be used to provide semantic annotation to existing BPMN models (the semantic labelling link between 'Domain Ontology' and 'BP Ontology' as shown in Figure 4)), and then various reasoning mechanisms can be applied to ensure the correctness of these annotations as proposed in [18]. Examples of concepts in FIBO are: 'Agent', 'Person', 'National id', 'real estate', 'agreement', 'contract', 'ownership', 'Asset', etc.; examples of relations are: 'manages', 'provides', 'represents', 'is issued by', 'is appointed by', etc. Examples of concepts in BPMNO are: 'Activity', 'Event types', 'Task', 'Gateway', etc.

Regulatory ontology is under development that addresses the regulatory domain by capturing concepts and relationships necessary to represent compliance patterns and compliance rules. Key concepts in this ontology are 'Input Data', 'Events', 'Condition', 'Action', 'Boolean Operator', etc. As discussed in Section II, the framework makes use of *compliance patterns* to represent compliance requirements mainly to provide an abstraction layer, so that experts do not have to go into the low-level and intricate details of the underlying formal/logical language. Therefore, the Regulatory ontology also maintains concepts such as 'Compliance Patterns', 'Occurrence Patterns', 'Timed Patterns', 'Operand', 'Label', etc. As shown in Figure 4, Compliance Management Ontology, BP ontology (e.g., BPMNO), Domain Ontology (e.g., FIBO) and Regulatory ontology represent the Terminological part (TBox) of the CMKB. Instances from these ontologies populate the Compliance Requirements Repository (CRR), and Business Process Repository (BPR), representing the Assertional part[5] (ABox) of the knowledge base.

The business process and compliance management activ-

[5] The terms ABox and TBox are used to describe two different types of statements in ontologies. TBox statements describe a conceptualization, a set of concepts and properties for these concepts. ABox are TBox-compliant statements about individuals belonging to those concepts.

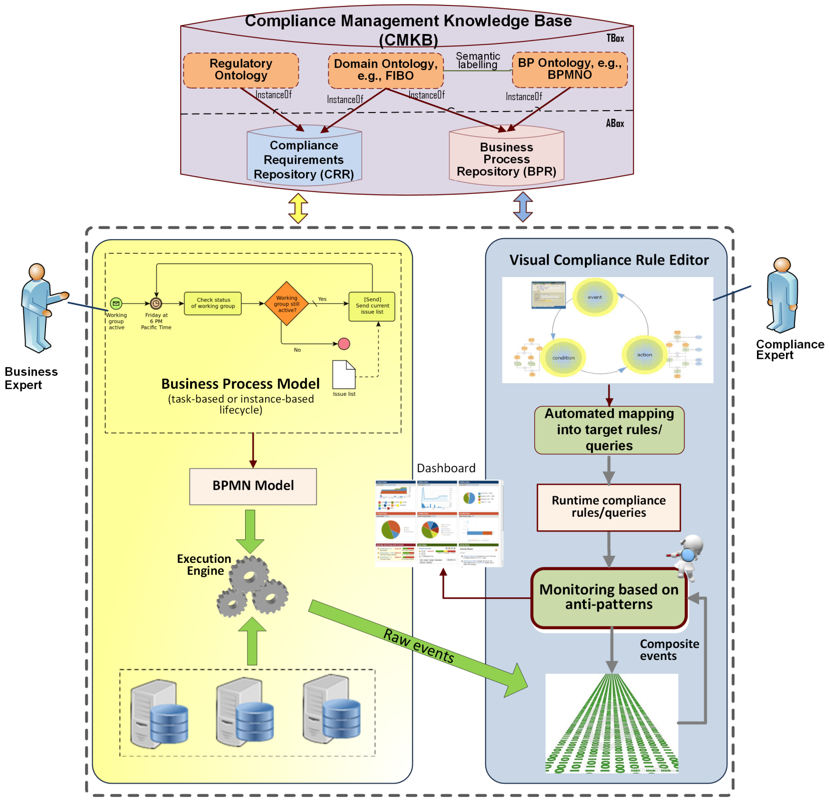

Compliance Management Knowledge Base (CMKB)
TBox
Regulatory Ontology
Domain Ontology, e.g., FIBO
Semantic labelling
BP Ontology, e.g., BPMNO
InstanceOf
InstanceOf
InstanceOf
InstanceOf
ABox
Compliance Requirements Repository (CRR)
Business Process Repository (BPR)
Business Expert
Working group active
Friday at 6 PM Pacific Time
Check status of working group
Working group still active?
Yes
[Send] Send current issue list
No
Issue list
Business Process Model
(task-based or instance-based lifecycle)
BPMN Model
Execution Engine
Raw events
Dashboard
Visual Compliance Rule Editor
event
condition
action
Compliance Expert
Automated mapping into target rules/ queries
Runtime compliance rules/queries
Monitoring based on anti-patterns
Composite events

Fig. 4: Framework Overview

ities start independently of each other, but they should be aligned with each other. This agrees with the necessity of the separation of compliance and business process management practices [51], [22], mainly because:(i) they have different objectives: ownership and governance perspectives , (ii) different lifecycles, (iii) different nature: compliance requirements are normative/descriptive by nature (describing what should be done), therefore, declarative languages are more suited to capture them, while business process specifications are prescriptive (describing how business activities should take place), therefore, procedural languages are the best to represent them, and (iv) the two specifications may conflict/contradict with each other. Therefore, they should be separated, however, their interrelationships should be carefully studied and maintained.

As mentioned above, Business process management practices commences with the business expert defining and modelling business process requirements using BPMN. The design process supports both task-based and instance-based lifecycles (cf. Section II). However, different execution environments could have slightly different lifecycle models, where adapters could be utilized to fill in this gap. In this case, adapters have to be implemented to correlate or supplement missing events that are expected by the monitoring component.

This step is highly iterative, such that the BPMN model follows multiple iterations of design and refinements to faithfully represent business logic and requirements. The outcome of this step is a BPMN model capturing the control and data flow of the business logic. The loan approval business scenario introduced in Section III (Figure 3) is an example of a BPMN model in the banking domain. The framework currently relies on the *Oryx* system [16] as its business process modeling environment that supports the BPMN 2.0 exchange format. However, the framework could be easily adapted to support other formats as well.

The BPMN model is then deployed on a business process (BP) engine in a possibly distributed environment. During execution, the BP engine emits different types of defined events reflecting the evolution of the execution of the BP model. These raw events are sent synchronously to the compliance management component (right-hand side of Figure 4), where business process monitoring comes into play. As shown in Figure 4, the interconnection between the business process management component and the compliance management component is through the emitted raw events.

From the compliance management side (right-hand side of Figure 4), the compliance expert starts by the specification of applicable compliance requirements using compliance patterns (as described in Section II) and using concepts from the CMKB. Table I in Section III shows an excerpt of some compliance requirements imposed on the loan approval model shown in Figure 3. To further ease the job of the compliance expert, we have developed a graphical compliance rule editor (on top of *Oryx*) that enables the compliance expert to intuitively build pattern-based expressions in a drag-and-drop fashion (more details about the implementation are presented in Section VI). Graphical pattern-based expression are then automatically mapped into the target formal/query language and then feed the monitoring component. As one possible realization of the framework, we have defined the mapping scheme from graphical pattern-based expressions into Event Processing Language (EPL) queries following a complex event processing approach (details are given next in Section VI). The framework currently adopts its customized format for storing and exchanging the compliance rules. Adopting a standard rule exchange format, e.g. RuleML [13], is left as a future work.

Raw events as data streams sent from the BP engine are then interrelated to generate composite events that are compliant with the lifecycle model on the stream (based on defined event-patterns of interest). The monitoring component then evaluates these rules/queries against the composite events and appropriate actions are taken in case a runtime violation is detected by the monitoring component. More specifically, when a new event is received on the stream, it automatically gets directed to the set of rules that have subscribed for that event for processing and detecting or predicting any possible non-compliance instances. As a result, the rule/query might generate another event on the stream to signal the compliance status of the instance that has generated the event. The monitoring is achieved by means of *anti-patterns*, a novel compliance monitoring evaluation approach we introduce in this paper (details are presented next in section V). Finally, compliance results are presented to the compliance/business experts on dashboards in detailed and aggregated manner with the support of various charts/indicators. Business and compliance experts scrutinize the results presented on the dashboards, which may initiate multiple iterations of the business process or compliance lifecycles for improvements.

Figure 4 highlights, by dashed lines, the components which represent the main focus of this paper. More specifically, the continuous runtime monitoring of compliance requirements on the basis for anti-patterns, which will be discussed in more detail next in Section V. The proof-of-concept of the proposed approach as a runtime monitoring tool-suite is presented in Section VI. This is followed by a discussion of validation and evaluation efforts in Section VII

V. Monitoring Compliance

In this section, we discuss the mechanics of the monitoring process in our framework. In principle, we follow the notion of *anti-patterns*, however, we are mainly focusing on the implementation of this notion for achieving the compliance requirements at the runtime phase instead of the design phase which has been considered by previous related work [4, 5]. In particular, we directly look for sequences of events that indicate that a violation has occurred or likely to occur. To achieve this, a set of anti-patterns are inferred for each compliance pattern. One advantage of looking for anti patterns is that the root cause of the violation is given without incurring additional costs [32]. Whenever a query finds a match, it generates a *Rule Violation Event*, see Definition 5.1, on the stream. This event can then be caught by another query to take an action against the violation.

Definition 5.1 (Rule Violation Event): A rule violation event is a tuple $(rule, case, isViolation, RC, timestamp)$ where:

- $rule \in CR$ is a reference to the rule for which a violation has been detected or predicted,
- $case \in PI$ is a reference to the process instance for which a violation has been detected or predicted,
- $isViolation \in \{true, false\}$ is a flag to indicate whether this an actual or a prediction of a violation,
- RC is the sequence of events that is the root cause of the violation,
- $timestamp \in \mathbb{N}$ is the time stamp at which the event occurred.

We define RVE as the set of rule violation events generated.

As per Definition 5.1, a rule violation event can be thrown even in the case where there is a possible violation. This is distinguished from the actual violation by the flag $isViolation$. Some of the patterns require that process tasks have to take place within a specified time span. Therefore, we assume that when a time span expires, a timer event is thrown to indicate that the time span has expired.

Definition 5.2 (Timer Event): A timer event is a tuple $(rule, case, time\ span, time\ stamp)$ where:

- $rule \in CR$ is a reference to the rule for which a violation has been detected or predicted,
- $case \in PI$ is a reference to the process instance for which a violation has been detected or predicted,
- $time\ span$ is the time span defined in $rule$,
- $time\ stamp \in \mathbb{N}$ is the time stamp at which the event occurred.

We use the notation $timer-event(r, c, time\ span)$ to refer to the timer event generated for rule r, case c and for $time\ span$. We define TE as the set of timer events generated.

Definition 5.3 (Event Stream): Event stream σ is a sequence of events on the form $e_1, e_2, \ldots, e_k$ where $e_i \in RE \cup RVE \cup TE, 1 \leq i \leq k$

As per Definition 5.3, the event stream is heterogeneous as it could hold raw events from the process execution, rule evaluation events and timer events. Moreover, raw events can belong to different process instances. We use Definition 5.4 to retrieve the history of execution of a certain process instance. It is also possible to retrieve a specific scope of the case history. Case history is used by the anti pattern queries to look for violations.

Definition 5.4 (Case History): Case history $\sigma_i(start, end)$, where $i \in PI$, $start, end \in RE$ is a projection on the event stream σ where only events related to instance i are projected. $start$ and end default to the case start and end events respectively. They are used to project a certain scope of the case history rather than the complete one. When an event e is a member of σ_i, we denote that as $e \in \sigma_i$.

Definition 5.5 (Count of Event Occurrence): Given that σ_i is a case history of process instance $i \in PI$ and $e \in RE$, we define a function $count(e, \sigma_i)$ that determines the number of occurrences of e in σ_i.

As shown in Figure 2, compliance rules can be either atomic or composite. In the following sections, we describe the anti-patterns, i.e., the violation scenarios for each atomic pattern.

A. Exists and Absence Anti Patterns

The *Exists* pattern requires that the *antecedent* event to occur a certain number of times within a certain scope in the process instance where a specific data condition holds. A violation to this rule occurs when the multiplicity constraint is no longer satisfied within the specified scope where the data condition holds with the antecedent event. Specifically, a violation takes place when either:

- A sequence of events where the scope start is observed and then the scope end is observed and in-between the antecedent occurs less than the minimum of the rules multiplicity.
- A sequence of events where the scope start is observed and then the antecedent occurs more than the maximum of the rules multiplicity.

These two possibilities are captured by Definition 5.6.

Definition 5.6 (Exists Violation): A violation to an Exists rule, $exists(pm, antecedent, \bot, condition, scope\_start, scope\_end, multiplicity, \emptyset, \bot, \bot, alert, alert)$, occurs in a process instance $i, i \in PI$ iff:

- $\exists r = \sigma_i(scope\_start, scope\_end) : multiplicity.min > count(antecedent, r)$. Or,
- $\exists r = \sigma_i(scope\_start, antecedent) : multiplicity.max < count(antecedent, r)$.

The *Absence* pattern requires that the *antecedent* event is never observed within a certain scope in the process instance where a specific data condition holds. Basically, the absence pattern can be seen as a special case of the *Exists* pattern where $multiplicity.min = multiplicity.max = 0$. Thus, the two possible violations scenarios in Definition 5.6 can be applied.

B. Response Anti Patterns

The *Response* pattern can be violated in the following cases:

- The rule's $antecedent$ is observed but the $consequent$ never occurs within the monitoring scope and time span, if defined, when $isBefore = true$,
- The rule's $antecedent$ is observed and then the $consequent$ is observed within the monitoring scope but before time span, if defined, where $isBefore = false$,
- The rule's $antecedent$ occurs and any of the forbidden events $with\ absence$ occurs before the rule's $consequent$,
- There is a (possible) violation if the (alert) time span elapses before the occurrence of the $consequent$ event, when $isBefore = true$.

Definition 5.7 (Response Violation): A violation of a Response rule, $response(pm, antecedent, consequent, condition, scope\ start, scope\ end, \bot, WA, time\ span, alert\ time\ span, isBefore, alert, alert)$, occurs in a process instance $i \in PI$ iff:

1) $\exists r = \sigma_i(scope\ start, scope\ end) : antecedent \in r \wedge consequent \notin r \wedge consequent.timestamp > antecedent.timestamp$. Or,
2) $\exists r = \sigma_i(scope\ start, timer\ event(response, i, time - span)) : antecedent \in r \wedge consequent \notin r \wedge consequent.timestamp > antecedent.timestamp$, where $isBefore = true$. Or,
3) $\exists r = \sigma_i(scope\ start, consequent) : antecedent \in r \wedge consequent.timestamp > antecedent.timestamp \wedge \nexists t = timer\ event(response, i, time\ span) : t.timestamp < consequent.timestamp$, where $isBefore = false$. Or,
4) $\exists r = \sigma_i(scope\_start, absent\ event) : absent\ event \in WA \wedge antecedent \in r \wedge consequent \notin r \wedge consequent.timestamp > antecedent.timestamp$.

Definition 5.7 formalizes the different violation scenarios for the *Response* pattern. The essence is that we can observe a sequence of events that starts with the $scope\ start$ event in which the $antecedent$ is observed but not the $consequent$ afterwords, the first, third and fourth case. This is where the time stamp is used for comparison. The difference between the three cases 1, 2 and 4 is in the closing event of the event sequence. In the first case, the $scope\ end$ is observed before having observed the $consequent$. In the third case, the $timer\ event$ which is generated when the time span elapses is observed. In the last case, one of the forbidden events $absent\ event \in WA$ is observed before $consequent$ and thus a violation has occurred. The second case however, detects that the $consequent$ has already occurred. However, there is a violation in this scenario if we fail to observe the timer event before the $consequent$. This is required only when the *Response* has its property $isBefore = false$ which means that the time span *must* elapse before observing the $consequent$, cf.Rule $R6.1$ in Table I.

The *One to One Response* is a special case of the *Response* pattern. In addition to the violation scenarios of the main pattern, *One to One Response* is violated if two occurrences of the $antecedent$ event are observed without observing the $consequent$ in between.

Definition 5.8 (One to One Response Violation): A violation of a One to One Response rule, $one\_to\_one\_response(pm, antecedent, consequent, condition, scope\ start, scope\ end, \bot, WA, time\ span, alert\ time\ span, isBefore, alert, alert)$, occurs in a process instance $i \in PI$ if a violation to its underlying Response rule occurs or $\exists r = \sigma_i(scope\ start, antecedent_1) : antecedent_2 \in r \wedge consequent \notin r \wedge consequent.timestamp > antecedent_1.timestamp$.

C. Sequence and Next Anti Patterns

The *Sequence* pattern requires that $antecedent$ occurs and is then followed by $consequent$ within a scope where the data condition holds. We can see that *Sequence* pattern can be a conjunction of the *Exists* and Response patterns. Thus, the violation of the *Sequence* pattern occurs if any of the violation scenarios of the *Exists* or the *Response* patterns occurs.

The *Next* pattern is a special case of the sequence pattern which in addition to the restrictions imposed by the *Sequence* pattern requires that no other events are observed in the process execution between $antecedent$ and $consequent$. In this regard, the anti patterns of the *Exists* and *Response* can be used. To enforce that no other events are allowed in-between, we can make use of the set WA that indicates the forbidden events. In this case, the set of forbidden events must refer to events of all other tasks in the process to ensure that nothing else happens between the $antecedent$ and the $consequent$.

D. Precedes Anti Patterns

The *Precedes* pattern can be violated in the following cases:

- The rule's $antecedent$ occurs but never the $consequent$ before it within the monitoring scope and time span, if defined,
- The rule's $antecedent$ occurs and the $consequent$ is observed before it within the monitoring scope but $before$ the time span elapses, in case the rule $isBfore = false$,
- The rule's $antecedent$ occurs and any of the forbidden events WA occurred before it and after the rule's $consequent$,

For *Precedes* rules, it is not possible to predict violations as it is a history-looking rule by nature.

Definition 5.9 (Precedes Violation): A violation of a Precedes rule, $precedes(pm, antecedent, consequent, condition, scope\ start, scope\ \ end, \perp, WA, time\ \ span, \perp, isBefore, alert, \perp)$, occurs in a process instance $i \in PI$ iff $\exists r = \sigma_i(scope\_start, antecedent)$:

- $scope\ end \notin r \wedge consequent \notin r$. Or,
- $scope\ end \notin r \wedge consequent \in r \wedge antecedent.timeStamp - consequent.timeStamp > time\ span$, if $isBefore = true$. Or,
- $scope\ end \notin r \wedge consequent \in r \wedge antecedent.timeStamp - consequent.timeStamp < time\ span$, if $isBefore = false$. Or,
- $\exists absent\ event \in WA : scope\ end \notin r \wedge consequent \in r \wedge absent\ event \in r \wedge absent\ event.timeStamp > consequent.timeStamp$.

Definition 5.9 formalizes the different violation scenarios for the *Precedes* pattern. The violation occurs when a projection on case history which starts with the $scope\ start$ event and ends with the $antecedent$ event and the scope is active, i.e., the $scope\ end$ was not observed before the $antecedent$ event $scope\ end \notin \sigma_i(scope\ start, antecedent)$ and:

- In the first case, the $consequent$ event was not observed as a member of that projection of case history.
- In the second case, the $consequent$ event was observed in the case history projection but the difference between the time stamp of the $antecedent$ and $consequent$ events is more than the prescribed time span in the rule when the rule's $isBefore$ property is set to $true$.
- The third case is similar to the case above with one difference is that the difference between the $antecedent$ and the $consequent$ events timestamps is less than the required time span. This is a violation only in case the rule has the $isBefore$ property set to $false$.
- In the last case, in addition to observing the $consequent$ event one of the forbidden events, $absent\ event \in WA$ was observed after $consequent$ was observed.

The *One to One Precedes* is a special case of the *Precedes* pattern. In addition to the violation scenarios of the main pattern, *One to One Precedes* is violated if two occurrences of the $antecedent$ event are observed without observing the $consequent$ in between.

Definition 5.10 (One to One Precedes Violation): A violation of a One to One Precedes rule, $one\_to_one\_precedes(pm, antecedent, consequent, condition, scope\ start, scope\ end, \perp, WA, time\ span, \perp, isBefore, alert, \perp)$, occurs in a process instance $i \in PI$ if a violation to its underlying Precedes rule occurs or $\exists r = \sigma_i(scope\ start, antecedent_1) \exists antecedent_2 \in r \nexists consequent \in r : consequent.timestamp > antecedent_2.timestamp$.

E. Separation/Bind of Duty Anti Patterns

The separation of duty pattern is violated whenever the consequent event has the same resource identifier as the antecedent. Note that the pattern does not care about the execution order of the tasks.

Definition 5.11 (Separation of Duty Violation): A violation of a Separation of Duty rule $Separation\_of\_Duty(pm, antecedent, consequent, \perp, scope\_start, scope\_end, \perp, WA, time\_span, \perp, alert, \perp$ occurs in a process instance $i \in PI$ iff:

- $\exists r = \sigma_i(scope\_start, antecedent) : consequent \in r \wedge consequent.resource = antecedent.resource$, or
- $\exists r = \sigma_i(scope\_start, consequent) : antecedent \in r \wedge consequent.resource = antecedent.resource$

As per Definition 5.11, there is a violation if the two events, $antecedent$ and $consequent$ related to the completion of the two separate tasks are observed in any order. In the first case, $consequent$ is observed and then $antecedent$, because we projected the case history until the occurrence of the $antecedent$ where $consequent$ was part of it. Whereas in the second case it was the other way around. In both cases violation occurs if they have the same resource performing the tasks. If $antecedent = completed(A, i)$ and $consequent = completed(B, i)$ then detecting the occurrences of the two events according to Definition 5.11 would signal a violation of a separation of duty rule between tasks A and B. However, there is also a possibility to predict a possible violation if two more separation of duty rules are used where in one of them $antecedent = started(A, i)$ and the $consequent$ remains the same. Whereas in the other rule the $antecedent$ remains the same and $consequent = started(B, i)$. With these rules if a sequence of events $completed(A, i), \ldots, started(B, i)$ is observed as defined by the second case in Definition 5.11, the violation is reported as a warning and there will be a chance to avoid the actual violation by reassigning task B to another resource. So, changing the type of events to detect in case of separation of duty rules helps to predict and avoid violations rather than just reporting its occurrence.

The *Bind of Duty pattern* is actually the negation of separation of duty. So, we can reuse the cases to detect the separation of duty with one change. That is to check that $antecedent.resource <> consequent.resource$.

VI. Implementation

Figure 5 presents the architecture of a proof-of-concept[6] implementation of the *BP-MaaS* framework for compliance monitoring of business processes[7]. In principle, the implementation of our framework consists of three main components: *Compliance rule modeler*, *compliance monitoring engine*, *monitoring dashboard*. Each of these components is described in the following subsections, respectively.

A. Compliance Rules Editor

The compliance rules editor provides the end users with a user-friendly modeling environment where the users can graphically model their compliance rules, in drag and drop style, using custom-built visual notations for the compliance patterns presented in Figure 2. In addition, the modeling environment is equipped with a plugin that exports the rules in XML format to be deployed to the monitoring engine. The *compliance rules editor*is built on top of the *Oryx* editor[8] [16], a popular open source environment for process modeling. Figure 6 shows a snapshot of the rule's editor capturing compliance requirement 3 from Table I, which visualizes compliance rules $R3.1$, $R3.2$ and $R3.3$.

B. Compliance Monitoring Engine

The compliance monitoring engine is responsible for *continuously* evaluating the compliance status of the running process instances against defined compliance rules. The monitoring engine receives the compliance rules in XML format from the rules editor and translates them into a set of queries that are continuously evaluated against the stream of events received from the process execution engine. The engine triggers the execution of the *compliance actions* for any detected violations of the compliance rules in addition to reporting the information of the violated rule and the violating business process instance to the end user by means of updating the monitoring dashboard.

The compliance monitoring engine was built as a service using C# .NET and has used an open source engine for complex event processing (CEP) [23] and event series analysis, *Esper*[9]. In principle, Esper was chosen because of its scalability, ease of modeling as reported in [37]. It provides an environment for developing applications that can process large volumes of incoming messages or events, regardless of whether incoming messages are historical or real-time in nature. It supports filtering and analyzing events in various ways, and responding to conditions of interest. In particular, Esper provides an SQL-like language, Event Processing Language (*EPL*)[10], that provides the standard *SELECT*, *FROM*, *WHERE*, *GROUP BY*, *HAVING* and *ORDER BY* clauses. In this context, streams replace tables as the source of data with events replacing rows as the basic unit of data.

A process execution engine provides the *raw* events the monitoring component needs to keep evaluating the compliance status. To provide these events, we have extended the Activiti [11] business process management platform with event emitters that propagate events reflecting the evolution of process instances and their tasks to the monitoring component. For this, every new instance created or a change of state in an existing process/task instance will be communicated to the monitoring component as a new entry on the respective event stream. From there, ESPER can evaluate the different anti pattern queries against these streams to detect violations. The built-in process model within the Activiti platform has been used to define process models to be executed.

In the rest of this section, we discuss the mapping of anti patterns described in Section V into EPL queries. These are shown as parameterized queries where parameters are enclosed in curly brackets{}. These parameters are actualized at rule registration time at the monitoring component.

```
insert into RuleViolationEvent (processID,Message,RuleID,
RuleType)
select s.ProcessID,'Event␣{Antecedent}({task})␣occurred␣less
than␣{MinOccurs}␣within␣{ScopeStartEvent}({ScopeStartTask})
and␣{ScopeEventEvent}({ScopeEventTask})␣in␣process␣',
 '{RuleID}','{RuleType}'
FROM PATTERN [
every(s={ScopeStartEvent}(cast(s.Task,string)=
'{ScopeStartTask}')->(e={ScopeEndEvent}(cast(e.Task,string)
='{ScopeEndTask}',ProcessID=s.ProcessID) ))] as scope
WHERE {MinOccurs} > (select count(*) from {Antecedent}.win:
keepall() as T
WHERE cast(T.Task,string)='{AntecedentTask}'
and(T.TimeStamp between scope.s.TimeStamp and
scope.e.TimeStamp))
```

Listing 1: Query to detect below-min-occurrences of a rule antecedent

```
insert into RuleViolationEvent (processID,Message,
RuleID,RuleType)
select s.ProcessID,'Event␣{Antecedent}({task})␣occurred␣more
than␣{MinOccurs}␣within␣{ScopeStartEvent}({ScopeStartTask})
and␣{ScopeEventEvent}({ScopeEventTask})␣in␣process␣',
'{RuleID}','{RuleType}'
FROM PATTERN [
every(s={ScopeStartEvent}(cast(s.Task,string)=
'{ScopeStartTask}')->(e={ScopeEndEvent}(cast(e.Task,string)
='{ScopeEndTask}',ProcessID=s.ProcessID)))] as scope
WHERE {MaxOccurs} < (select count(*) from {Antecedent}.win:
keepall() as T
WHERE cast(T.Task,string)='{AntecedentTask}'
and(T.TimeStamp between scope.s.TimeStamp and
scope.e.TimeStamp))
```

Listing 2: Query to detect above-max-occurrences of a rule antecedent

Listings 1 and 2 define the anti patterns in Definition 5.6 as EPL queries that detect *Exists*, *Absence*, *Sequence* and *Next* anti patterns by specifying the values for $MinOccurs$ and $MaxOccurs$. For the *Absence* anti pattern, we can use the query in Lisitng 1 where $MinOccurs = 0$. To detect the first anti pattern for the *Sequence* rule, we can set the $MinOccurs = 1$.

[6] A video demonstration of our framework implementation is available on https://www.youtube.com/watch?v=wRdZKsOi5x4

[7] The source code of BP-MaaS is available on https://github.com/BP-MaaS/BP-MaaS.

[8] https://code.google.com/p/oryx-editor/

[9] http://esper.codehaus.org/

[10] http://esper.codehaus.org/esper-4.2.0/doc/reference/en/html/epl_clauses.html

[11] http://www.activiti.org/

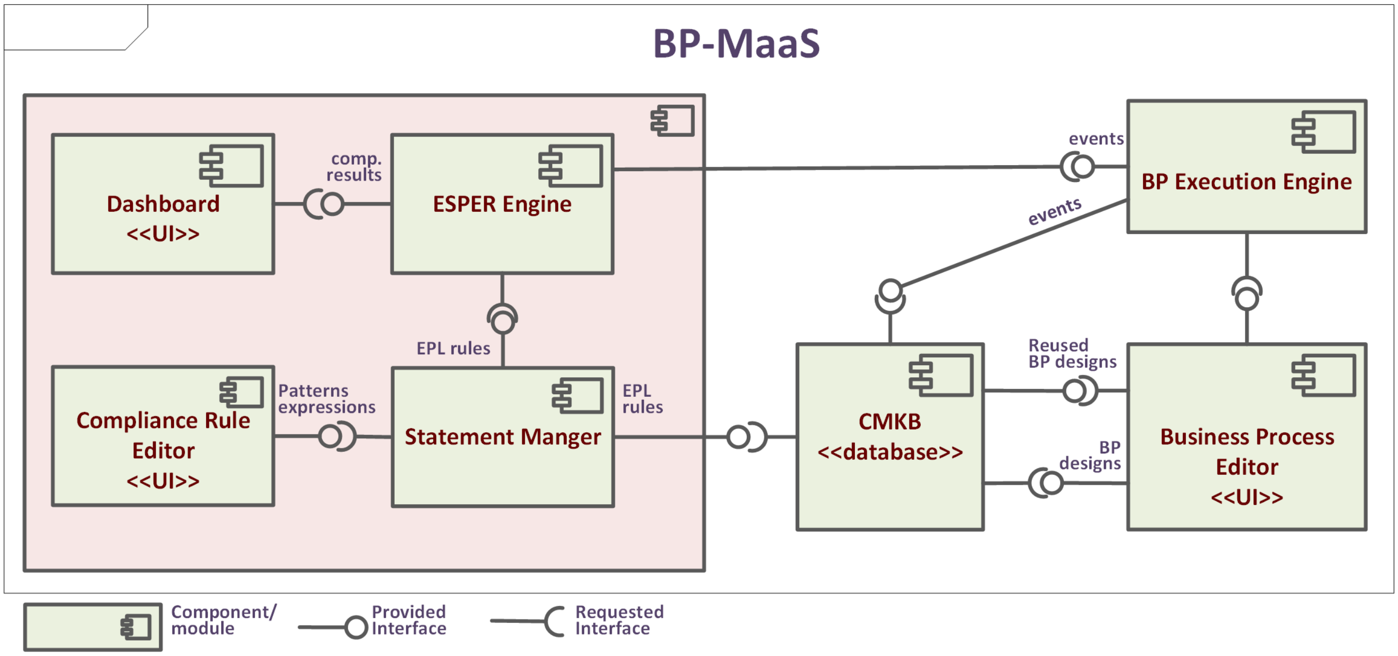


Fig. 5: BP-MaaS architecture represented as a UML component diagram

In Listing 1, EPL *PATTERN* clause is used to look for a sequence of events where the rule's *scope start* is observed and then followed by $->$ *scope end*, this sequence is being matched continuously to the event stream, by means of the $every$ keyword. Whenever there is a match, we get the event instance s matching the start of the scope and the event instance e matching the end of the scope. Then, all event instances of a type matching rule's $antecedent$ whose data payload satisfies the rule's condition and whose $timeStamp$ lies between s and e are counted and compared to the rule's $MinOccurs$. Listing 2 does a similar thing but comparing the occurrences of the $antecedent$ event to the $MaxOccurs$ parameter.

Listing 3, defines an EPL statement for anti patterns of *Sequence* and *Response* rules, cf. 5.7, where the pattern *scope start* followed by $antecedent$ then followed by any of *scope end*, one of the forbidden events, or time span of the rule is observed but never the $consequent$. This query detects the violation for $isBefore = true$ response rules.

```
insert into RuleViolationEvent (processID , Message , RuleID ,
RuleType )
select s . ProcessID , 'Event␣{ AntecedentEvent }({ AntecedentTask })
was␣not␣followed␣by␣{ ConsequentEvent }({ ConsequentTask }) within
␣{ ScopeStartEvent }({ ScopeStartTask })␣and
␣{ScopeEndEvent }({ ScopeEndTask })␣in␣process␣' , '{RuleID}' ,
 '{RuleType}'
FROM PATTERN [
every ( s= { ScopeStartEvent }( cast ( s . Task , string )=
{ ScopeStartTask})−>(every ( Antecedent={ AntecedentEvent }
( cast ( Task , string )={ AntecedentTask } ,
processID=s . processID , Implies ( Data , { condition }))
−>((e={ ScopeEndEvent }( cast ( Task , string )={ ScopeEndTask } ,
processID=s . processID ) or absent={ Absent }( cast ( Task , string )
in {WA} , processID=s . processID ) or timer : interval ({ TimeSpan })
and not Consequent = { ConsequentEvent }( cast ( Task , string )=
{ ConsequentTask } , processID=s . processID ))))]
```

Listing 3: Query to detect *Sequence* and *Response* violations

The query in Listing 4 detects the violation when a *Response* rule requires that the time span elapses before the consequent can take place, i.e., $isBefore = false$.

```
insert into RuleViolationEvent ( processID , Message , RuleID ,
RuleType )
select s . ProcessID , 'Event␣{ AntecedentEvent }({ AntecedentTask })
was␣followed␣by{ ConsequentEvent }({ ConsequentTask })␣within
{ ScopeStartEvent }({ ScopeStartTask })␣and␣{ ScopeEndEvent }
({ ScopeEndTask })␣in␣process␣but␣before␣{TimeSpan}␣elapses ' ,
'{RuleID}' , '{RuleType}'
from pattern
[ every ( s={ ScopeStartEvent }( cast ( Task , string )=
{ ScopeStartTask})−>(every Antecedent={ AntecedentEvent }
( cast ( Task , string )={ AntecedentTask } , processID=s . processID ,
Implies ( Data , { condition }))−>((e={ ConsequentEvent }
( cast ( Task , string )={ ConsequentTask } , processID=s . processID )
and not timer : interval ({ TimeSpan } ))))))]
```

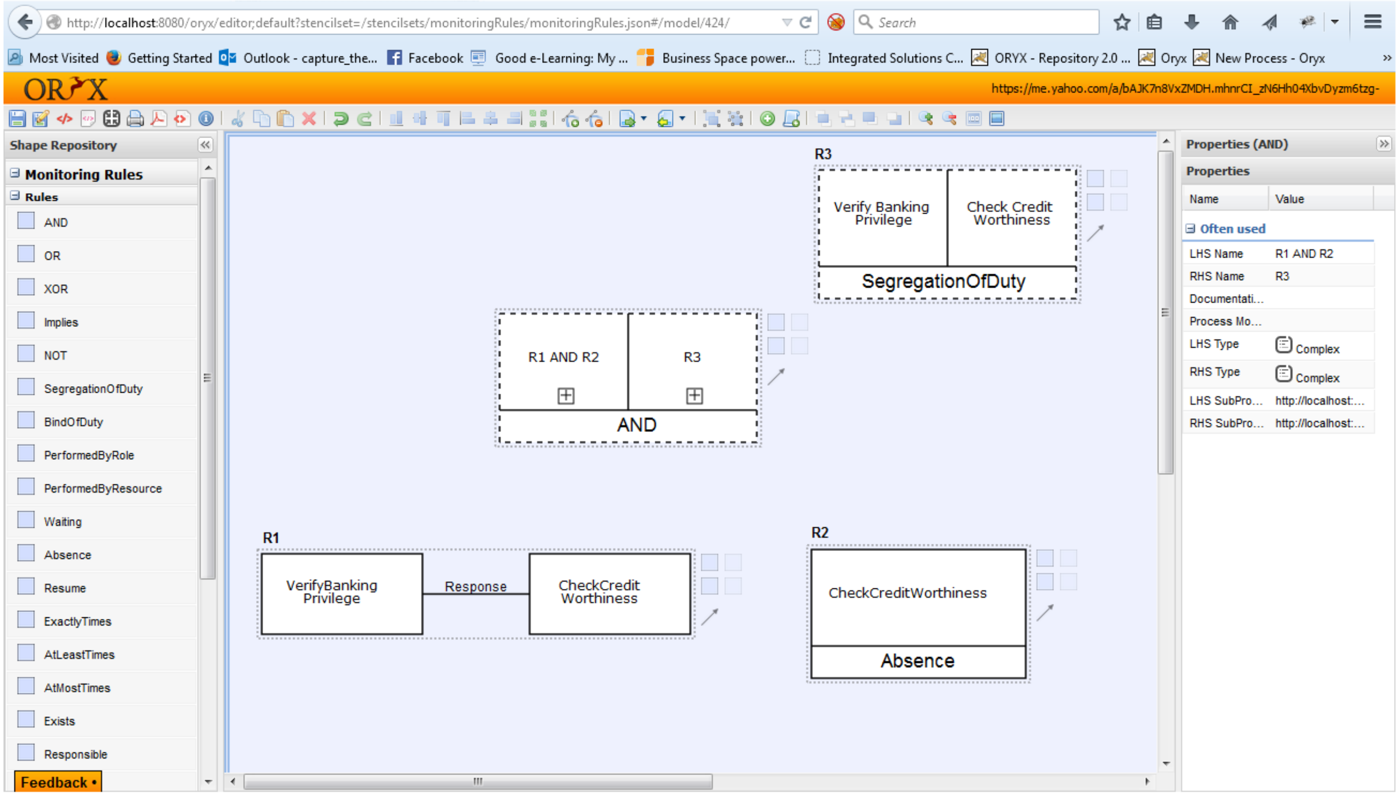
Listing 4: Query to detect *Response* violations for $isBefore = false$

```
insert into RuleViolationEvent ( processID , Message , RuleID ,
RuleType )
select s . ProcessID , 'One␣to␣one␣response␣violation ,␣successive
occurrences␣of␣Event␣{ AntecedentEvnt }({ AntecedentTask })
were␣detected␣without␣detecting␣Event
{ ConsequentEvent }({ ConsequentTask })␣in␣between␣within
{ ScopeStartEvent }({ ScopeStartTask })␣and␣{ ScopeEndEvent }
({ ScopeEndTask })␣in␣process ' , '{RuleID}' , '{RuleType}'
from pattern
[ every ( s={ ScopeStartEvent }( cast ( Task , string )={ ScopeStartTask })
−>(every Antecedent={ AntecedentEvent }( cast ( Task , string )=
{ AntecedentTask } , processID=s . processID )
−>(Antecedent2={ AntecedentEvent }( cast ( Task , string )=
{ AntecedentTask } , processID=s . processID )
and not Consequent={ ConsequentEvent }( cast ( Task , string )=
{ ConsequentTask } , processID=s . processID ))))]
```

Listing 5: Query to detect one to one response anti pattern

Listing 5, is dedicated to the detection of *One to One Response* anti pattern, cf. Definition 5.8. The query looks for two occurrences of the $antecedent$ event, after the *scope start* without having any occurrences of $consequent$ event in between.

```
insert into RuleViolationEvent ( processID , Message , RuleID ,
```



Name	Value
LHS Name	R1 AND R2
RHS Name	R3
Documentati...	
Process Mo...	
LHS Type	Complex
RHS Type	Complex
LHS SubPro...	http://localhost:...
RHS SubPro...	http://localhost:...

Fig. 6: Visual representation of compliance req. 3 ($R3.1$, $R3.2$, $R3.3$) from Table I

```
RuleType)
select  s.ProcessID,'Precedes␣Rule␣violation:
Event␣{ConsequentEvent}({ConsequentTask})␣never␣occurred
before␣{AntecedentEvent}({AntecedentTask})␣within
{ScopeStartEvent}({ScopeStartTask})␣and
{ScopeEndEvent}({ScopeEndTask})␣in␣process','{RuleID}',
'{RuleType}'
FROM PATTERN [
every(s={ScopeStartEvent}(cast(Task,string)=
'{ScopeStartTask}')->((not Consequent={ConsequentEvent}
(cast(Task,string)='{ConsequentTask}', ProcessID=s.ProcessID)
and not e={ScopeEndEvent}(cast(Task,string)='{ScopeEndTask}',
ProcessID=s.ProcessID))
 until Antecedent={AntecedentEvent}(cast(Task,string)=
 '{AntecedentTask}', ProcessID=s.ProcessID)))]
```

Listing 6: Query to detect Precedence anti pattern where consequent never occurred

Listing 6, realizes the *Precedes* anti pattern, cf. Definition 5.9, where the *scope start* is observed and later on followed by *antecedent* where there is neither *scope end* nor *consequent* events observed in-between. Listing 7, on the other hand detects the other possibility of violation where after *scope start consequent* is observed but no *scope end* followed by either a timer indicating that the rule's *timeSpan* has elapsed or one of the forbidden events and then observing the rule's *antecedent* event occurrence.

```
insert into RuleViolationEvent (processID, Message, RuleID,
RuleType)
select s.ProcessID, 'Event␣{ConsequentEvent}
({ConsequentTask})␣is␣observed␣and␣either␣{TimeSpan}␣or
one␣of␣the␣tasks␣in␣{WA}␣were␣observed␣and␣before␣that
␣{AntecedentEvent}({AntecedentTask})␣was␣observed␣within
␣{ScopeStartEvent}({ScopeStartTask})␣and
␣{ScopeEndEvent}({ScopeEndTask})in␣process',
 '{RuleID}','{RuleType}'
FROM PATTERN [
every(s={ScopeStartEvent}(cast(s.Task,string)=
'{ScopeStartTask}')->(every(Consequent={ConsequentEvent}
(cast(Consequent.Task,string)='{ConsequentTask}',ProcessID=
s.ProcessID)->(e={ScopeEndEvent}(cast(e.Task,string)=
'{ScopeEndTask}',ProcessID=s.ProcessID)
or absent={Absent}(cast(absent.Task,string) in ({WA}),
ProcessID=s.ProcessID)
or timer:interval({TimeSpan}))
->(Antecedent={AntecedentEvent}(cast(Antecedent.Task,string)=
'{AntecedentTask}',ProcessID=s.ProcessID)))))]
```

Listing 7: Query to detect Precedes anti pattern where forbidden or timer events occur

Listing 8 is an EPL query to detect a violation to a *Precedes* rule where not enough time has elapsed between the occurrence of *antecedent* and *consequent*, $isBefore = false$.

```
insert into RuleViolationEvent (processID, Message, RuleID,
RuleType)
select s.ProcessID, 'Event␣{ConsequentEvent}({ConsequetTask})
occurred␣before␣{AntecedentEvent}({AntecdedentTask})␣within
{ScopeStartEvent}({ScopeStartTask})
and␣{ScopeEndEvent}({ScopeEndTask})␣but␣sooner␣than␣the␣time
span␣{TimeSpan}␣in␣process','{RuleID}','{RuleType}'
FROM PATTERN [
every(
s={ScopeStartEvent}(cast(s.Task,string)='{ScopeStartTask}')
->(every(Consequent={ConsequentEvent}(cast(Consequent.Task,
string)='{ConsequentTask}',ProcessID=s.ProcessID)
and not e={ScopeEndEvent}(cast(e.Task,string)=
'{ScopeEndTask}',ProcessID=s.ProcessID)
->(Antecedent={AntecedentEvent}(cast(Antecedent.Task,string)=
'{AntecedentTask}',ProcessID=s.ProcessID)
and not timer:interval({TimeSpan})))))]
```

Listing 8: Query to detect Precedes anti pattern where time span has not elapsed between antecedent and consequent

Listing 9, is dedicated to the detection of *One to One Precedes* anti pattern, cf. Definition 5.10.

```
insert into RuleViolationEvent (processID, Message, RuleID,
RuleType)
select s.ProcessID,'Two␣or␣more␣occurrences␣of␣Event
{AntecedentEvent}({AntecedentTask})␣were␣detected␣without
detecting␣Event␣{ConsequentEvent}({ConsequentTask})
in␣between␣within␣{ScopeStartEvent}({ScopeStartTask})␣and
{ScopeEndEvent}({ScopeEndTask})␣in␣process',
'{RuleID}','{RuleType}'
FROM PATTERN [
every(s={ScopeStartEvent}(cast(s.Task,string)=
'{ScopeStartTask}')->(every
(Antecedent={AntecedentEvent}(cast(Antecedent.Task,string)=
'{AntecedentTask}',ProcessID=s.ProcessID)->(
(not Consequent={ConsequentEvent}(cast(Consequent.Task,
string)='{ConsequentTask}',ProcessID=s.ProcessID))
until Antecedent2={AntecedentEvent}(cast(Antecedent2.Task,
string)='{AntecedentTask}',ProcessID=s.ProcessID))))]
```

Listing 9: One to one Precedes anti pattern in EPL

```
insert into RuleViolationEvent (processID, Message, RuleID,
RuleType)
select s.ProcessID, 'Events␣{AntecedentEvent}
({AntecedentTask})␣and␣{ConsequentEvent}({ConsequentTask})
were␣performed␣by␣Resource' + s.Antecedent.resource+ '␣within
{ScopeStartEvent}({ScopeStartTask})␣and
{ScopeEndEvent}({ScopeEndTask})␣in␣process',
'{RuleID}','{RuleType}'
FROM PATTERN [
every(
s={ScopeStartEvent}(cast(s.Task,string)='{ScopeStartTask}')
->(every(Antecedent={AntecedentEvnt}(cast(Antecedent.Task,
string)='{AntecedentTask}',ProcessID=s.ProcessID)
-> every(Consequent={ConsequentEvent}(cast(Consequent.Task,
string)='{ConsequentTask}',ProcessID=s.ProcessID))))]
WHERE Consequent.Resource = Antecedent.Resource
```

Listing 10: Separtion of Duty anti pattern

10 is the EPL statement to detect *Separation of Duty* anti patterns. Actually, there has to be another query where the match of $antecedent$ and $consequent$ are swapped to address the fact that the two events can occur in an arbitrary order. Requirement 2 " If a manager needs to travel, the request has to be approved by another manager." can be represented as a separation of duty rule.

1) Detecting Violation of Composite Patterns: In principle, we define an event stream for *rule violation* events as there are event streams for the different process and activity lifecycle events. In particular, activity event streams are supplied by events from the execution environment, cf. 4, whereas the *rule violation* stream is supplied by events based on the matching of the different anti pattern queries. This is shown as starting with `Insert into RuleViolationEvent` in all previous EPL statements. So, whenever a match to the anti pattern occurs, a new instance of the complex event *RuleViolationEvent*, cf. 5.1, is created and sent over that stream.

In case that the initial rule was a composite rule, another query can pick the violation event and act upon it based on the rule's logic. For instance if the rule is of the form $r : r1 \wedge r2$, the query in Listing 11 is defined to detect if there is any occurrence of a rule violation event for either $r1$ or $r2$ and generates another rule violation event for r.

```
insert into RuleViolationEvent
select processID,Message,{r.RuleID} from RuleViolationEvent
where RuleID = {r1} or RuleID = {r2}
```

Listing 11: Monitoring violation of AND composite rule

For complex rules on the form $r = r1 \vee r2$, a query as in Listing 12 is defined to detect the occurrences of violation events of both $r1$ and $r2$ to generate a violation event of the composite rule r. We actually define another query where the sequence of violation events of $r1$ and $r2$ is swapped as the violations might occur in any order. If the composite rule would have more operands, we depend on the associativity of the OR operator to break it into a sequence of binary operators, i.e., $r1 \vee r2 \vee r3 \equiv ((r1 \vee r2) \vee r3)$.

```
insert into RuleViolationEvent
select processID,Message, {r.RuleID} from pattern
[every(
r1=RuleViolationEvent(RuleID={r1})
->(
r2=RuleViolationEvent(RuleID={r2},processID={r1}.processID)))]
```

Listing 12: Monitoring violation of OR composite rule

C. Monitoring Dashboard

The monitoring dashboard is a user-friendly interface for the end-user to monitor the stream of events and manipulate (e.g., adding, removing, activating, deactivating) the set of registered compliance rules in addition to being able to receive the notifications and statistics about the running process instances and detected non-compliance instances (Figure 7). The monitoring dashboard component has been implemented as a .NET program which is responsible for communicating the information between the compliance monitoring engine and the end user. All the three component are loosely coupled as there are no direct dependencies among them. They communicate via messages so that each of them can be replaced as long as the message contract is kept the same.

It also should be noted that our framework remains agnostic towards the different systems which can be used for implementing the different components. For example, any process execution environment can be the source of raw events as long as the events are generated according to the reference lifecycle models of processes and tasks. The Esper engine can be replaced with any other stream processing engine (e,g. *StreamBase*[12], *Apache Storm*[13]).

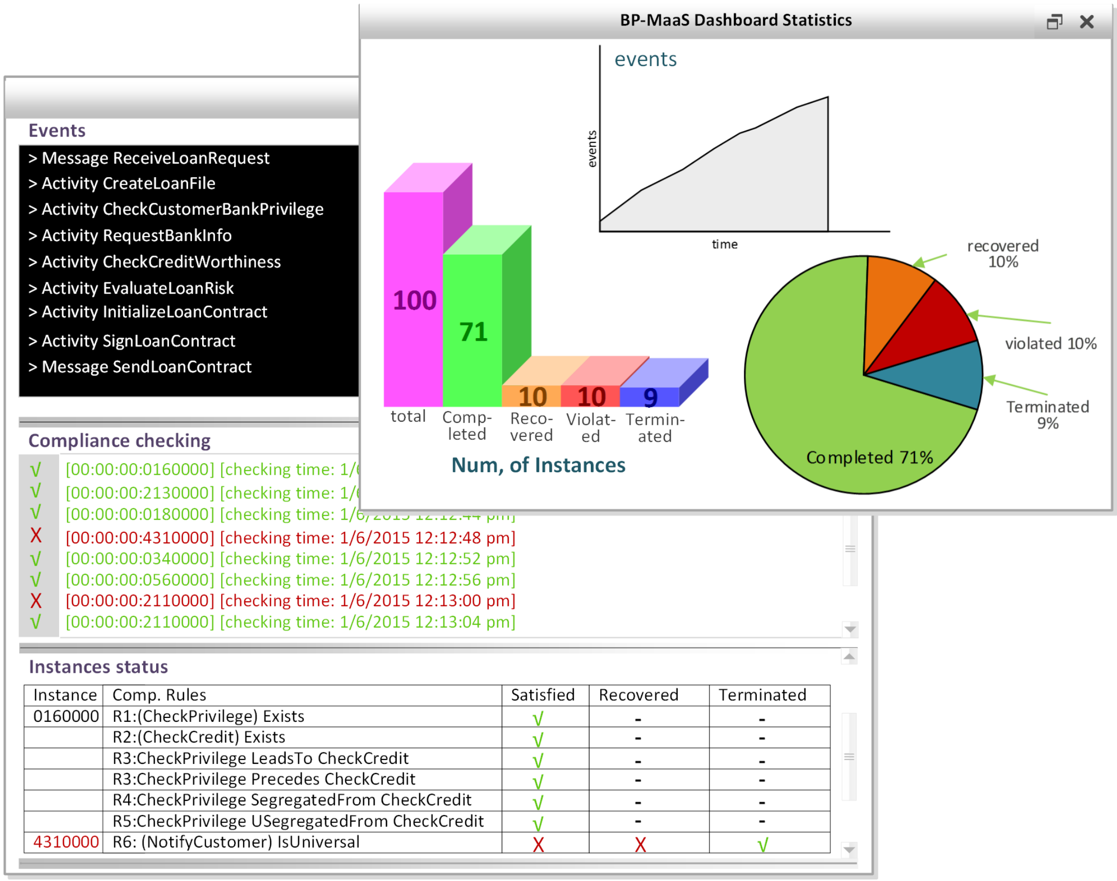
VII. Evaluation

The utility of a design artifact must be rigorously demonstrated via well-executed evaluation methods [28]. Observational methods, such as case studies and field studies, allow an in-depth analysis of the artifact and the monitoring of its use in multiple projects within the technical infrastructure of the business environment.

In Section III, we have introduced two case studies that involved processes operating in the banking domain, addressing different business concerns and entailing a faithful set of rich and diverse compliance requirements that any financial institution is required to comply with. The case studies involved representing relevant compliance requirements in compliance patterns (introduced in Section II-B) with the main objective of investigating the applicability and utility of the overall monitoring approach proposed in this paper, to represent and continuously monitor applicable compliance requirements against running BPMN instances, by means of

[12] http://www.streambase.com/

[13] http://storm.incubator.apache.org/



Events

> Message ReceiveLoanRequest
> Activity CreateLoanFile
> Activity CheckCustomerBankPrivilege
> Activity RequestBankInfo
> Activity CheckCreditWorthiness
> Activity EvaluateLoanRisk
> Activity InitializeLoanContract
> Activity SignLoanContract
> Message SendLoanContract

Compliance checking

```
√ [00:00:00:0160000] [checking time: 1/(
√ [00:00:00:2130000] [checking time: 1/(
√ [00:00:00:0180000] [checking time: 1/6/2015 12:12:44 pm]
X [00:00:00:4310000] [checking time: 1/6/2015 12:12:48 pm]
√ [00:00:00:0340000] [checking time: 1/6/2015 12:12:52 pm]
√ [00:00:00:0560000] [checking time: 1/6/2015 12:12:56 pm]
X [00:00:00:2110000] [checking time: 1/6/2015 12:13:00 pm]
√ [00:00:00:2110000] [checking time: 1/6/2015 12:13:04 pm]
```

Instances status

Instance	Comp. Rules	Satisfied	Recovered	Terminated
0160000	R1:(CheckPrivilege) Exists	√	-	-
	R2:(CheckCredit) Exists	√	-	-
	R3:CheckPrivilege LeadsTo CheckCredit	√	-	-
	R3:CheckPrivilege Precedes CheckCredit	√	-	-
	R4:CheckPrivilege SegregatedFrom CheckCredit	√	-	-
	R5:CheckPrivilege USegregatedFrom CheckCredit	√	-	-
4310000	R6: (NotifyCustomer) IsUniversal	X	X	√

Fig. 7: BP-MaaS Monitoring Dashboard Screenshot.

anti-patterns. This validation and evaluation step also enabled us to identify the limitations of the approach and the potential enhancement points.

The loan approval case study used as the running scenario throughout this paper compromised 7 high-level compliance requirements (compliance constraints as they originate from various compliance sources). By applying the compliance refinement methodology in [60, 59], this has resulted in 15 organization-specific compliance requirements based on the BPMN model shown in Figure 3. Examples of these refined/internalized compliance requirements are given in Table I.

The compliance patterns used to realize the 15 requirements are the *Response* pattern, which constitutes 7 compliance rules (note that a compliance requirement can be represented by one or more compliance rules) as shown in Table I). *Exists* pattern is used in one rule, *Precedes* pattern in 3 compliance rules, *Next* and *segregationOfDuty* patterns are used in one compliance rule, respectively. *PerformedBy* resource pattern is utilized in 5 rules, *Mutual Exclusive Choice* composite pattern is used to represent one compliance rule. And *Time span and isBefore = false* real time pattern is used in one rule in conjunction with the *Response* pattern.

The second case study targets the anti-money laundering banking scenario. The U.S. Patriot act of 2001 [44] was considered as the main source of compliance requirements of this case study, which constitutes a large number of compliance requirements structured into twelve sections. As discussed previously in Section III, the interpretations of embedded requirements and encoding them in the Semantics of Business Vocabulary and Business Rule (SBVR) standard [45] as a structured natural language is ongoing work in GRCTC [20]. For this case study, we have only considered the suspicious activity detection and reporting part of the AML practices.. From the U.S. Patriot act, we found 27 compliance requirements relevant to the suspicious activity detection and reporting business process.

The compliance patterns used to realize the 27 requirements are the *Response* pattern 11 requirements, the *Exists* pattern 1 requirement, the *Performed By Role* pattern 4 requirements, the *Precedence* pattern 3 requirements, the *Absence* pattern 1 requirement, and the *Next* pattern 1 requirement. Among the 11 *Response* rules, one rule used the *time span* property whereas one other rule used the *with absence* property. Out of the three *Precedence* rules, two rules used the *with absence* property. Table II shows a excerpt of these requirements, by referring to its clause number inside the U.S. Patriot act.

Table III gives the type and number of compliance requirements covered within the case studies, and whether the case study participants were able to express these requirement effectively using compliance patterns and continuously monitor them against running business process instances by means of anti-patterns, using the prototypical implementation discussed

TABLE II: An excerpt of the Compliance Requirements relevant to the AML case study

Compliance Requirements	Source	Control	Pattern Representation
R1: Required Standards for AML MSB programs	Section $1022.210 (d)(1)(i)(A)(B)(C)(D)	Anti-money laundering programs for money services business: policies, procedures, and internal controls	R1.1: Response(Send Payment Order ,Verify Customer Identification) R1.2: Response(Send Payment Order ,Retain Supporting Documents)
R2: Identity and Reporting related provisions	Section $1022.210 (d)(1)(iv)	It is necessary that customer identification and verification includes name, date of birth, address and identification number of a person.	R2.1: Exists(Verify Customer Identification, Check{ CustomerIdentification.name is not null, CustomerIdentification.DoB is not null, CustomerIdentification.address is not null, CustomerIdentification.ID is not null })
R3: Identity and Reporting related provisions	section $1022.210 (d)(1)(iv)	It is obligatory that customer identification for amounts above $10,000 to take place within one calendar day	R3.1: Response(Initiate Money Transfer, Verify Customer Identification, time span = before 1 day)
R4: Retention of Record of related provisions	section $1022.320(c)	It is obligatory that each money service business maintains copy of each Suspicious Activity Report-MSB filed and business record of any supporting documentation for 5 calender years.	R4.1: Absence(Delete Suspicious Transaction Records) R4.1: Next(Receive Add Doc Request, Send Add Doc)
R5: Confidentiality	Section$1022.320 (d)(ii)(A)1	It is permitted that each money service business to disclose a suspicious activity report to FinCEN or an appropriate law enforcement agency if the person involved in a suspicious transaction is not notified that a suspicious activity report has been filed.	R5.1: Precedence(Disclose Suspicious Activity Report, Process.Start, WA={ Receive Deferment Notification})

TABLE III: Categories and Numbers of Compliance Req. Covered in the Case Studies

		Number of Comp. Requirements			Supported by our approach?	
Type of Controls		Case St.1: Loan Processing	Case St.2: Anti-money Laundering	TOTAL	Yes	No
PROCESS	- Control flow	1	2	3	3	-
	- Data Requirements	-	3	3	3	-
	- Resources	2	-	2	2	-
	- Control flow- Data	1	7	8	8	-
	- Control flow- Resources	3	1	4	4	-
	- Control flow- Real time	1	1	2	2	-
	- Resource-data	1	-	1	1	-
	- Control flow-Resources-Real time	1	-	1	1	-
	- Control flow-Resources-Data	1	2	3	3	-
	- Control flow-Data-Real time	-	1	1	1	-
TOTAL		11	17	28	28	-
TECHNICAL/ MANUAL	- Data Requirements	3	3	8	-	8
	- Others	-	7	7	-	7
PHYSICAL		1	-	1	-	1
TOTAL		15	27	42	28	16

in Section VI. As shown in Table III, compliance requirements are classified into three distinct classes:

- *Process*: compliance requirements that are relevant to the policies and practices concerning the design and execution of business process models. Authorizations, approvals, inspections, segregation of duties applied through business tasks and other elements are examples of such requirements.
- *Technical/Manual*: are requirements that involve the use of devices or systems mainly for authentication, encryption or security purposes. Examples include firewalls and intrusion prevention/detection systems.
- *Physical*: are requirements that involve largely the institution of physical means, such as locks, fences and alarms, to guard critical assets.

As shown in Table III, the requirements that are classified as process constituted the majority. These requirements were of particular interest to us, as this type of constraints is the main target for the runtime monitoring approach introduced in this paper. They involved rules concerning mainly segregation of duties, access-rights, condition-based sequencing of activities, data processing requirements and real time constraints. For this category, we further classified it into classes corresponding to the four structural facets of business processes; i.e., control flow, data requirements, employed resources, and real-time, and their combination, as a single compliance requirements might be, for example, addressing the control flow and data aspects of BPs. Real-time constraints usually do not exist by their own, that is why it was not included as one of the subclasses of the process category. As show in Table III, 28 compliance requirements in total from the two case studies belong to this category, and they all could be effectively modelled, executed and monitored by applying our approach.

Technical requirements mainly involve constraints regarding data processing, e.g., rules that are related to the structure and integrity of the data manipulated within the processes. They typically demanded for sequential numbering of certain business objects, such as orders or invoices, or the retention of data for a specific period of time. Also, it involved the requirements of manual management reviews and reconciliations, which are inherently manual by nature. Other constraints in this category mandate the existence of authentication, encryption or security devices and/or software components, which could be checked on other enterprise systems level, but not on this business process level. Requirements from the technical and physical categories are subsequently can not be supported by our approach. In total, 28 compliance requirements out of 42 are fully supported by our approach, which represents a ratio of around 70%. Despite the limitations discussed above, we can conclude that within the process category of compliance requirements, the proposed runtime compliance monitoring approach is an effective means for expressing runtime compliance requirements in a user-friendly and abstract form, and supports the continuous monitoring of these constraints by applying a novel evaluation mechanism based on anti-patterns. We can also conclude that compliance requirements fall in the *process* category represent a major subset of the compliance requirements imposed on real-world scenarios. Future work involves intensifying this validation and evaluation step by applying our approach on a larger scale case studies in different domains.

VIII. Related Work

Compliance management has been an active research area recently from both the academic and industrial communities, given the high-cost associated with non-compliance, including business failures, bankruptcy, significant fines and even criminal penalties. In the following, prominent related-work efforts in runtime compliance monitoring is summarized and appraised against the work proposed in this paper. For a detailed comparison framework, we refer the reader to [32].

Runtime monitoring requires business process models to be reduced to some abstract representation, which are built up by collecting runtime information (e.g. exchanged messages sequences, performed activities). On the other hand, runtime monitoring also requires compliance requirements to be structurally/formally represented using a formal/structural language, e.g. LTL, CTL, ECA rules. In addition, various querying languages could also be utilized, such as *BP-Mon* [11] and XPath [62]. The actual compliance checking between abstract traces and formal rules/queries is performed by a runtime compliance checker (engine), which is usually an external component that is incorporated into the execution environment, but could also be an internal component. The checker can check the adherence to the requirements either after the execution is completed, or synchronous with the execution, following a more proactive approach. In the following, we classify related work into four categories; *graph-based* approaches, *formal-based* approaches, *XML querying* approaches and *complex-event* processing, which will be discussed in the next subsections and appraised against the work presented in this paper.

A. Graph-based Monitoring Approaches

Graph-based approaches mainly target the design-time phase of the business process lifecycle for (sub-)process models querying, substitution,and compliance checking; examples are: [29], [52], [3], [53], [17]. On the other hand,few studies have addressed runtime compliance monitoring, which include [11, 33]. *Business Process Monitoring* (*BP-Mon*) is a graphical query language proposed in [11] to visually represent monitoring requirements against BPEL models, abstracted into event traces. Graph matching techniques (*homomorphism*) are then exploited to evaluate the compliance of *completed* running BPEL instances, focusing on control-flow and timing constraints. Similarly, the study in [33] adopts a graph-based compliance rule language to capture compliance requirements, supporting sequence, data and real-time constraints, where runtime compliance checking is done *synchronously* with the execution.

B. Formal Monitoring Approaches

Influential *formal monitoring* approaches are reported in [27] , [36], [38], [9], [10], [24], [35] by founding compliance requirements on a formal/mathematical language. The study in [36] uses *Event Calculus* (EC) as the formal basis of monitored constraints against BPEL models. EC is an expressive language; however it is excessively difficult to be used. Monitoring is implemented as integrity-checking technique on *completed* executions. EC is also used in [38], however to cope with the complexity of EC, Declare language [47] is utilized as a graphical intermediate representation. Logic programming reasoning is then used to dynamically reason about partial, evolving execution traces. These approaches [7, 38] focus on control-flow and timing constraints.

Model-checking formal approaches is adopted in [7, 27, 34]. *LTL-FO+* is proposed in [27] as an extension to LTL that includes full first order quantification over data, focusing on control-flow and data requirements. In [34], *Declare* [47] is used, which is mapped into LTL, only supporting control-flow constraints, where monitoring is done synchronously with the execution. The same approach is applied in [35] using Declare and LTL to capture compliance requirements, while declarative process models are considered instead, mainly to detect conflicting compliance requirements.

Metric first-order temporal logic is used in [10], supporting past and bounded future operators. This approach [10] provides an optimized monitoring technique addressing control-flow and timing constraints, however the complexity of the adopted logic is not tackled. An extension is made in [9] to support data-constraints. The REALM model is proposed in [24] which constitutes (among others) a conceptual model and metadata. The conceptual model captures the concepts and relationships related to a certain domain (domain ontology), which are used to build compliance rules. To ensure the rigor of the framework, compliance rules are first represented formally using Past LTL, then mapped to proprietary notations. Compliance checking is also performed by a proprietary component (Active Correlation Technology (ACT)) that correlate events to detect runtime violations. The approach supports control-flow and real-time constraints.

C. Query-based and Rule-based Approaches

Prominent *XML querying* approaches are [26], [61], [54]. In [26] and [61], requirements in LTL are translated into equivalent XQuery expressions, and an XQuery engine is used to evaluate the compliance, focusing on sequence and data constraints. *BPath* [54] is proposed as an XPath extension with LTL modalities. BPath expressions are then mapped into XPath, and a native XML query engine is utilized, supporting sequence and timing constraints.

Influential Rule-based proposals include [41], [8], [14], [42], [7]. In [8], desired properties and constraints on BPEL systems are specified in WS-CoL (Web Service Constraint Language), a special-purpose assertion specification language that borrows its roots from JML (Java Modeling Language) and extends it with constructs to gather data from external sources. WSCoL are interweaved into BPEL specification, and a dedicated monitoring manager evaluates the compliance by focusing on data constraints. In [14], compliance requirements are represented in Prolog and verified against a workflow language, supporting sequence and timing constraints using a rule engine.

In [41] a generic runtime compliance management framework is proposed, which is based on a set of Dwyer′s property specification patterns [19], and provides a high-level conceptual model for compliance requirements refinements and the definition of recovery actions, as response to detected violations. The framework is realized by implementing it using BPMN models and Event-Condition-Action (ECA) rules. This approach is closely related to the work presented in this paper, however, our work relies on a wider set of novel compliance patterns; moreover, our approach addresses the four structural facets of the BP lifecycle; and we define a novel evaluation approach based on anti-patterns.

D. Complex Event Processing Monitoring Approaches

Complex Event Processing (CEP) technology is utilized in [40], [57], [64], [56]. Prominent efforts in this direction use Event Pattern Languages (EPLs) to capture relevant requirements and constraints. In [40], a model-driven engineering approach is adopted, such that a high-level DSL language is introduced for the abstract specification of compliance constraints, with support for sequence and resource constraints.

The work in [64] only considers sequential requirements, where an approach is also introduced to filter and aggregate query results to provide compact feedback on deviations. Business processes are modelled in [57] as event flows where compliance requirements are structurally represented in a conceptual graphical rule model the authors also proposed; and then a CEP engine (SARI) [39] is utilized to check the compliance, with support for sequence and timing constraints. Major approaches in this category check compliance synchronously with the execution.

E. Discussion

Guided by [28] to help validating the utility, novelty and applicability of the approach proposed in this paper, we have investigated, categorized and analyzed related work efforts in the area of Business process and Web services runtime monitoring, and aligned them against the work proposed in this paper. Table IV presents a summary and evaluation of prominent related work proposals discussed in the above sub-sections. The criteria used in the comparative analysis presented in Table IV are as follows:

- *Category*: refers to the five categories discussed above; i.e.,Graph-based, Formally-based, Query-based, rule-based and CEP.
- *Generic*: takes the value ”Yes” or ”No”, which signifies whether the approach provides a generic compliance monitoring framework.
- *BP Language*: the Business Process Language considered by the approach.
- *Observer*: indicates whether the observer that listens to event-patterns of interest is an internal or external component.
- *Runtime info*: the collected runtime information as events, against which compliance rules are evaluated.
- *Rule Support*: indicates whether the approach propose a solution to tackle with the complexity of the underlying formal/query/rule specification language. ’N/A’ means that no support is provided.
- *Rule language*: the formal/query/rule language as the formal foundation for compliance requirements specification.
- *Evaluation approach*: the runtime verification approach used.
- *Synchronous*: takes the value ”Synchronous” or ”Asynchronous”, and indicates whether the monitoring is done step-by-step synchronous with the execution, or after the execution is completed, respectively.
- *BP facets*: lists the support of the approach to the four structural facets of the BP lifecycle; i.e., control-flow, data, employed resources and real time.
- *Recovery Actions*: takes the value ”Yes”, ”No” or ”Partial”, and indicates whether the approach provides a mechanism to reason about detected violations and/or invoke (semi-) automated recovery actions. Partial means that the framework partially support this criterion; e.g., halting the execution and informing appropriate personnel, however, automated recovery actions are not supported to resolve the non-compliance anomalies, etc.

We can conclude from Table IV that the approach proposed in this paper is a generic compliance monitoring approach that could be concretized to any of the approaches in the above sub-sections. As a proof-of-concept, we have implemented it in CEP (Section VI), which is one of the possible realizations. Table IV distinguishes our work by:

1) We adopt a graphical high-level pattern-based specification compliance language, incorporating a wide range of rich compliance patterns accepted by the community. Compliance patterns are mostly used in the literature in design-time compliance checking, whilst in this paper we applied them to runtime monitoring.
2) We have implemented the adopted patterns in an intuitive graphical manner, which further ease the work of the business and compliance experts.
3) Our approach supports the four structural facets of BPs; control-flow, data, employed resources and timing. The highlighted related-work approaches only support a subset of these classes.
4) Compliance monitoring is performed step-by-step and synchronously with the executions, which is crucial for

TABLE IV: Summay and Evaluation of Related Work

Category	Approach	Generic	BP Language	Observer	runtime info	Rule support	Rule language	Evaluation Approach	Synchronous	BP facets	Recovery Actions
Graph-based	[11]	No	BPEL	Internal	Event traces (activities start /completion)	Graphical BP-Mon	Graph representation	Graph homomorphism	Synchronous	control-flow real time	No
	[33]	No	ADEPT [49]	External	Event traces (activities start /completion)	graphical	Compliance Rule Graphs	Graph Pattern matching	Synchronous	control-flow Data real time	Yes
Formally -based	[27]	No	Workflow language	External	Event traces (activities start /completion)	N/A	LTL-FO+	Model Checking	Asynchronous	control-flow Data	No
	[36]	No	BPEL	External	Event traces (activities start /completion)	N/A	Event Calculus	integrity-checking	Asynchronous	control-flow real time	No
	[38]	No	distributed control systems	External	Event traces (defined events)	Declare [47]	Event Calculus	Logic programming	Synchronous	control-flow real time	No
	[34]	No	Workflow language	External	defined event traces	Declare [47]	LTL	Model Checking	Synchronous	control-flow	Yes
	[10], [9]	No	distributed systems	External	Event traces (activities start /completion)	N/A	MFOTL	Model Checking	control-flow Data real time	Asynchronous	No
	[24]	No	proprietary systems	Internal	defined Event traces	Textual patterns	PLTL mapped into proprietary IBM notations	proprietary correlation engine(ACT)	Data control-flow Data real time	Synchronous	No
Query-based	[26], [61]	No	Web service choreograph	External	event traces (exchanged messages)	N/A	XQuery	XQuery evaluation	Asynchronous	control-flow Data	No
	[54]	No	BPEL	External	event traces (exchanged messages)	N/A	BPath	XPath evaluation	Synchronous	control-flow real time	No
Rule -based	[8]	No	BPEL	External	event traces (exchanged messages)	N/A	WS-CoL	Rule-based reasoning	Asynchronous	control-flow Data	No
	[41]	Yes	BPMN	External	Event traces (defined events)	Dwyer's Patterns [19]	ECA	Rule-based reasoning	Synchronous	control-flow Data employed resources	Yes (No impl.)
	[14]	No	Workflow language	External	Event traces (activities start /completion)	Rule template	Prolog	Rule-based reasoning	Asynchronous	control-flow real time	No
	[7]	No	BPEL	External	Event traces (activities start /completion)	RTML	Java Code	Code-based matching	Asynchronous	control-flow real time	No
CEP	[40]	No	BPMN	Internal	Event traces (activities start /completion)	textual DSL	EPL	Esper engine	Synchronous	control-flow employed resources real time	No
	[56]										
	[64]	No	BPMN	Internal	Event traces (activities start /completion)	N/A	EPL	Esper engine	Synchronous	control flow	Yes
	[57]	No	event-driven process chains model	Internal	Event flows	N/A	Compliance rule model	CEP engine (SARI) [39]	Synchronous	control-flow real time	Yes
Our Approach		Yes	BPMN	Internal	event traces (task-based and instance-based lifecycle)	visual compliance patterns	EPL (PoC)	Anti-Patterns evaluation approach	Synchronous	control-flow Data employed resources real time	Partial

a proactive compliance support. The majority of the approaches discussed check compliance on completed executions.

5) The approach presented in this paper supports event traces complying with both the task-based, and instance-based lifecycles [50] (presented in Section II-A), as opposed to other proposals, which mainly consider the start/completion of activities, or the sending/receiving of exchanged messages.
6) The evaluation technique based on the concept of anti-patterns is novel and technology independent as it doesn't assume specific technologies in place, and can be implemented in various platforms. In addition, the concept of anti-patterns could also be applied to verify compliance in other BP life cycle phases.
7) As discussed in Section VII,the utility and applicability of the approach has been validated on two real-life case studies addressing the banking domain, and the findings indicate that our approach supports a major subset of real-life compliance requirements imposed on the considered scenarios.

As discussed in Section VI, compliance violations reasoning and analysis is supported partially by the proposed framework, such that, when a runtime violation is detected, a dedicated actor is notified to take appropriate action and the violated execution is halted. Extending the proposed framework with an efficient root-cause analysis approach that reason about compliance violations, and enables the (semi-) automatic invocation of defined recovery actions is considered as an ongoing work direction.

IX. Conclusions and Future Work

Business processes form the foundation for all organizations, and as such, are impacted by industry regulations. Business process compliance management is an emergent *business need*, as it has been witnessed that without explicit BP definitions, effective and expressive compliance frameworks, organizations may face litigation risks and even criminal penalties. Therefore, Compliance management should be one of the integral parts of business process management. This paper contributes by presenting a *generic proactive* compliance monitoring framework;i.e. BP-MaaS, addressing the runtime phase of the BP lifecycle. The framework adopts a wide range of expressive high-level compliance patterns for the abstract specification of monitoring requirements. From high-level pattern expressions, corresponding runtime queries can be automatically generated for the actual monitoring. As a PoC, we have adopted the CEP technology as the structural basis of runtime queries. Compliance monitoring is then performed based on the notion of *anti-patterns*, a novel runtime evaluation approach that is technology-independent. As an instantiation artifact, anti-patterns are implemented that evaluates CEP queries against running BP instances to detect runtime compliance anomalies.

Some of the lessons learned are that managing compliance of business processes is a complex and multi-disciplinary task. It requires a multi-faceted approach to the problem involves not only technical aspects rooted in various fields that have to be bridged (such as computer science, business process management, formal methods, and legal studies) but also, social and organizational aspects as it highly involves knowledge work. No matter how sophisticated the offered solutions and the underlying technologies are, BP compliance management cannot be fully automated. Having efficient techniques and solutions in place can only facilitate and improve the quality of the work involved. As also shown by the case studies we conducted, the automated verification and monitoring of compliance are possible only for a certain segment of requirements. Technical and physical related requirements necessitate checks and controls that have to be performed manually by compliance experts.

Future research and development are on-going in several directions to enhance and fully-support the compliance monitoring framework proposed in this paper. First, defining automated recovery actions that could take place whenever a violation is detected, which would necessarily involve the notion of business transactions [46]. Second, providing an efficient technique that enables the prediction of potential runtime violations that support not only that of timing constraints, but also other compliance classes as well. This will require the application of some statistical and analytical models, such as Bayesian networks. Third, self-adapting the running BP instance once a compliance violation has occurred or is predicted to occur, to recover the impacts of the violation, and continue its execution normally after the adaptation.

Future work also includes basing the compliance monitoring framework on semantic repositories. This involves building a set of interrelated semantic ontologies (e.g. business process ontology, compliance ontology, organizational ontology, etc.) using the Ontology Web Language (OWL2.0) standard as part of a central compliance management knowledge base. This will allow us to: (i) conduct a set of preliminary structural analysis using the reasoning tools associated with these technologies, (ii) ensure the ontological alignment between compliance and business specifications, (iii)facilitate the communication between different stakeholders with diverse backgrounds and removes any ambiguity, and (iv)assists in the integration between heterogeneous systems. Last but not least, The validation of the proposed approach will be further intensified by its application on larger scale real-life case studies in different domains, such as healthcare and manufacturing, which is expected to raise some interesting challenges. For example, in the healthcare domain, physicians should be provided with higher levels of flexibility to override some compliance rules (weak constraints), as the patient treatment process is tightly related to the physicianś knowledge and judgment.

Acknowledgment

This work is supported by King Abdulaziz City for Science and Technology (KACST), NSTIP (MAARIFAH) project 11-INF1991-03.

References

[1] W M P Van Der Aalst and A K A De Medeiros. Process Mining and Security : Detecting Anomalous Process Executions. In *(WISP)*, 2004.

[2] Ahmed Awad, Ahmed Barnawi, Amal Elgammal, Radwa Elshawi, Abduallah Almalaise, and Sherif Sakr. Runtime detection of business process compliance violations: An approach based on anti patterns. In *ACM SAC*, 2015.

[3] Ahmed Awad and Sherif Sakr. On efficient processing of BPMN-Q queries. *Computers in Industry*, 63(9):867–881, 2012.

[4] Ahmed Awad, Matthias Weidlich, and Mathias Weske. Specification, Verification and Explanation of Violation for Data Aware Compliance Rules. In *ICSOC/ServiceWave*, 2009.

[5] Ahmed Awad and Mathias Weske. Visualization of compliance violation in business process models. In *BPM Workshops*, 2009.

[6] R. Baldwin, M. Cave, and M. Lodge. *Understanding regulation: theory, strategy, and practice*. Oxford University Press, 2011.

[7] Fabio Barbon, Paolo Traverso, Marco Pistore, and Michele Trainotti. Run-time monitoring of instances and classes of web service compositions. In *ICWS*, 2006.

[8] Luciano Baresi and Sam Guinea. Towards dynamic monitoring of ws-bpel processes. In Boualem Benatallah, Fabio Casati, and Paolo Traverso, editors, *Service-Oriented Computing - ICSOC 2005*, volume 3826 of *Lecture Notes in Computer Science*, pages 269–282. Springer Berlin Heidelberg, 2005.

[9] David Basin, Matus Harvan, Felix Klaedtke, and Eugen Zalinescu. Monpoly: Monitoring usage control policies. In *Proceedings of the 2nd International Conference on Runtime Verification (RV 2011)*, pages 360–364, 2012.

[10] David Basin, Felix Klaedtke, Samuel Müller, and Birgit Pfitzmann. Runtime Monitoring of Metric First-order Temporal Properties. In Ramesh Hariharan, Madhavan Mukund, and V Vinay, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 2 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 49–60, Dagstuhl, Germany, 2008. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[11] Catriel Beeri, Anat Eyal, Tova Milo, and Alon Pilberg. Monitoring business processes with queries. In *VLDB*, 2007.

[12] M. Bennett. Fibo: Best practice in big data. *J Bank Regul*, 14(3-4):255–268, Jul 2013.

[13] Harold Boley, Said Tabet, and Gerd Wagner. Design Rationale for RuleML: A Markup Language for Semantic Web Rules. In *SWWS*, 2001.

[14] Federico Chesani, Paola Mello, Marco Montali, Fabrizio Riguzzi, Maurizio Sebastianis, and Sergio Storari. Checking compliance of execution traces to business rules. In Danilo Ardagna, Massimo Mecella, and Jian Yang, editors, *Business Process Management Workshops*, volume 17 of *Lecture Notes in Business Information Processing*, pages 134–145. Springer Berlin Heidelberg, 2009.

[15] EDM Council and OMG-FDTF. Financial industry business ontology (fibo). Technical report, 2013.

[16] Gero Decker, Hagen Overdick, and Mathias Weske. Oryx - an open modeling platform for the BPM community. In *Business Process Management, 6th International Conference, BPM 2008, Milan, Italy, September 2-4, 2008. Proceedings*, volume 5240 of *Lecture Notes in Computer Science*, pages 382–385. Springer, 2008.

[17] Patrick Delfmann, Sebastian Herwig, Lukasz Lis, Armin Stein, Katrin Tent, and Jrg Becker. Pattern specification and matching in conceptual models - a generic approach based on set operations. *Enterprise Modelling and Information Systems Architectures*, 5(3):24–43, 2010.

[18] Chiara Di Francescomarino, Chiara Ghidini, Marco Rospocher, Luciano Serafini, and Paolo Tonella. Reasoning on semantically annotated processes. In Athman Bouguettaya, Ingolf Krueger, and Tiziana Margaria, editors, *Service-Oriented Computing ? ICSOC 2008*, volume 5364 of *Lecture Notes in Computer Science*, pages 132–146. Springer Berlin Heidelberg, 2008.

[19] Matthew B. Dwyer, George S. Avrunin, and James C. Corbett. Patterns in property specifications for finite-state verification. In *ICSE*, 1999.

[20] A. Elgammal and T. Butler. Towards a framework for semantically-enabled compliance management in financial services. In *International workshop on Knowledge-Aware Service-Oriented Applications, ICSOC2014 workshops*, 2014.

[21] Amal Elgammal and Tom Butler. Towards a framework for semantically-enabled compliance management in financial services. In *1st International Workshop on Knowledge Aware Service Oriented Applications (KASA?15), co-located with ICSOC 2015*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014.

[22] Amal Elgammal, Oktay Turetken, Willem-Jan van den Heuvel, and Mike Papazoglou. Formalizing and applying compliance patterns for business process compliance. *Software and Systems Modeling*, pages 1–28, 2014.

[23] Opher Etzion and Peter Niblett. *Event processing in action*. Manning, 2010.

[24] Christopher Giblin, Samuel Mueller, and Birgit Pfitzmann. From regulatory policies to event monitoring rules: Towards model-driven compliance automation, 2006.

[25] Object Management Group. Business process model and notation specification 2.0.2. Technical report, 2013.

[26] Sylvain Hall and Roger Villemaire. XML Methods for Validation of Temporal Properties on Message Traces with Data. In *OTM*, 2008.

[27] Sylvain Hallé and Roger Villemaire. Runtime monitoring of message-based workflows with data. In *EDOC*, 2008.

[28] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design science in information systems research. *MIS Q.*, 28(1):75–105, March 2004.

[29] Stefan Kühne, Heiko Kern, Volker Gruhn, and Ralf Laue. Business process modeling with continuous validation. *Journal of Software Evolution and Process*, 22(7):547–566, 2010.

[30] Barbara Staudt Lerner, Stefan Christov, Leon J. Osterweil, Reda Bendraou, Udo Kannengiesser, and Alexander Wise. Exception Handling Patterns in Process-Aware Information Systems. *IEEE TSE*, 36(2), 2010.

[31] David Luckham. *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley, 2002.

[32] Linh Thao Ly, Fabrizio Maria Maggi, Marco Montali, Stefanie Rinderle-Ma, and W M P Van Der Aalst. A Framework for the Systematic Comparison and Evaluation of Compliance Monitoring Approaches. In *EDOC*, 2013.

[33] Linh Thao Ly, Stefanie Rinderle-Ma, David Knuplesch, and Peter Dadam. Monitoring Business Process Compliance Using Compliance Rule Graphs. In *OTM*, 2011.

[34] Fabrizio Maria Maggi, Marco Montali, Michael Westergaard, and W M P Van Der Aalst. Monitoring Business Constraints with Linear Temporal Logic: An Approach Based on Colored Automata. In *BPM*, 2011.

[35] FabrizioMaria Maggi, Michael Westergaard, Marco Montali, and WilM.P. van der Aalst. Runtime verification of ltl-based declarative process models. In Sarfraz Khurshid and Koushik Sen, editors, *Runtime Verification*, volume 7186 of *Lecture Notes in Computer Science*, pages 131–146. Springer Berlin Heidelberg, 2012.

[36] Khaled Mahbub and George Spanoudakis. A framework for requirements monitoring of service based systems. In *ICSOC*, 2004.

[37] Vuk Mijovic and Sanja Vranes. A survey and Evaluation of CEP Tools. In *YUINFO*, 2011.

[38] Marco Montali, Fabrizio Maria Maggi, Federico Chesani, Paola Mello, and Wil M. P. van der Aalst. Monitoring business constraints with the event calculus. 2013.

[39] Emmanuel Mulo, Uwe Zdun, and Schahram Dustdar. Monitoring web service event trails for business compliance. In *SOCA*, pages 1–8. IEEE, 2009.

[40] Emmanuel Mulo, Uwe Zdun, and Schahram Dustdar. Domain-specific language for event-based compliance monitoring in process-driven SOAs. *Service Oriented Computing and Applications*, 7(1), 2013.

[41] Kioumars Namiri and Nenad Stojanovic. Pattern-based design and validation of business process compliance. In *Proceedings of the 2007 OTM Confederated International Conference on On the Move to Meaningful Internet Systems: CoopIS, DOA, ODBASE, GADA, and IS - Volume Part I*, OTM'07, pages 59–76, Berlin, Heidelberg, 2007. Springer-Verlag.

[42] N.C. Narendra, V.K. Varshney, S. Nagar, M. Vasa, and A. Bhamidipaty. Optimal control point selection for continuous business process compliance monitoring. In *Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on*, volume 2, pages 2536–2541, Oct 2008.

[43] OASIS. Web services business process execution language version 2.0. Technical report, 2007.

[44] FinCEN-United States Department of the Treasury. Usa patriot act, 2001.

[45] OMG. Semantics of business vocabulary and business rules (sbvr), version 1.0, 2008.

[46] M Papazoglou. Web services and business transactions. *World Wide Web*, 6(1):49–91, 2003.

[47] Maja Pesic, Helen Schonenberg, and Wil M. P. van der Aalst. Declare: Full support for loosely-structured processes. In *EDOC*, 2007.

[48] P. Reuter and E.M. Truman. *Chasing Dirty Money: The Fight Against Money Laundering*. 2004.

[49] Stefanie Rinderle, Manfred Reichert, and Peter Dadam. Flexible support of team processes by adaptive workflow systems. In *Distributed and Parallel Databases*, pages 91–116, 2004.

[50] Nick Russell, W M P Van Der Aalst, Arthur H. M. ter Hofstede, and David Edmond. Workflow Resource Patterns: Identification, Representation and Tool Support. In *CAiSE*, 2005.

[51] Shazia Sadiq, Guido Governatori, and Kioumars Namiri. Modeling control objectives for business process compliance. In Gustavo Alonso, Peter Dadam, and Michael Rosemann, editors, *Business Process Management*, volume 4714 of *Lecture Notes in Computer Science*, pages 149–164. Springer Berlin Heidelberg, 2007.

[52] Sherif Sakr and Ahmed Awad. A framework for querying graph-based business process models. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 1297–1300, New York, NY, USA, 2010. ACM.

[53] Sherif Sakr, Ahmed Awad, and Matthias Kunze. Querying Process Models Repositories by Aggregated Graph Search. In *Business Process Management Workshops*, 2012.

[54] Samir Sebahi and Mohand-Said Hacid. Business process monitoring with bpath - (short paper). In *OTM Conferences (1)*, 2010.

[55] Financial Action Task. The fatf recommendations, 2012.

[56] R. Thullner, S. Rozsnyai, J. Schiefer, H. Obweger, and M. Suntinger. Proactive business process compliance monitoring with event-based systems. In *Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2011 15th IEEE International*, pages 429–437, Aug 2011.

[57] Robert Thullner, Szabolcs Rozsnyai, Josef Schiefer, Hannes Obweger, and Martin Suntinger. Proactive business process compliance monitoring with event-based systems. In *EDOC Workshops*, 2011.

[58] O. Turetken, A. Elgammal, W. van den Heuvel, and M. Papazoglou. Capturing compliance requirements: A pattern-based approach. *IEEE Software, special issue on Software Engineering for Compliance*, 29(3):28–36, 2012.

[59] Oktay Türetken, Amal Elgammal, Willem-Jan van den Heuvel, and Michael P. Papazoglou. Capturing compliance requirements: A pattern-based approach. *IEEE Software*, 29(3), 2012.

[60] Oktay Turetken, Amal Elgammal, Willem-Jan van den Heuvel, and Mike Papazoglou. ENFORCING COMPLIANCE ON BUSINESS PROCESSES. In *ECIS 2011 PROCEEDINGS*, 2011.

[61] Marcus Venzke. Specifications using xquery expressions on traces. *Electron. Notes Theor. Comput. Sci.*, 105:109–118, December 2004.

[62] W3C. Xml path language (xpath) 2.0 (second edition), 2011.

[63] W3C. Owl 2 web ontology language structural specification and functional-style syntax (second edition). Technical report, December 2012.

[64] M Weidlich, H Ziekow, and J Mendling. Event-based Monitoring of Process Execution Violations. In *BPM*, 2011.

Comprehensive Survey on Dynamic Graph Models

Aya Zaki
Faculty of Computer
and Information Science
Ain Shams University
Cairo, Egypt

Mahmoud Attia
Faculty of Computer
and Information Science
Ain Shams University
Cairo, Egypt

Doaa Hegazy
and Safaa Amin
Faculty of Computer
and Information Science
Ain Shams University
Cairo, Egypt

***Abstract*—Most of the critical real-world networks are continuously changing and evolving with time. Motivated by the growing importance and widespread impact of this type of networks, the dynamic nature of these networks have gained a lot of attention. Because of their intrinsic and special characteristics, these networks are best represented by dynamic graph models. To cope with their evolving nature, the representation model must keep the historical information of the network along with its temporal time. Storing such amount of data, poses many problems from the perspective of dynamic graph data management. This survey provides an in-depth overview on dynamic graph related problems. Novel categorization and classification of the state of the art dynamic graph models are also presented in a systematic and comprehensive way. Finally, we discuss dynamic graph processing including the output representation of its algorithms.**

Keywords*—*dynamic graphs; evolving networks; evolving graphs; temporal graphs; data management

I. Introduction

Most real-world networks like social networks [1], [2], [3], wireless networks [4], transportation networks [5], and other networks contain a vast amount of information. These networks are mostly represented by graphs. These graphs model the network entities and their relations in the form of vertices and edges respectively. The majority of current network graph representations rely on static graphs. Such type of graphs fails to handle the real-time changes of networks. That's why there is a significant interest in providing a dynamic graph model that stores the network historical changes and gives the ability to query these changes [6].

Temporal Relational Database ”TRD” shares the power of storing historical changes with dynamic graphs. A lot of literature on TRD focus on temporal data model and temporal query language [7], [8], [9], [10], [11], [12], [13]. The two main basic concepts of TRD are valid time and transaction time. Valid time represents the time period that indicates when the fact is true in the real world. Transaction time represents the time period of storing and removing the fact from the DB.

In this survey, we focus on valid time, where the goal is to retrieve the graph entities that are valid at any given time instant. We start off by reviewing the main issues of dynamic graphs: temporal evolution and dynamic graph queries, as well as their related terminologies. Then, we present an overview to categorize the existing dynamic graph models proposed by other researchers. Finally, we provide a brief survey on processing including both: dynamic graph algorithms and output representation. Fig 1 summarizes our proposed classification of the accomplished work in dynamic graphs.

The rest of this paper is organized as follows: Section 2 provides an overview about how dynamic graphs evolve with time. In section 3, we present a study of dynamic graph queries. Section 4 categorizes the existing dynamic graph models. Section 5 states the output representations of dynamic graph algorithms and overviews some of the most important graph problems. Finally, Section 6 concludes this paper and discusses the future research plans.

II. Temporal Evolution

Temporal evolution shows how dynamic graphs evolve with time and the changes that happen to its components. This type of evolution can be categorized into two categories: topological evolution and attributes evolution, which will be discussed in details in the following subsections II-A and II-B.

A. Topological Evolution

A dynamic graph undergoes continuous changes with time in its structural components: nodes and edges. Such changes are called topological evolution. Due to the topological evolution, the graph structure is reshaped based on the following:

- Edge Evolution: dynamic graph changes related to its edges only. These changes can be represented by the actions: add edge or remove edge. In some cases, the evolution of a network is modeled by edge evolution only, where the network nodes are constant over time [14].
- Node Evolution: dynamic graph changes related to its vertices only. These changes can be represented by the actions: add node or remove node. Contrary to edge evolution, there are no cases where a networks evolution is modeled by node evolution only.

Most networks that have topological evolution, involve both edge evolution and node evolution [13]. In these networks, when a remove node action occurs, both the out and in edges of the removed node are implicitly removed from the graph before the removal of the node itself.

B. Attributes Evolution

Dynamic graph continuous changes occur on graphs attributes (i.e., internal attributes of nodes and edges) and do not affect the graph topology.

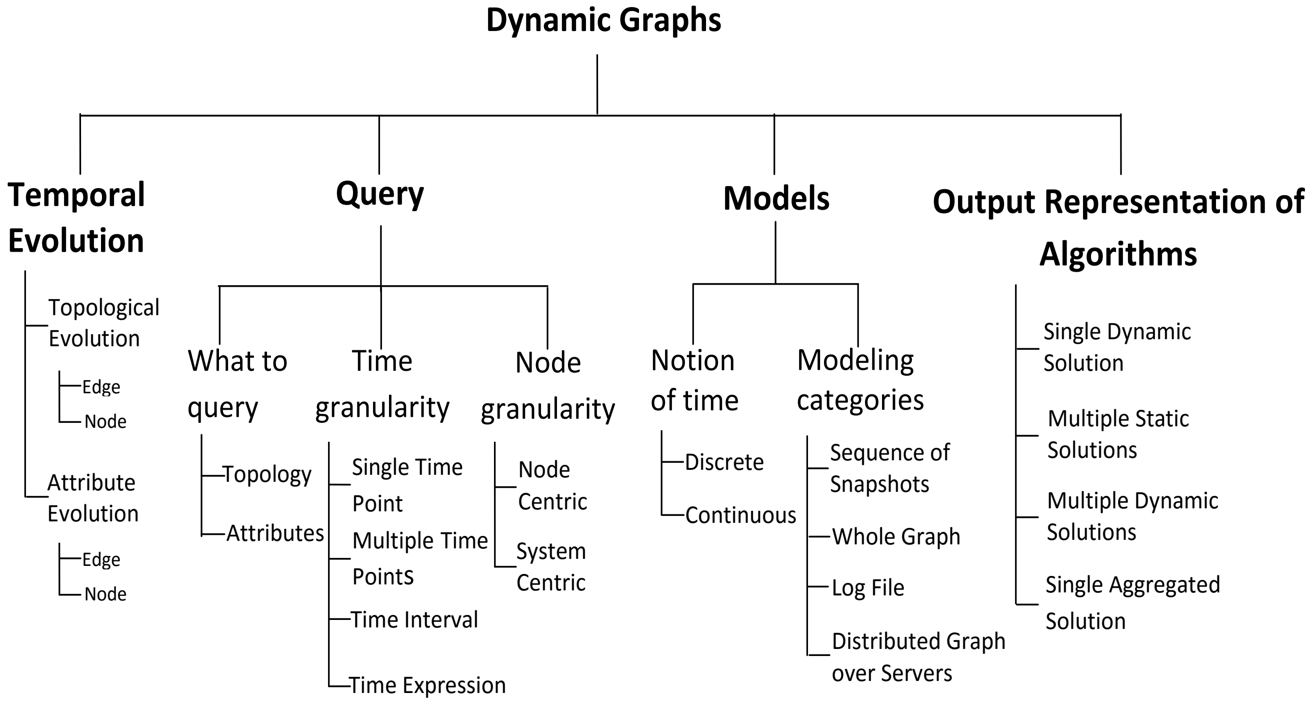


Fig. 1: Dynamic Graphs

- Edge Attributes Evolution: dynamic graph changes related to its edges attributes. These changes can be represented by the actions: add attribute value, remove attribute value or update attribute value [15].
- Node Attributes Evolution: dynamic graph changes related to its node attributes. These changes can be represented by actions similar to edge attributes evolution [16].

According to this classification, the evolving graphs have two types: fully evolved dynamic graphs and partially evolved dynamic graphs. Fully evolved dynamic graphs have both topological and attributes evolution in edges and nodes. In contrast, partially evolved dynamic graphs have a partial mix of them.

III. QUERY

In this section, the query operators that have been proposed in the literature of dynamic graphs are presented. We provide a novel classification based on three criterions that will be detailed in the following subsections.

A. *What to query*

According to the interest of applications, we classify the query functionality on dynamic graphs into two classes:

- Topology: includes queries that ask about historical graph structure. In other words, queries are asking about nodes and edges of the graph at a previous time. For example, querying a node neighbors ”retrieve friends of Jon at Oct 10, 2000” [17].
- Attributes: in this class, queries are used to retrieve a graph components attributes by asking about nodes attributes or edges attributes at a previous time. For example, querying the sent packets from node A to node B ”retrieve packets that are sent from A to B at Oct 10, 2014” [15].

B. *Time granularity*

Dynamic graph queries differ from static graph queries because of the inclusion of the time dimension. The time dimension has several forms, which are classified into four types as follows:

- Single Time Point: where query is used to ask about graph historical information valid at a specific time point. For example ”retrieve the graph structure at Oct 10, 2000 [18].
- Multiple Time Points: this query is asking about retrieving graph historical information that is valid at multiple time points. For example ”retrieve the graph structure at every monday between 2000 and 2005 or ”retrieve Jon friends at Oct 10, 2000 and Dec 24, 2003” [19].
- Time Interval: this type of query has more complex form. It is asking about graph historical information valid at an interval of time. For example ”how does the average number of Jon friends change over [2000, 2005]” [19].
- Time Expression: in which, a boolean expression is applied over a set of multiple discrete point. For example, ”retrieve Jon friends that are valid at $(2000\wedge\neg 2001)$” [19].

C. Node granularity

The third criterion that distinguishes dynamic graph queries is the node granularity classified into two types:

- Node-Centric Queries: involve one or more graph node related to a specific node. In other words, these queries need to access a part of the graph. For example, retrieve friends of Jon at Oct, 10, 2009 [17].
- System-Centric Queries: are also known as Global Queries. These queries involve all graph nodes. For example, compute the graph diameter at Oct 10, 2008 [19].

IV. Models

A dynamic graph model is a mapping $G_t = (V, E)$ that yields the state of the graph (i.e., the set of nodes and set of edges) at a given time instant *t*. Both directed and undirected dynamic graphs can be represented by most of the existing discrete and continuous models. In a discrete model, snapshots are taken periodically at every fixed time period (e.g., every 30 minutes, every day, and every week). This type of model provides complete accurate mapping at specific time instants and gives the nearest state (e.g., time-based, changes-based) at any other instant. On the other hand, the continuous model keeps track of all changes by representing every one of them. Therefore, it can map every instant into a completely accurate valid graph state. In the following subsections, we classify the relevant literature into four categories, and describe each one of them.

A. Dynamic Graph Models Categorization

The existing dynamic graph models that were proposed by other researchers can be categorized into four basic categories:

- Sequence of Snapshot: the graph historical changes are stored as a sequence of snapshots. Each snapshot represents the graph state at a single instant of time. The snapshot consists of a set of vertices V and a set of edges E. However, the existing models in this category are discrete.
- Whole Graph: the graph historical changes are stored as one large graph. The changes (i.e., vertex/edge deletion, vertex/edge insertion and their attributes updates) are applied and stored in the same graph. Moreover, each graph element (i.e., vertex, edge or attribute) is accompanied with a valid time point or valid time interval according to its model. Models of this category are either discrete or continuous.
- Log File: the latest snapshot as well as key snapshots are kept while the graph historical changes between any two consecutive key snapshots are stored in a log file. Each log file is accompanied with its valid time interval. The existing models in this category are continuous.
- Distributed Graph over Servers: dynamic graph distribution can be categorized based on two parameters:
  - Time: divides the graph historical changes according to time over a set of servers. Each server is responsible for a period of time.
  - Structure: divides the graph structure (e.g., vertices distribution) over a set of servers. Each server is responsible for managing its vertices by storing and retrieving their historical changes.

  This survey discusses distributed graph over servers from a structural perspective. Thats because the distribution based on time can be merged with any of the other categories as an improvement, without affecting the management (i.e., storing and retrieving) of the used model of the merged category. In contrast, distribution based on structure affects the management of the used model of the merged category.

B. Sequence Of Snapshots

Varieties of this category have been proposed in Rossi's model[20], FVF[21], and Yang's model[14]. This category models dynamic graphs as a sequence of snapshot $G_{[t_1,t_n]}$= $\{G_1, G_2, G_3, ..., G_n\}$. Each snapshot G_i is a static graph that represents the valid state of the dynamic graph at time point t_i. The snapshot is represented by a triple $< V_i, E_i, t_i >$ and is stored by its time point t_i.

Storing sequence of snapshots naively as in Yang's model[14], and Rossi's model [20] would clearly require a prohibitively large storage. FVF[21] proposes Find Verify and Fix "FVF" framework which takes sequence of snapshots that are produced in a compressed storage model as input. This compressed storage model stores a set of key snapshots and the associated set of deltas. The set of key snapshots is intended to be much smaller than the original set of all snapshots. A set of deltas stores only changes that are needed to completely construct a snapshot from its related key snapshot by merging the key snapshot with the proper delta.

In FVF, the sequence of snapshots is divided into clusters based on similarities among them. Each cluster has two representative graphs ($G_\cup$ and $G_\cap$), where $G_\cap$ is the largest common sub-graph of all snapshots in the same cluster, and $G_\cup$ is the union of the smallest sub-graphs of all snapshots in the same cluster. FVF needs to access the two representative graphs of the G_i's cluster as well as the snapshot G_i itself for answering a query (e.g., node-centric) on G_i snapshot. Compressed Storage Models "SM" have been discussed in FVF[21] for storing these clusters. However, the most efficient one is called SM-FVF. It saves four deltas for each cluster C, which has k snapshots.

- First, D($G_\cap, G_{P\cap}$): the needed edges to be inserted or deleted from $G_{P\cap}$ to get $G_\cap$, where the $G_{P\cap}$ is the $G_\cap$ of the previous cluster.
- Second, $\triangle(G_\cup, G_\cap)$: the set of edges that exists in $G_\cup$ and does not exist in $G_\cap$.
- Third, $\triangle(G_1, G_\cap)$: the set of edges that exists in G_1 and does not exist in $G_\cap$.
- Finally, D(G_i, G_{i-1}), $\forall\ 2 < i < k$: the needed edges to be inserted or deleted from G_{i-1} to get G_i.

Resulted storage of one cluster can be computed as in (1).

$$SM-FVF(C) = \{D(G_\cap, G_{P\cap}), \triangle(G_\cup, G_\cap), \triangle(G_1, G_\cap), D(G_i, G_{i-1}), \forall 2 < i < k\}[21] \quad (1)$$

The SM-FVF compressed model has no redundancy among delta files of the same cluster, but it has redundancy among delta files of different clusters. This redundancy is ignored relatively to the naive model which has a lot of redundant data among its sequence of snapshots. A downside of this compressed model is that it needs to access $[G_{1\cap}, G_{(i-1)\cap}]$ for constructing $G_{\cap_i}$. Therefore, it is suitable for queries that need to be applied over the whole sequence of snapshots. We suggest replacing the delta file D$(G_{\cap}, G_{P\cap})$ of each cluster i with a delta file that contains $G_{i\cap}$ edges and vertices for improving queries that need to access only one snapshot or few snapshots.

Evaluating a snapshot G_t is straight forward procedure in sequence of snapshots category. It is done by either finding the exact time point *t* or the nearest time point to it in order to return its associated pair $< V, E >$ in the naive model or by constructing it in the compressed model.

In conclusion, the compressed model in FVF[21] is more efficient with regard to the used storage than the naive model described by Yang[14], and Rossi[20]. However, the naive model is faster than the compressed model in query performance due to the consumed construction time in the compressed model.

C. Whole Graph

In this category, dynamic graphs is modeled as one large graph $G_{[t_1,t_n]} = < V_{[t_1,t_n]}, E_{[t_1,t_n]} >$, where $V_{[t_1,t_n]}$ and $E_{[t_1,t_n]}$ are set of all vertices instances and edges instances respectively. Alternatives of this category have been proposed in koloniari's model1 [22], Huo's model[23], [24], TPM[25], FSDNs[15], and Evo-graph[16].

Dynamic graphs in this category can be basically represented as a two sets methodology: vertices set and edges set as models presented by koloniari in [22], and Huo in[23], [24]. Each element of the edges set and the vertices set can be represented as a triplet $< srcID, desID, [t_s, t_e] >$ and $< vID, \{att\}, [t_s, t_e] >$ respectively, where the accompanied interval $[t_s, t_e]$ represents only one valid interval. This methodology has downsides in its storage by storing the same element multiple times with different valid time intervals which happens in case of an element existence and re-existence. Alternatively, the temporal provenance model "TPM" [25] provides another representation of the two set methodology. Each vertex of the vertices set can be either entity instance with a triplet $< vID, \{att\}, t_s >$ or timed folder/path node with a triplet $< vID, \{att\}, (t_s, d) >$. Timed folder node and timed path node are containers that are used for storing queries results. Each edge of the edges set represents a relationship between two vertices and it does not have any time notion. Fig. 2 provides more information about vertices and edges notions of the TPM (i.e., the relation types as well as the entities types). TPM suffers from a number of drawbacks regarding to storage that are concluded a follows:-

- TPM Stores redundant vertices. For example, when an entity starting connection with another entity, new two vertices are created corresponding to them with new IDs and the same data.
- Some results of queries are stored also in the same graph as Path/Folder nodes which increase the graph size.

The two set methodology drawbacks are avoided in Fixed Schedule Dynamic Networks "FSDNs"[15]. The paper proposes a data structure that allows storing only the new valid time interval deduced from the re-existence rather than restoring the whole elements. The data structure consists of set of vertices, where every vertex element is a triplet $< vID, I, \{Neighbor\} >$. The vertex contains a set of its valid time intervals I as well as a set of its neighbors. Each neighbor is a triplet $< nID, I, att >$, where the set I represents valid time intervals of this neighbor's connection. This data structure has downsides regarding to its storage by storing each edge twice (i.e., once at each end point vertex). However, in both methodologies, element intervals do not intersect.

Furthermore, there are models in the whole graph category that not only store the historical changes of dynamic graphs but also, store the types of changes themselves like Evo-graph [16]. The Evo-graph model states the vertices versions of dynamic graphs and the types of changes themselves that produced these versions. Evo-graph consists of two interconnected components: data-graph and change-graph. Data-graph comprises all vertices versions of the actual data. Each data vertex version is a pair$< vID, \{att\} >$ and is connected to another data vertex by data-graph edge. The change-graph contains the change types that produce new vertices' versions in data-graph as shown in Fig. 3. Each change-graph vertex is a triplet $< vID, changeType, timeStamp >$ and is connected to another change vertex by a change-graph edge. Evo-graph components are connected by evolution edges. An evolution edge connects two data-graph vertex versions (i.e., before and after a change operation) with a change-graph vertex (i.e., the change operation itself). The Evo-graph model suffers from a number of drawbacks, the main of such drawbacks can be summarized as follows:-

- It is not applicable for all graph networks due to its structure
- Regarding to storage, the same vertex is stored several times using any type of change except the update type as shown in Fig. 3.

Retrieving a snapshot G_t in the whole graph category is costly because the search space is the whole graph. The snapshot is evaluated by traversing the whole graph elements to capture its valid elements only at time *t*. The valid element at time *t* is an element whose time point is *t* or its time interval contains *t*. Huo's model [23], [24] proposes a temporal partitioning for improving time point query performance. However, the node-centric query is efficient in FSDNs[15] due to its data structure, where a node is selected then the search is expanded to its neighbors list.

In temporal partitioning, the whole graph time interval is distributed over partitions. For example, given n time instants and the fact that each partition can have m time instants, then [n/m] partitions will be generated. Edges and vertices that have overlapped time intervals over more than one partition will be duplicated in these partitions. While this duplication allows time point query to access one partition rather than accessing the whole graph, it increases the TPM storage size. On the other hand, this portioning makes the small time interval query access two partitions in the worst case. Therefore, Huo in [24] applies overlapped partitioning for improving the performance of the small interval query and reducing the number of accessed partitions. However, the overlapped partitioning consumes more storage (i.e., 50% overlapping will lead to 100% redundancy).

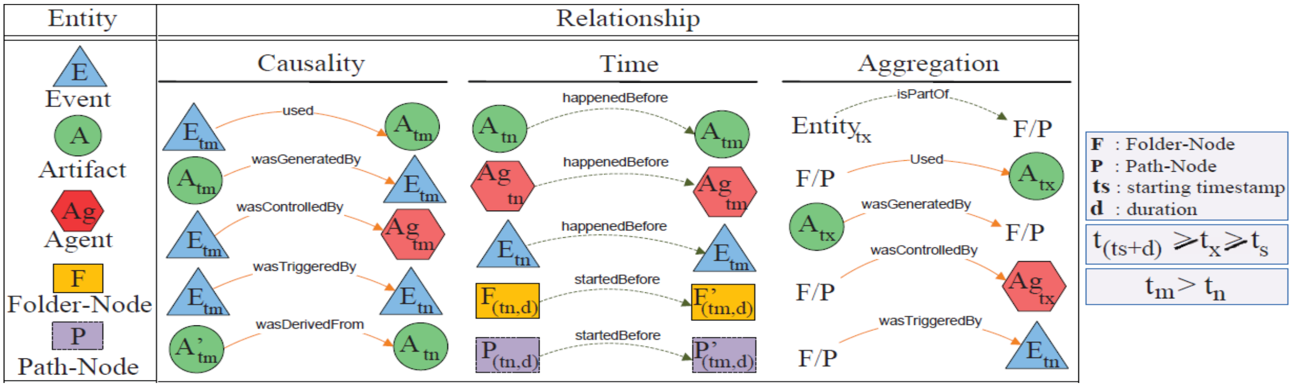


Fig. 2: Temporal Provenance Model notions [25]

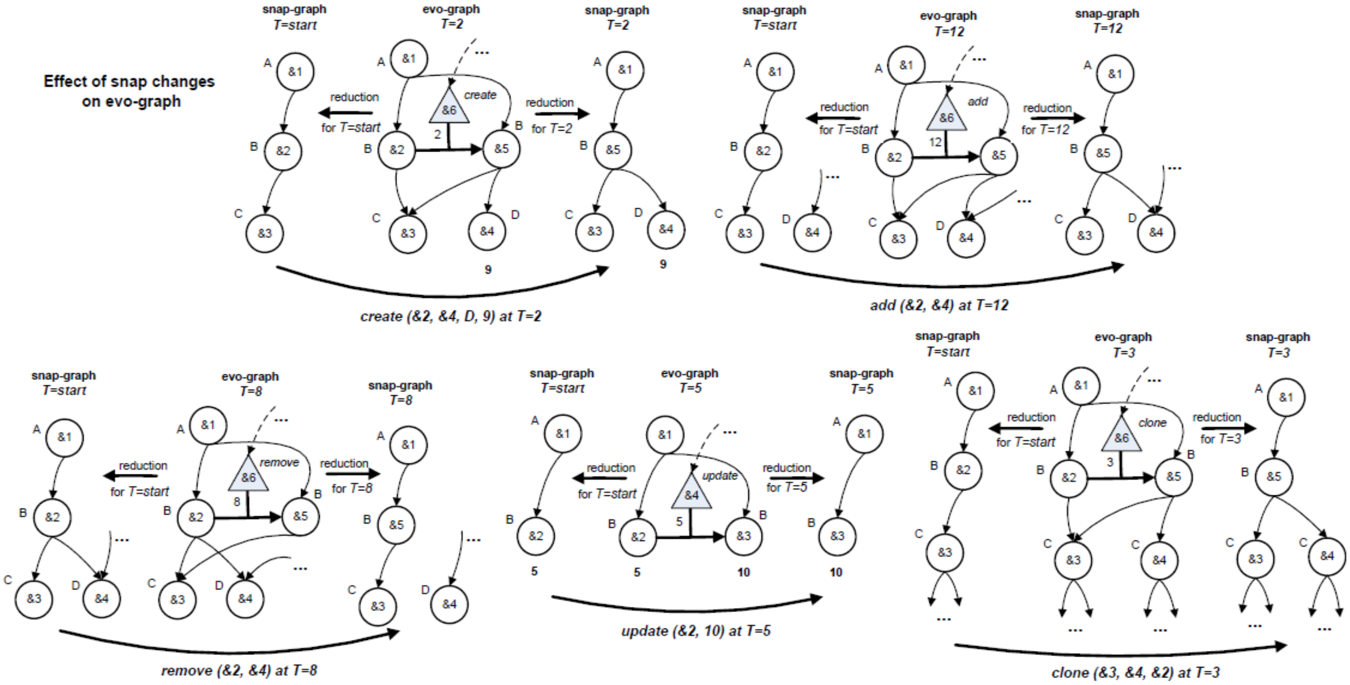


Fig. 3: Effect of snap change operations on the Evo-graph [16]

In conclusion, it is found that FSDNs[15] is the most efficient model in the used storage but, Huo's model [23], [24] is the most efficient model in query performance due to the overlapped partitioning.

D. Log File

The main idea of this category is based on materialization, which is a process of storing a set of dynamic graph snapshots and delta log files between them. Varieties of this category have been presented in Koloniari's model2[19], [17], Khurana's model [18], and Chronos[26]. This category models dynamic graphs as $G_{[t_1,t_{n=crr}]} = < G_{t_{crr}}, L, M >$, where $G_{t_{crr}}$ is a snapshot that represents the current dynamic graph state, L is a set of log files $L = \{L_{1[t_{s_1},t_{e_1}]}, L_{2[t_{s_2},t_{e_2}]}, ..., L_{n[t_{s_n},t_{e_n}]}\}$, each representing the historical changes, which occurred during its associated time interval $[t_{s_i}, t_{e_i}]$ and M is set of materialized snapshots, each representing dynamic graph state at the beginning of a log file.

Materializing snapshots has three types [19]:

- Time-based: the duration between any two consecutive materialized snapshots is constant.
- Operation-based: the number of historical changes of any log file is constant [18].
- Similarity-based: similarities between any two consecutives snapshots do not exceed a threshold value [19] [26].

The details of each type are summarized in TABLE. I showing the advantages and the disadvantages of each of them.

The vital issues of the log-file category are: storing the materialized snapshots and structuring the log files. Firstly, materializing snapshots raises a problem of how the model can efficiently store them considering their growing number. Chronos[26] stores each materialized snapshot as a block at the beginning of its corresponding log file. Koloniari [19],

TABLE I: Materialization types comparison

	Time-based	Operation-based	Similarities-based
Advantage	The overhead for deciding a new snapshot materialization is minimal.	The overhead for deciding a new snapshot materialization is minimal.	- The threshold value is defined by the user so, the materialized snapshots redundancy ratio is acceptable by the user. - It balances between the redundancy ratio and the log file size.
Disadvantage	When the changes of a dynamic network do not occur uniformly, time periods that have many changes produce a very big log file size and time periods that have few changes produce a very small log file size. The big file size leads to an increase in the construction cost.	Big redundancy between materialized snapshots for example [19]:- - When a file contains changes and their reverse, the two bounded snapshots have a lot of redundancy. - When a file contains changes of specific graph elements (edges, vertices), the two bounded snapshots have a lot of redundancy.	High overhead for deciding a new snapshot materialization because of computing snapshots similarities periodically.

[17] provides the naive solution of storing the materialized snapshots as a sequence of snapshots. The two mentioned strategies materialize snapshots commonalities in a redundant manner as they store the whole materialized snapshot every time. This redundancy can be avoided using the graph pool component proposed in Khuranas model[18].

The Graph-pool component has two parts: overlaid snapshots and Graph-ID bit mapping. Graph-pool stores all snapshots elements (e.g., edges, vertices and attributes of them) in one large graph, in a compact manner. It stores each different element only once associated with a mapping string that maps the element to its related active snapshots including both materialized snapshots and retrieved historical snapshots. For processing a snapshot of the graph-pool, Graph-ID bitmapping stores for each overlaid snapshot s the following information:-

- Snapshot ID.
- Bits' indices: they provide the indices of bits that represent the s snapshot at all elements mapping strings. For example, each materialized snapshot is represented by 1 bit in the mapping string to decide if the element of the mapping string is related to the materialized snapshot or not.
- Dep-ID: the overlaid historical snapshot marks the used materialized snapshot in construction as dependent. That happens when the size of the commonality between the historical snapshot s and the used materialized snapshot is large relative to the materialized snapshot size. This prevents traversing the whole graph-pool elements for setting the corresponding bits of the historical snapshot.

While this compression eliminates the stored materialized snapshots redundancy by storing each different element only once, it leads to processing overhead on the graph-pool. The overhead comes from traversing the whole graph-pool elements for overlaying a new snapshot or deleting an existing one. For example, when a snapshot is pulled to memory, it is overlaid on the graph-pool edge by edge and node by node (i.e., it traverses the whole graph-pool elements to fill its related bits in the mapping string). The graph-pool periodically deletes the unnecessary snapshots when there is no query load or when memory is needed for a better storage management. The snapshot deletion is accomplished by traversing the whole graph-pool elements to reset its corresponding bits' indices of each element's mapping string. A graph-pool element is removed, when its mapping string contains only one snapshot that will be deleted. Removing those unnecessary elements of the graph-pool decreases the used storage.

Secondly, it is important to examine the log file structure to decrease the construction time, since the log files store the graph historical changes and they are used in snapshots reconstruction. A log file can be structured as storing all graph historical changes according to their occurrence time as in Koloniaris model [19]. This needs much construction time for traversing all changes of the log file till the target time. Alternatively, Khurana [18] proposed a delta-graph component that isolates topological evolution from attributes evolution as well as any other type that can be defined by the user. The isolation improves the construction performance by specifying the target evolution type. chronos [26] also isolates the log file's historical changes, but it isolates nodes evolutions from edges evolutions (e.g., edges files and vertices files). It stores the historical changes after the materialized part block in a locality layout for a better performance, which speeds up the reconstruction time of a specific vertex or edge at a particular time *t*.

The vertex file structure as well as the edge file structure is similar. For example, an edge file in Chronos [26] stores the vertices identifiers in its headers. For every vertex, the edge file stores a detail block. A detail block starts by listing all edges of its vertex at the beginning of the associated time interval. Then it stores a list of changes to this vertex edges (i.e., add edge, delete edge and update edge). This structure enforces constructing the corresponding materialized snapshot before constructing the target snapshot itself.

The Delta-graph component[18] is a directed graphical structure, which is maintained as a weighted graph and is stored in memory. It is used for managing the log files and reconstructing snapshots with the lowest possible number of historical changes. Delta-graph provides multiple hierarchies for improving the reconstruction phase; each is corresponding to an evolution type. Every hierarchy contains statistics about the log files but not the actual data [18]. However, the actual files content are stored on disk. The lowest level nodes of the delta-graph are corresponding to the materialized snapshots. The edges between the leave nodes represent the log files that are needed to construct two consecutive snapshots from each other. However, each interior node has k children and is corresponding to a graph that is constructed by applying a differential function {intersection, skewed, balanced, empty ....} over its children. The edges between the interior nodes represent the files that are needed to construct a child node from its parent node. Moreover, the highest level node (root) is not corresponding to a graph. Each hierarchy can have a different differential function. While delta-graph component improves the performance of the reconstruction phase by isolating the evolution types, it stores overhead files for the interior edges.

Khuranas model [18] suffers from two drawbacks related to storage and performance because of the overhead induced by its delta-graph, and graph-pool components respectively. First, the storage overhead is caused from saving the files of the delta-graph edges connecting between the interior nodes. These files do not store the correct graph historical changes as mentioned previously. From our study, we deduce that, the number of these files can be computed as in (2).

$$Numberoffiles = k * \Sigma_{i=1}^{(log_k n)-1} k^i \tag{2}$$

(k: delta-graph arity, n: # of nodes of the lowest level) $\log_k n$ is the delta-graph levels number and $\Sigma_{i=1}^{(\log_k n)-1} k^i$ is the number of the delta-graph interior nodes.

Furthermore, the graph-pool stores at most k snapshots for every delta-graph level during constructing it (i.e., when a delta-graph level completes its k snapshots, it constructs their parent. Then graph-pool deletes them).Also each graph-pool element stores at most two overhead bits for every unrelated snapshot in its mapping string. Second, the models performance is affected by the processing over head required by the graph-pool to traverse all the graph-pool elements in order to overlay or delete a single snapshot, even though, some of the graph-pool elements may not exist in the snapshot to be overlaid or deleted. From our study, we deduce that, there is a relation between the storage and performance of khuranas model as presented in TABLE. II.

TABLE II: Storage and performance behaviors based on Parameter

Prameter	Storage	performance
- Log file size (L) increased ↑	graph-pool size decreased ↓	snapshot construction cost increased ↑
- delta-graph arity (k) increased ↑	graph-pool size increased ↑	# of delta-graph level decreased ↓ and snapshot construction performance decreased ↓
- # of isolated delta-graphs increased ↑	delta-graph size increased ↑	snapshot construction performance decreased ↓

Sometimes the retrieved historical snapshots need to be stored for analytical processing or processing in general. These snapshots can be stored in a compact manner like Khurana's model [18] in which, they are saved in graph-pool with the materialized snapshots. This needs much time for processing as mentioned before. Chronos [26] proposed a model for storing the retrieved snapshots, which pays attention to time locality. This enhances snapshots processing performance. The model consists of two arrays: vertex-array, and edge-array. The vertex-array groups the vertex's versions across all the snapshots placing multiple versions of the same vertex one after the other. The edge-array groups the edges with source vertex or destination vertex. Every edge-array element contains {edge Id, target-node Id, a mapping string: which contains one bit for each snapshot as a flag, and the corresponding weight of each related snapshot (i.e., w_{ij}^0 represents the weight of edge e_{ij} at snapshot 0)}. While this structure decreases the number of cash miss during processing, which improves processing performance, it has redundancy among every vertex versions in the vertex-array.

In Chronos [26] the log file structure as well as the proposed model of the retrieved historical snapshots is compatible in locality layout design. To increase locality benefit, Chronos proposed a Locality-Aware Batch Scheduling "LABS" that makes processing execution aligned with the underlying layout design [26]. LABS enables accessing the edge-array once for all snapshots rather than accessing N times one for each snapshot. LABS improves the parallelization and incremental computations performance, for more details check the following paper [26].

Evaluating a snapshot G_t in the log file category is accomplished by construction using the corresponding log file and the corresponding materialized snapshot, which is very costly. In Chronos model[26], constructing the corresponding materialized snapshot is required before constructing the historical snapshot itself. However, in the materialized graph sequences of Koloniari's model2 [19], the materialized snapshot is retrieved, and then the needed snapshot is constructed directly. Moreover, Koloniari provides partial reconstruction in [17] for constructing the target sub-graph rather than constructing the whole snapshot in case of node-centric query, which improves the node-centric query performance. In [18], it uses delta-graph component to construct the target snapshot as follows:

- A temporary node is created between the corresponding two nodes of the selected materialized snapshots.
- The corresponding edges log files of the created node are estimated.
- Dijkstra's shortest path algorithm is applied on delta-graph.
- Finally, the historical changes of the resulted path with the lowest cost from the root to the temporary node is used to reconstruct the target snapshot. This path contains the minimum number of changes that are needed to construct the target snapshot.

The edge cost represents its file size, which depends on the query evolution type (e.g., topological, node attributes,... etc). For retrieving more than one snapshot, rather than applying Dijkstras shortest path algorithm multiple times for finding the path with the lowest cost of every target snapshot, it computes the lowest-weight steiner tree that connects the root and the added temporary nodes.

As a conclusion, it is found that Khurana's model[18] is more efficient than the other two models regarding to the used storage of the materialized(or historical) snapshots in memory. However, Chronos's model [26] is faster than the other two models regarding to processing phase over the retrieved historical snapshots because of LABS. On the other hand, Khurana's model [18] is the fastest in the snapshot reconstruction phase.

E. Distributed Graph over Servers

Dynamic graph vertices in this category are divided over a set of servers S. Each server is responsible for managing (i.e., storing and retrieving) its vertices historical changes. Varieties of this category have been proposed in G^*[6], MG^*[27], and Kineograph[28], where the vertices are assigned to servers based on their hash-value.

The two main issues of the distributed graph over servers category are: the way of assigning vertices to servers, and the way of storing historical changes at each server. Primarily, the current existing criterion that controls the decision of assigning vertices to servers raises servers communication problem. So, it is important to consider a new vertices distribution technique, which decreases assigning connected vertices pairs to different servers.

The second issue of this category is representing the historical changes in a server efficiently taking into account the continuous growing. The historical changes of a server can be represented by any of the previously mentioned categories. For example, every server can represent its historical changes as a sequence of snapshots like Kineograph [28]. A snapshot consists of a set of vertices and provides topological evolution only, where each vertex is accompanied with an adjacent list that representing the vertex outing edges. Storage wise, Kineograph is not efficient as it stores a lot of redundancy among the stored sequence of snapshots. This redundancy is avoided in G^* [6]. G^* provides compact graph index "CGI", which is compressed data structure for storing the historical changes of the server's sub-graph as a sequence of snapshots in a compact manner. Each snapshot in G^* model is represented as a triplet $< Id, \{att\}, vertexSet >$. The compression comes from storing every vertex version only once. G^* represents a vertex version as a vertex location pair "VL-pair" $< vertexID, Disk\_Location >$ that map each vertex Id to the vertex version location on-disk, where the vertex version actual data is stored. To accomplish this compression idea, G^* combines every common set of vertices versions over snapshots in one map index. After that, it associates every map to its relevant snapshots. This architecture enables computations sharing across snapshots, which accelerates query processing. For example, when processing is applied on a vertex version, the obtained results are shared with all the other snapshots containing this version.

However storing data on disk requires many disk access, G^* proposes a schema for the on-disk data to minimize the number of disk access by grouping the vertex and its outing edges as one unit. Therefore, loading and storing a vertex and its outing edges occur at the same time. Every vertex version is stored as V(graph.Id, Id, att_1, att_2,..., edgeSet), where edgeSet are the vertex outing edges. Moreover, every edge schema is E(graph.Id, vertex.Id, des.Id, att_1, att_2,...). one more gain of this schema is that it enables a vertex versions to share their attributes commonalities as mentioned in G^* [6].

Due to the continuous increase in snapshots number, G^* has update time overhead in CGI. The overhead comes from finding the commonalities between the newly added snapshot and the stored snapshots to keep storing each vertex version only once. G^* proposes a split CGI as an enhancement by splitting the CGI to a set of CGIs using a threshold value of the maximum update time. In the split CGI, each CGI is responsible for a set of snapshots. Therefore, the split CGI has redundancy among CGIs. CGI is better than the split CGI in the used storage. On the other hand, the split CGI is better than CGI in the performance.

MG^*, a modified model of G^*, efficiently solved the update overhead problem [27]. MG^* represents the historical changes of every server as represented in the log file category. Every server materialized snapshot in MG^* is represented by the CGI data structure and the vertices actual data are stored on disk using the same schema of G^* [6]. The in-between historical changes of the materialized snapshots are stored as log files with fixed size for bounding the construction time. The MG^* stores the historical changes in a temporary list, to reduce the number of writing them to disk from O to one time instead (i.e., O is number of events that will be stored in the log file). According to such modification, in the update process, the MG* only appends the historical changes to the temporary list, consuming almost no time in comparison to G^*. Moreover, the total used memory of MG^* is better than that of G^*, where the difference is order of magnitude. That is because MG^* transfers a huge part of the historical changes from memory to disk in the form of log files.

A snapshot G_t is evaluated in the distributed graph over server category by instructing each server to retrieve its part at time point *t*, then the master server returns the aggregation of the retrieved parts as the final results. Even though, models in this category have the same flow of retrieving a snapshot, they have different performance due to the different ways of representing the historical changes in the corresponding servers. For example, Kineograph has the best performance in this category due to its naive structure. On the other hand, MG^* has the worst performance due to the snapshot construction phase. However, MG^* provides completely accurate snapshots, thus, it produces completely accurate results of queries as it stores all historical changes. In contrast the other two models do not guarantee complete accuracy of the retrieved snapshots.

F. Dynamic Graph Models Summary

For each previously mentioned model, we present a summary of the remaining dynamic graph properties that are considered independent from the category to which the model belongs to, as shown in TABLE III.

To provide a full overview of the discussed dynamic graph models in this survey, TABLE IV summarizes these models based on five criterions: retrieving performance of a snapshot G_t, update performance of a historical changes unit, existing redundancy, used memory storage, and used disk storage. The evaluation of each criterion is denoted by five levels: Very-Low, Low, Medium, High, and Very-High.

V. Dynamic Graph Algorithms Output Representation

Because of the addition of time parameter to dynamic graphs, algorithms on dynamic graphs are more complicated

TABLE III: Dynamic graph models properties.

	Model Type		Graph Type		What to Query		Evolution			
Dynamic graph Model	Discrete	Cont	Directed	Undirected	Topology	Attribute	Topological		Attribute	
							node	edge	node	edge
FVF model	✓		✓	✓	✓		✓	✓		
Rossi’s model	✓		✓	✓	✓		✓	✓		
Yang’s model	✓			✓	✓			✓		
koloniari’s model1		✓		✓	✓		✓	✓		
Huo’s model	✓		✓	✓	✓	✓	✓	✓		✓
FSDNs model	✓		✓		✓		✓	✓		
TPM model		✓	✓		✓	✓	✓	✓	✓	
Evo-graph model	✓		✓		✓	✓	✓	✓	✓	
Koloniari’s model2		✓	✓	✓	✓		✓	✓		
Khurana’s model		✓	✓	✓	✓	✓	✓	✓	✓	✓
Chronos model		✓	✓	✓	✓	✓	✓	✓	✓	✓
G^* model	✓		✓	✓	✓	✓	✓	✓	✓	✓
MG^* model	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Kineograph model	✓		✓	✓	✓		✓	✓		

TABLE IV: Dynamic graph models performance.

Dynamic graphs Model	Retrieve performance of G_t	Update Performance	Redundancy	Memory Storage	Disk Storage
FVF model	High	High	Low	Medium	Low
Rossi’s model	Low	Low	Very-High	High	-
Yang’s model	Low	Low	Very-High	High	-
koloniari’s model1	High	Low	Medium	High	-
Huo’s model	Medium	Low	Medium	High	-
FSDNs model	High	Low	Very-High	Very-High	-
TPM model	High	Low	Medium	Medium	-
Evo-graph model	High	Low	Very-High	Very-High	-
Koloniari’s model2	High	Low	Medium	Medium	Low
Khurana’s model	High	Medium	-	Low	Low
Chronos model	Very-High	Low	Medium	Medium	Medium
G^* Server	Medium	High	Low	Medium	Medium
MG^* Server	High	Low	Very-Low	Low	Medium
Kineograph Server	Low	Low	High	High	-
G^* model	Low	Medium	Low	Medium	Medium
MG^* model	Medium	Very-Low	Very-Low	Low	Medium
Kineograph model	Very-Low	Very-Low	High	High	-

compared to algorithms on static graphs. Therefore, the output representation of dynamic graph algorithms is different.

Since static graph elements have no validation time, static algorithms of the same problem can produce the same output elements given the same input elements. On the other hand, the added time parameter on dynamic graphs introduces multiple validation times to dynamic graph elements. The validation time of each element belongs to the dynamic graph time interval. Furthermore, dynamic graph algorithms of the same problem cannot produce the same output elements given the same input elements because of the way of handling the time parameter. On light of such difference, we introduce a novel classification for dynamic graph output representation, assuming that the dynamic graph time interval is $[t_i, t_j]$, where $i < j$ as follows:

- Single dynamic solution: results from applying a dynamic graph algorithm only once over a dynamic graph, where its elements are valid at time interval $[t_i, t_j]$. This solution elements have different validation times, which belong to the time interval $[t_i, t_j]$ [14][15][29][30].
- Multiple static solutions: this type of output results from applying a static graph algorithm once at each snapshot of the sequence of snapshots that are valid at time interval $[t_i, t_j]$. All elements of each solution are valid at single time point [21].
- Multiple dynamic solutions: this type is similar to the first type with a little difference. While dynamic graph algorithm is also applied once on a dynamic graph, the output is a set of dynamic solutions. Each solution is valid at a sub-interval, wich belongs to the time interval $[t_i, t_j]$. Furthermore, each solution is the optimal solution during its sub-interval time [23].
- Single aggregated solution: this output is obtained by applying an aggregation function over a set of static or dynamic solutions. Also, it is considered as a second stage after getting the solutions set by the second or the third type of the output representation at the first stage [23].

VI. Conclusion and Future Work

In the real world, networks continuously evolve with time and need to be stored in a dynamic graph that handles such intrinsic property. The contribution of this survey is to provide a complete and thorough overview of dynamic graphs and its related problems, and conclude and compare the prominent work done on this topic. We accomplished our goal by explaining the related terminologies of dynamic graph temporal evolution and query operators. The survey also proposes a novel categorization of the existing dynamic graph models while discussing the main issues of each category. Another novel classification, regarding the output representation of dynamic graph algorithms, is also presented.

Hence, we can provide a comprehensive toolbox for a generic dynamic graph model. We suggest a toolbox for a dynamic graph model that contains the following properties:-

- The dynamic graph model should keep continuous evolutions.
- Support directed and undirected structure.
- No redundancy in the stored data.
- Support temporal evolution with its both types: topological and attributes.
- Support all types of query's attributes: time granularity, what to query and node granularity.
- Has a good query processing performance.
- The model structure should allow developing dynamic graph algorithms easily.

References

[1] R. L. Breiger, ”The analysis of social networks”, na, 2004, pp. 505-526.

[2] K. Musia, P. Kazienko,” Social networks on the internet”, World Wide Web, vol. 16, pp. 31-72, 2013.

[3] S. Wasserman, K. Faust, ”Social network analysis: Methods and applications”, Cambridge university press, vol. 8, Nov 1994.

[4] M. J. Neely, E. Modiano, C. E. Rohrs, ”Dynamic power allocation and routing for time-varying wireless networks”, Selected Areas in Communications, IEEE Journal, vol. 23, pp. 89-103, 2005.

[5] E. Khler, K. Langkau, M. Skutella, ”Time-expanded graphs for flow-dependent transit times”, In AlgorithmsESA, Springer Berlin Heidelberg, pp. 599-611, Jan 2002.

[6] A. G. Labouseur, et al, ”The g* graph database: efficiently managing large distributed dynamic graphs”, Distributed and Parallel Databases, Springer, pp. 1-36, 2014.

[7] A. Bolour, T. L. Anderson, L. J. Dekeyser, H. K. T. Wong, ”The role of time in information processing: a survey”, ACM SIGART Bulletin, pp. 28-46, Apr 1982.

[8] J. Clifford, S. Gadia, S. Jajodia, A. Segev, R. Snodgrass, ”Temporal databases: theory, design and implementation”. Redwood City: Benjamin-Cummings, 1993.

[9] C. J. Date, H. Darwen, N. Lorentzos, ”Temporal data and the relational model”, Elsevier, 2002.

[10] G. zsoyolu, R. T. Snodgrass, ”Temporal and real-time databases: A survey”, Knowledge and Data Engineering, IEEE Trans on, vol. 7, pp. 513-532, Aug 1995.

[11] B. Salzberg, V. J. Tsotras, ”Comparison of access methods for time-evolving data”, ACM Computing Surveys (CSUR), vol. 31, p.p 158-221, Jun 1999.

[12] R. Snodgrass, I. Ahn, ”A taxonomy of time databases”, In ACM Sigmod Record, ACM, vol. 14, pp. 236-246, May 1985.

[13] R. T. Snodgrass, ”Tsql2 tutorial”, In The TSQL2 Temporal Query Language, Springer, US, pp. 33-47, Jan 1995.

[14] Y. Yang, J. X. Yu, H. Gao, J. Pei, J. Li, ”Mining most frequently changing component in evolving graphs”, World Wide Web, Vol. 17, pp. 351-376, May 2014.

[15] B. B. Xuan, A. Ferreira, A. Jarry, ”Computing shortest, fastest, and foremost journeys in dynamic networks”, International Journal of Foundations of Computer Science, vol. 14, pp. 267-285, Apr 2003.

[16] G. Papastefanatos, Y. Stavrakas, T. Galani, ”Capturing the history and change structure of evolving data”, In Proc. of the 5th Int. Conf. on Advances in Databases, Knowledge, and Data Applications, pp. 235-241, 2013.

[17] G. Koloniari, E. Pitoura, ”Partial view selection for evolving social graphs”, In Proc. of the First International Workshop on Graph Data Management Experiences and Systems, ACM, p.9, Jun 2013.

[18] U. Khurana, A. Deshpande, ”Efficient snapshot retrieval over historical graph data”, In Proc. of the 9th IEEE Int. Conf. Data Engineering (ICDE), IEEE, pp. 997-1008, Apr 2013.

[19] G. Koloniari, D. Souravlias, E. Pitoura, ”On graph deltas for historical queries”, In WOSS, Feb 2013.

[20] R. A. Rossi, B. Gallagher, J. Neville, K. Henderson, ”Modeling dynamic behavior in large evolving graphs”, In Proc. of the 6th ACM Int. Conf. on Web search and data mining, ACM, pp. 667-676, Feb 2013.

[21] C. Ren, E. Lo, B. Kao, X. Zhu, R. Cheng, ”On querying historical evolving graph sequences”, Proceedings of the VLDB Endowment, vol. 4, pp. 726-737, 2011.

[22] G. Koloniari, K. Stefanidis, ”Social search queries in time”, In PersDB, 2013.

[23] W. Huo, V. J. Tsotras, ”Efficient temporal shortest path queries on evolving social graphs”, In Conference on Scientific and Statistical Database Management, SSDBM, vol. 14, p.38, Jun 2014.

[24] W. Huo, ”Query processing on temporally evolving social data”, PhD dissertation, University of California, Riverside, 2013.

[25] S. M. R. Beheshti, H. R. Motahari-Nezhad, B. Benatallah, ”Temporal Provenance model (TPM): model and query language”, CoRR, abs/1211.5009, Nov 2012.

[26] W. Han, and et al, ”Chronos: a graph engine for temporal graph analysis”, In Proc of the 9th European Conf. on Computer Systems, ACM, p. 1, Apr 2014.

[27] A. Zaki, M. Attia, D. Hegazy, S. Amin, ”Efficient distributed dynamic graph system”, In Proc. of the 7th Int. Conf. on Intelligent Computing and Information System, IEEE, pp. 667-676, 2015.

[28] R. Cheng, and et al, ”Kineograph: taking the pulse of a fast-changing and connected world”, In Proc. of the 7th ACM European Conf. on Computer Systems, ACM, pp. 85-98, Apr 2012.

[29] M. A. Sakr, R. H. Gting, ”Group spatiotemporal pattern queries”, GeoInformatica, vol. 18, pp. 699-746, Oct 2014.

[30] S. Huang, J. Cheng, H. Wu, ”Temporal graph traversals: definitions, algorithms, and applications”, CoRR, abs/1401.1919, 2014.

Evolutionary Algorithms Based on Decomposition and Indicator Functions: State-of-the-art Survey

Wali Khan Mashwani
Department of Mathematics,
Kohat University of Science & Technology,
Khyber Pakhtunkhwa (KPK), Pakistan

Abdellah Salhi
Department of Mathematical Sciences,
University of Essex,
Wivenhoe Park,Colchester, UK

Muhammad Asif jan
Department of Mathematics,
Kohat University of Science & Technology,
Khyber Pakhtunkhwa (KPK), Pakistan

Muhammad Sulaiman
Department of Mathematics,
Abdul Wali Khan University, Mardan,
Khyber Pakhtunkhwa (KPK), Pakistan

Rashida Adeeb Khanum
Department of Mathematics,
Jinnah College for Women Peshawar,
Khyber Pakhtunkhwa (KPK), Pakistan

Abdulmohsen Algarni
King Khalid University,
College of Computer Science,
Abha, Asir, Saudi Arabia

***Abstract*—In the last two decades, multiobjective optimization has become mainstream because of its wide applicability in a variety of areas such engineering, management, the military and other fields. Multi-Objective Evolutionary Algorithms (MOEAs) play a dominant role in solving problems with multiple conflicting objective functions. They aim at finding a set of representative Pareto optimal solutions in a single run. Classical MOEAs are broadly in three main groups: the Pareto dominance based MOEAs, the Indicator based MOEAs and the decomposition based MOEAs. Those based on decomposition and indicator functions have shown high search abilities as compared to the Pareto dominance based ones. That is possibly due to their firm theoretical background. This paper presents state-of-the-art MOEAs that employ decomposition and indicator functions as fitness evaluation techniques along with other efficient techniques including those which use preference based information, local search optimizers, multiple ensemble search operators together with self-adaptive strategies, metaheuristics, mating restriction approaches, statistical sampling techniques, integration of Fuzzy dominance concepts and many other advanced techniques for dealing with diverse optimization and search problems**

Keywords*—Multi-objective optimization, Multi-objective Evolutionary algorithms (MOEAs), Pareto Optimality, Multi-objective Memetic Algorithm (MOMAs), Pareto dominance based MOEA, Decomposition based MOEA, Indicator based MOEAs.

I. Introduction

Multi-objective optimization deals with problems involving two or more conflicting objectives. In general, optimization problems can be combinatorial or continuous. The Traveling Salesman Problem (TSP) [165] and the Minimum Spanning Tree (MST) are two well-known combinatorial problems. Combinatorial optimization has various applications [38], [35], [170], [31], [181] in air traffic routing, the design of telephone networks, electrical engineering, hydraulic networks, cable TV and computer systems and others. Continuous optimization is widely used in mechanical design problems [109]. This study is concerned with multi-objective optimization problems (MOPs) including continuous variables. The general formulation of a MOP is:

$$\begin{aligned} &\text{minimize } F(x) = (f_1(x), \ldots, f_m(x))^T \\ &\text{subject to } x \in \Omega \end{aligned} \tag{1}$$

where Ω is the decision space, $x = (x_1, x_2, \ldots, x_n)^T$ is a decision vector and x_i, $i = 1, \ldots, n$ are decision variables, $F(x) : \Omega \to R^m$ includes m real valued objective functions in the objective space R^m. If Ω is a closed and connected region in R^n and all objective functions involve continuous variables then problem (1) is called a continuous MOP.

In real world multi-objective optimization problems, objective functions are usually in conflict or mostly incommensurable. Consequently, there is not a unique solution that can minimize all the objective functions at the same time. The problem must be solved in terms of Pareto optimality. This concept was first devised by Francis Ysidro Edgeworth in 1881 and then later on generalized by Vilfredo Pareto in 1896. To describe this concept, we will introduce a few definitions.

A solution $u = (u_1, u_2, \ldots, u_n) \in \Omega$ is said to be Pareto optimal if there does not exist another solution $v = (v_1, v_2, \ldots, v_n) \in \Omega$ such that $f_j(u) \leq f_j(v)$ for all $j = 1, \ldots, m$ and $f_j(u) < f_j(v)$ for at least index k. An objective vector is Pareto optimal if the corresponding decision vector is Pareto optimal. All the Pareto optimal solutions in the decision space form the Pareto Set (PS) and their image in the objective space forms a Pareto Front (PF), [136], [37], [41].

In the last few years, several multi-objective evolutionary algorithms (MOEAs) have been developed [98], [84], [123], [127], [189], [167], [85], [121] and they have proven their power in many demanding real-world optimization tasks [36], [35], [31], [123], [189], [127], [96]. Classical MOEAs can generally be divided into three main paradigms: the Pareto dominance based MOEAs [42], [193], [192], [149], [65], the indicator based evolutionary algorithms (IBEAs) , [199], [20], [12], [19], [157], [178] and the decomposition based MOEAs [183], [101], [185], [129], [132], [125], [86], [130]. MOEAs operate on a population and approximate the set of optimal solutions in a single simulation run, maintaining diversity among these solutions using different measures such as fitness sharing techniques, the niching approach, the Kernel approach, the nearest neighbour approach, the histogram technique, the crowding/clustering estimation technique, the relaxed form of dominance and the restricted mating and many others.

The fast Non-dominated Sorting Genetic Algorithm II (NSGA-II), [42], SPEA2 [192], the Pareto Archive Evolution Strategy (PAES), [88], the Multi-Objective Genetic Algorithm (MOGA), [52], and the Niched Pareto Genetic Algorithm (NPGA), [65], are well known Pareto dominance based MOEAs. Among them, NSGA-II [42] is an improved version of the Non-dominated Sorting Genetic Algorithm (NSGA), [80] for dealing with MOPs. It generates offspring with crossover and mutation and selects the next generation according to non-dominated sorting and crowding distance comparison. SPEA2 [192] is an improved version of Strength Pareto Evolutionary Algorithm (SPEA), [194]. SPEA2, [192], incorporates a fine-grained fitness assignment strategy, a density estimation technique, and an enhanced archive truncation method in contrast to SPEA [194]. It incorporates a mechanism like k-Nearest Neighbour (kNN) and a specialized ranking system to sort the members of the population, and select the next generation of population, from combination of current population and offsprings population created by crossover and mutation. Both SPEA2 [192] and NSGA-II [42] have shown excellent performances on various real-world, scientific and engineering problems.

Memetic Algorithms (MAs) are a growing area of research motivated by the meme concept introduced by Richard Dawkins. MAs are hybrid algorithms that combine local search optimizers and genetic algorithms for solving NP-hard problems. The first multi-objective MA was developed by Ishibuchi and Murata [67]. That was then improved by Jaszkiewicz, [1], [77]. Basically, these algorithms reformulate the given MOP as simultaneous optimization of all weighted Tchebycheff functions or all weighted sum functions. The Adaptive Multi-objective optimization using Genetically Adaptive Multimethod search (AMALGAM) [177] blends multiple search operators to evolve populations.

This paper provides a state-of-the-art survey of MOEAs that employ indicator and decomposition functions for guiding their search and evolve their populations. We have included, especially, those approaches which are recently developed and found in the existing specialized literature of the evolutionary computation (EC).

The rest of this paper is organized as follows. Section II provides the latest and enhanced variants of MOEA/D. Section III is related to Indicator based EAs. Section IV will finally conclude this paper with some future research directions.

II. Decomposition Based Multi-objective Evolutionary Algorithm

Decomposition is a procedure that break down the given system or task into smaller pieces and then optimize them either sequentially or parallel [112]. This concept is already incorporated in many meta-heuristics, namely, tabu search technique [54], simulated annealing (SA) [175], ant colony optimization (ACO) [58], differential evolution [103], particle swarm optimization (PSO) [156], genetic algorithm (GA) [43], evolutionary strategy (ES) [102] for solving various test suits of optimization and search problems. Metaheuristics are higher-level procedure or heuristics designed that efficiently provide good set of solutions for the given optimization problems [22], [169].

In [180], a decomposition-based multi-objective differential evolution particle swarm optimization (DMDEPSO) is developed and intelligently resolved the problems of design of a tubular permanent magnet linear synchronous motor (TPMLSM).Two-Phase Local Search (TPLS) is proposed in [155] in order tackle TSP with bi-objective functions. This proposed algorithm in their first phase of optimization process generates good initial solution based on single objective function and then forward that solutions to their second phase, where local search technique is used to apply on them by using aggregation of objective functions till the time no optimum solutions are found. Improved TPLS are devised in [46], [47] aiming at to improve its anytime performance by employing regular distribution of the weight vectors in order to equally distribute the effort of the objective space in all directions. In [142], [141], cellular multi-objective genetic algorithm (CMOGA) is suggested. It uses canonical cellular model of GAs (cGA) as a baseline model and habituating the weighted sum approach in order to convert the MOP at hand into scalar optimization problems. Moreover, CMOGA also implanted in multi-objective genetic algorithm (MOGA) framework the cellular structures for residing each individual of its population in a cell of spatially structured space. It then locally utilizes genetic operations in the neighbourhood of each cell for creating an offspring population. Another novel and promising cellular genetic algorithm called MOCell developed in [144] for dealing with multi-objective continuous optimization problems. This algorithm maintain an external archive for storing non-dominated set of solutions and utilize them again at certain stages of population evolution. Another algorithm of same nature was proposed in [3] which does not adopt an external archive as like MOCell does [144] in their algorithmic framework.

Nowadays, memetic algorithms (MAs) is an another emerging and hot area of research inspired by Darwinian principles of natural evolution and Dawkins' notion of a meme [61], [94]. This was first introduced by Pablo Moscato in 1989 [138]. They are also called Baldwinian evolutionary algorithms (EA), Lamarckian EAs [150], cultural algorithms, genetic local search or hybrid algorithm [139], [147], [146], [151]. This class of algorithms have shown great performance and achievements in solving real-world problems and various complicated test suites of MOPs [140], [60], [68], [53], [137], [126], [128].

Ishibuchi and Murata were the first researchers whose proposed the multi-objective genetic local search (MOGLS) [70], [67] for solving multi-objective combinatorial optimization problems. This proposed methodology has utilized the scalar fitness function with random weights generation strategy to guide their population until stoping criteria not satisfied. Jaszkiewicz has further improved MOGLS [78] by incorporating the modified parent selection mechanism in Ishibuchi's MOGLS. Jaszkiewicz examined the performance of their MOGLs using test suit of multiple-objective 0/1 knapsack problems (MOKPs). Later on, Ishibuchi further enhanced its MOGLS [63], here the authors apply local search upon limited number solutions as per probability value, P_{LS} with aim to minimize the computation burden of the existing Ishibuchi's MOGLS [70], [67].

Multi-objective evolutionary algorithm based on decomposition (MOEA/D) [183] is well-known developed paradigm in evolutionary computation field. This novel and robust stochastic technique bridges traditional mathematical programming and evolutionary computing (EC) and transforms the MOP at hand into N different scalar optimization sub-problems (SOPs). For this purposes, it employs an aggregation based techniques including weighted sum function, Tchebycheff approach and many others [136] and then optimizes all these SOPs simultaneously rather than solving a MOP directly. Aggregation based fitness assignment strategy of MOEA/D naturally handles convergence and diversity issues contrary to non-decomposition based approaches. The main feature of MOEA/D is its neighborhood relationships among their subproblems that defines based on the distances between their aggregation coefficient vectors.

The original MOEA/D was then further enhanced in [101] by replacing simulated binary crossover (SBX) [81] with differential evolution (DE) [162]. Moreover, in this improved version of MOEA/D called MOEA/D-DE [101], two neighbourhoods are used and each child solution to replace minimum number of old solutions in its neighbourhood structure. Moreover, general guidelines for the formulation of multi-objective continuous test instances are also main part of this study [101]. Recently, in [32], multiple neighbourhood replacement strategies, reference point determination and different decomposition methods are explicitly analyzed upon multi-objective flowshop scheduling test problems. Dynamical resources allocation scheme for the subproblems are introduces in [185], where each subproblems get resources in dynamical manner based on suggested utility function.Gaussian process model is integrated in MOEA/D framework and as result MOEA/D-EGO is developed in [186] to handle an expensive MOP by converting into a number of single-objective optimization subproblems. Gaussian process models provides a probabilistic non-parametric modelling approach for black-box identification of non-linear dynamic systems. The Gaussian processes can highlight areas of the input space, where prediction quality is poor, due to the lack of data or its complexity, by indicating the higher variance around the predicted mean [90].

In [72], [71], [73], [74], simultaneous use of aggregation functions along with neighborhood structures are incorporated in an original MOEA/D framework [183], [101] for solving combinatorial optimization problems with many objective functions. In [154], each subproblem have been associated more than one solution to maintain search diversity.

Mostly MOEA/D [183] frameworks use Tchebycheff and the weighted sum approaches with fixed weight vectors strategy. However, in [99], the wights of aggregation functions are adjusted in adaptive as well as fixed manners. This algorithm make use of external archive and stores non-dominated solutions using modified ϵ-dominance strategy and to utilize that solutions in the generating process of weight vectors. The use of fixed weight vectors adjustment is sometimes creates hurdles that are expected in solving problems with complicated Pareto fronts (PFs), discontinuity or sharpness or low tail in their structure. This novel procedure for adaptive weight vectors generation adopted in MOEA/D-AWA: MOEA/D-adaptive weight adjustment have offered promising results coping with MOPs used in [161].

Aggregation approaches like weighted sum approach [136] and the Tchebycheff approach [136] are widely employed in the framework of MOEA/D. However, they are sometimes unable to deal with problems having had disparately scaled objective functions. In [163], Normal Boundary Intersection (NBI) style Tchebycheff approach is utilized in MOEA/D-DE [101] as resultant new algorithm called MOEA/D-NBI has been developed for solving portfolio management MOPs. The same NBI-style Tchebycheff approach is also implemented in [184] for also dealing with portfolio management problems.

In [97], simulated annealing (SA) is integrated in MOEA/D and the combined impact analyzed upon both constrained multi-objective knapsack problems and unconstrained multi-objective traveling salesman problem. In this algorithm, simulated annealing (SA) has been used as local search with adaptive procedures.

A novel smart multi-objective particle swarm optimization based decomposition (SDMOPSO) is recently suggested in [2] for solving ZDT test problems [198]. Noting that different only decomposition strategies hare incorporated particle swarm optimization (PSO) [49] framework. Like MOEA/D, this algorithm transforms the given MOP into N numbers of SOPs. In [111], PSO injected to MOEA/D to handle the multi-objective $0/1$ knapsack problems and also continuous multi-objective optimization problems. In [125], [132], differential evolution [162] and particle swarm optimization [49] are incorporated simultaneously with self-adaptive procedures in MOEA/D [183] to handle five standard ZDT test problems [198] and CEC'09 test instances [187].

In [33], two enhanced mechanisms such as guided mutation and priority update schemes are introduced in the original framework MOEA/D [101]. This algorithm was denoted by MOEA/D-GM and efficiently tackled CEC'09 test instances [187] consist of both unconstrained and constrained MOPs. MOEA/D-GM creates its new population with guided mutation (GM) rather than differential evolution (DE) [162] and also introduces update mechanism based on priority queue of subproblems.

An interactive decomposition based multiobjective evolutionary algorithm (IMOEA/D) is recently proposed in [55] that incorporates preference mechanism for selecting the preferred sub-problems rather than the preferred region in the objective space. At each interaction, iMOEA/D provides a set of current solutions to decision maker (DM) to pick out the most preferred one for guiding search towards the neighborhood of the selected ones. iMOEA/D are tested upon benchmark functions and various utility functions are used to simulate the DM's responses. Iterative threshold based MOEA/D framework developed in [100] for optimizing sparse signal recovery in compressive sensing.

The synthesis problem of the difference patterns of monopulse antenna arrays are modeled as MOP in [153] with composed of two objective functions including the maximum side-lobe level (MSLL) and beam width (BW) of principal lobe. MOEA/D-DE [101] is applied on these modeled problems and approximated their Pareto Fronts (PFs) with different number of elements and sub-arrays. Linear antenna array design is an electromagnetic optimization problems and can be formulated as a MOP with two objectives: the minimum average side lobe level (SLL) and null control in specific directions. MOEA/D-DE [101] is also applied on these problems as well and for it an optimized spacing between the elements of linear array while achieving the best possible trade-off between the above mentioned two design objective functions [152]. Furthermore, MOEA/D-DE [101] is applied on problems formulated in [57] and have found better optimally sized two mixed-mode circuits including positive second generation current conveyor and current feedback operational amplifier as compared to NSGA-II [42].

A thread-based parallel implementation of MOEA/D framework is devised recently in [145], [48] by executing on modern multi-core processors. Parallel Decomposition (PaDe) is recently developed in [117]. It has habituated a asynchronous generalized island model for solving various decomposed problems. In [164], a parallel version of MOEA/D has been developed that assigns the computational resources for generating solutions in the minimum overlapped update ranges of solutions and tournament selection based on the scalarizing function value to strengthen the selection pressure of its parent population. A new fine grained message passing schemes for the distribution the MOEA/D computations are implemented in MOEA/D-MP framework [44]. In [119], new selection and replacement strategies have been adopted in MOEA/D for solving bi-objective combinatorial optimization problems.

In [86], [129], [84], [85], different multiple search operators are being engaged in MOEA/D framework [185] and handled the test instances designed for the special session of MOEAs competition in IEEE Congress of Evolutionary Computation (CEC'09) [187]. Two different structured and well-established MOEAs in evolutionary computing (EC) field, namely, MOEA/D [183] and NSGA-II [42] are combined at population and generation levels in [124], [122] for two different benchmark functions expressed in continuous variables. The same two well-known algorithms are also engaged altogether for dealing with a hard multiobjective optimization problem in [134]. The concepts of fuzzy dominance are recently introduced in the MOEA/D framework [143] for enhancement of MOEA/D paradigm. The impact of the ensemble use of the different neighbourhood sizes are recently investigated in [188] based on self-adaptive procedures. Very recently, the impact of multiple crossovers are examined in [131] using MOEA/D-DRA [185] as global search technique and experiment carried over CEC'09 test instances [187].

In [4], Tabu Search (TS) is used within MOEA/D framework to solve the multiobjective permutation flow shop scheduling problems. In [91], several problem-specific operators are also investigated in MOEA/D framework [101] to tackle multiobjective mobile agent routing problems. An encoding representation and various genetic operators are designed for wireless sensor networks deployment and power assignment problems (DPAPs) in [92]. Several constraint handling techniques are also adopted in the framework of MOEA/D algorithmic structure for solving constrained K-connected DPAP in WSNs [93]. Covariance matrix adaption evolution strategy (CMA-ES) has been injected to MOEA/D as a local search optimizers. The resultant algorithm abbreviated as MOEA/D-CMA developed in [182] for handling the CEC'09 box-constrained benchmark functions [187].

In [82], Ant colony optimization (ACO) is incorporated within MOEA/D framework [183] and an algorithm called MOEA/D-ACO is developed, where the effects of grouping, neighborhood, and the location information of current solutions are explicitly analyzed over multiobjective $0/1$ knapsack test problems. In [190], a generator based on multivariate Gaussian models is engaged in MOEA/D framework, where probability models samplings new trial solutions and Gaussian distribution models extracts both local and global population distribution information in robust manners. An efficient multiobjective memtic algorithm called MOMADA is recently developed in [83] by utilizing modified Pareto local search methods [113] to explore the neighborhoods of different locally optimal solutions of the subproblem. In [120], Nelder and Meads algorithm also known as nonlinear simplex search method has employed as local search optimizer in MOEA/D framework for solving Zitzler-Deb-Thiele (ZDT) [198] and Deb-Thiele Laumanns-Zitzler (DTLZ) benchmark functions [40].

Decomposition-based memetic algorithm with extended neighbourhood search (D-MAENS) is developed in [135] for solving multi-objective capacitated arc routing problem (MO-CARP). D-MAENS also decomposes the given MO-CARP as like MOEA/D [183] into a number of scalar subproblems by employing the weighted sum approach with a set of uniformly distributed weight vectors adjustment. A new replacement mechanism and the assignment mechanism for offspring solutions are introduced in improved D-MAENS [166]. In [160], a decomposition based memetic algorithm is proposed and examined their performance upon multi-objective vehicle routing problem with time windows (MO-VRPTW). The suggested algorithm accommodates three types of local search methods periodically in combination with novel selection operator. In [132], [125], multi-objective memetic algorithm based on decomposition is developed for solving multi-objective continuous optimization problems. The proposed algorithm employs particle swarm optimization (PSO) and deferential evolution (DE) with self-adaptive manner for population evolution in their suggested enhanced MOEA/D version.

In [133], artificial bee colony (ABC) and teaching-learning-based optimization (TLBO) are engaged within MOEA/D framework to tackle the ZDT test problems [198] and seven unconstrained MOPs of the test suite of the 2009 IEEE congress on evolutionary computation [187]. ABC [62] works on the foraging behavior of a honey bee and TLBO [148] works on the philosophy of teaching and learning process. Opposition-based learning is a fast growing area of research developed in [172]. OBL has been incorporated in MOEA/D framework as resultant an algorithm called MOEA/D-OBL has been sprang up recently in [115]. The suggested opposition-based initial population and opposition-based learning strategy to generate an offspring population have improved the convergence ability of original MOEA/D [183]. An improved MOEA/D algorithm denoted by TMOEA/D is developed in [110]. The proposed algorithm has utilized a monotonic increasing function and transformed each individual objective function into the one as resultant the curve shape of the non-dominant solutions of the transformed multi-objective problem get closed to the hyperplane whose intercept of coordinate axes is equal to one in the original objective function space.

In [6], [5], a decomposition based evolutionary algorithm is developed for solving both benchmark functions with many objectives and also real-world problems including the car side impact problem, the water resource management problem and the constrained ten-objective general aviation aircraft (GAA) design problem. Moreover, the proposed algorithm have been employed Latten hypercube sampling (LHS) mechanism for reference points generation and adaptive epsilon scheme to establish balance between convergence vs diversity dilemma.

MOEA/D with uniform decomposition measurement (MOEA/D-UDM) developed in [116] for many-objective problems (MAPs). In MOEA/D-UNM [116], the authors have been highlighted two main issues concerned with original MOEA/D paradigm dealing with MAPs, firstly, the number of constructed weight vectors are not arbitrary and mainly distributed on the boundary of weight space; secondly, the relationship between the optimal solution of subproblem and its weight vector is nonlinear for the Tchebycheff decomposition approach. To address aforementioned issues, a novel weight vectors initialization method based on the uniform decomposition measurement and modified Tchebycheff decomposition function have introduced in the MOEA/D-UNM framework while coping with MAPs. In [105], both dominance and decomposition concepts have been combined in order to exploit the merits of both paradigms for the purposes to maintain balance between convergence and diversity in the process of population evolution while coping with MAPs.

The concepts of bandit-based operator selection (AOS) method and fitness-rate-rank-based multiarmed bandit (FRRMAB) are borrowed from the existing literature and have been incorporated in MOEA/D framework as consequence MOEA/D-FRRMAB is developed in [106] to handle many-objective optimization problems. A stable matching model based on the preference articulations is employed in [108] as a resultant an algorithm called MOEA/D-STM is developed. The only difference in MOEA/D-STM and [107] are, MOEA/D-STM considers the perpendicular distance between x and the weight vector of subproblem p. However, the algorithm suggested in [107] make use of niche count of p is an additional term with perpendicular distance measurement for diversity improvement and promotion in their current population. MOEA/D-STM uses a stable matching model to find a suitable matching between subproblems and solutions.In [107], the selection of an appropriate solution for each subproblem is based on the interrelationship between subproblems and solutions.

Hybrid MOEA is developed in [104] which maintain different selection principles and two separate co-evolving archives to hold non-dominated solutions. One archive stores solutions with competitive selection pressure and other preserves a population with a satisfied distribution in the objective space. Furthermore, to exploit guidance information towards the Pareto-optimal set (PS), a restricted mating selection mechanism is employed in this algorithm for selecting mating parents from each archive to produce an efficient offspring solutions. Recently, [130], multiobjective algorithm based on multi-method is developed. MMTD employs two well-known algorithms, MOEA/D [183] and NSGA-II [42], for population evolution for dealing with CEC'09 test instances [187] and ZDT test problems.

III. Indicator Based Evolutionary Algorithm

Hypervolume metric was first introduced by Zitzler and Thiele as an indicator or measurement function. It measures both convergence and diversity which are desirable in the context of multi-objective evolutionary optimization [197], [193], [195]. It measures volume dominated by non-dominated solutions in objective space and extremely suitable for assessing the dominance levels of multiple set of solutions in multi-objective evolutionary optimization. For example, consider two Pareto sets (i.e. $A$and B), then the hypervolume indicator values of set A will be higher than B if A dominates B. Hypervolume is also called S metric or size of the space covered [191] or Lebesgue measure [95], [50]. It has been used as part of the selection or archiving process of MOEAs.

In the recent few years, the indicator based evolutionary algorithms (IBEAs) have gained growing attention and much popularity due to its strong theoretical support and background [30], [11], [26], [7], [10]. They have shown high search ability in various investigations [179], [13], [17]. The main features of these algorithms are, they do not need any diversity maintenance mechanism because indicator functions are automatically recover the issue of diversity promotion among their population solutions [176]. The first Hypervolume indicator based EAs called "hypervolume by slicing objectives (HSO)" was developed by Eckart Zitzler in his seminal work [191] and also seminal work of J.Knowels [45]. Subsequently, various indicator based evolutionary algorithms were developed by employing various procedures as like preferences based information [34], [25], [171], [8], [9], [196], different local search optimizers and many others [158]. One practical drawbacks of IBEA is that it needs much time for hypervolume calculation while dealing with many objectives problems. To address the mentioned drawbacks, in the recent past several faster and enhanced of IBEAs have been developed for solving both continuous and combinatorial optimization problems [114], [24], [173], [14], [28]. The algorithms developed in [114], [24] use objectives in place of points and as resultant they are quickly determine their solution contribution to the front. A methodology based on Monte Carlo sampling simulation has been recently introduced in [14] to estimate the ranks of their individuals induced by the hypervolume indicator for not determining the exact indicator values. In [15], [16], HypE: hypervolume estimation algorithm is devised for multi-objective Optimization. HypE uses the same Monte Carlo simulation method to approximate the exact hypervolume values. A fast and enhanced version of HypE is developed [18] for solving many objectives problems. In [29], a generalized methodology for preference-directed hypervolume-based multi objective search called W-HypE is developed. W-HypE also relies on Monte Carlo sampling and thereby allows to tackle problems with an arbitrary number of objectives.

In [76], an idea of scalarizing function-based hypervolume approximation method are introduced IBEAs framework to dealt with many objectives optimization problems. A simple and fast hypervolume indicator-based MOEA (FV-MOEA) is proposed [79] that selects partial solutions rather than the whole solution set of solutions and quickly update the exact hypervolume contributions of different solutions. An iterative approach to indicator-based evolutionary multiobjective optimization is proposed in [69] with main feature in which only a single solution is obtained within single run. The proposed algorithm needs multiple runs for finding a set of solutions.

In [21], an indicator based evolutionary algorithm (IBEA) are compared with well-known algorithms including a fast nondominated sorting genetic algorithm II (NSGA- II) [42] and strength Pareto evolutionary algorithm (SPEA-II) [192] over standard portfolio optimization problems. An algorithm called Prospect indicator based evolutionary algorithm (PIBEA) is developed in [23] for solving DTLZ MOPs [40].The algorithmic structure of PIBEA is somewhat similar to well-known NSGA-II framework, however, it measures the potential of each individual to reproduce offspring that dominate itself and spread out in the objective space of the given MOPs. In [168],a directed search (DS) is incorporated as local search method within global indicator based optimization algorithms. In [89], Newton steepest descent method [51] and Hooke &Jeeves [64] have been concurrently integrated within SMS-EMOEA [20] framework to handle ZDT test problems [198].

An indicator-based ant colony optimization algorithm which is abbreviated as IBACO is devised [118] for solving the multi-objective knapsack problem (MOKP). IBACO engages binary quality indicators to reinforce their best solutions and not to eliminate the worst ones as exercised in the selection phase of IBEA [200].

An indicator-based EMOA called R2-IBEA is proposed in [157], [174], [27] for solving both ZDT [198] and DTLZ test problems [39]. R2-IBEA eliminates dominance ranking and performs selection with the $R2$ indicator [59].The $R2$ indicator usually requires a set of weight vectors that are uniformly distributed in the objective space. R2-IBEA is similar to MOEA/D [183] in a sense, it also habituates aggregation function like the Tchebycheff function with uniformly distributed weight vectors. R2IBEA [157] dynamically adjusts the location of the reference point according to the extent of the current generation individuals in the objective space. The key feature of $R2$ indicator is that it is computationally much less expensive than the hypervolume indicator. Another IBEA that avoids dominance ranking and uses a binary ϵ_2 indicator in the selection process of suggested $IBEA - \epsilon_2$ [199]. The ϵ_2 is also called binary hypervolume indicator which is also computational cost exponentially grows deal with problems having many objective functions.

In [66], the algorithmic behaviors of MOEAs belongs to Pareto dominance-based, decomposition based and hypervolume-based categories are experimentally analyzed upon many-objective knapsack problems. The gathered experimental results are clearly indicate the superiority of decomposition based MOEAs against Pareto dominance and indicators based EAs dealing with knapsack benchmark functions with many objectives.

IV. Conclusion

Pareto dominance-based EAs are main streams in the field of evolutionary computation (EC). However, their performance are greatly degraded on many objectives problems [87], [159], [56], [75]. Indicator based and decomposition based EAs are promising paradigms of EC. Indicator based EAs (IBEAs) mostly use the hypervolume as the indicator function to guide their process of population evolution. Decomposition based MOEAs use aggregation approaches as fitness functions and neighbourhood relationship to structure their scalar optimization problems (SOPs). MOEA/D uses several aggregation functions (i.e., weighted sum approach, Tchebycheff approach and normal-boundary intersection method and the normalized normal constraint method) for converting the problem of approximating the PF into a number of scalar optimization problems functions. This paper provides the latest review of MOEAs that integrate the decomposition concept and indicator functions in their framework along with other efficient techniques like the incorporation of preference based information, local search optimizers, multiple search operators with self-adaptive strategies, metaheuristics, mating restriction approaches, statistical sampling techniques, Fuzzy dominance concepts to tackle issues of convergence and diversity in an efficient manner for dealing with different real-world problems and diverse and complicated MOPs test suites.

References

[1] A.Jaszkiewicz, “Genetic local search for multi-objective combinatorial optimization,” *European Journalof Operational Research*, vol. 137, no. 1, pp. 50–71, 2002.

[2] N. Al Moubayed, A. Petrovski, and J. McCall, “A Novel Smart Multi-objective Particle Swarm Optimisation Using decomposition,” in *Proceedings of the 11th international conference on Parallel problem solving from nature: Part II*, ser. PPSN’10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 1–10.

[3] E. Alba, B. Dorronsoro, F. Luna, A. J. Nebro, P. Bouvry, and L. Hogie, “A Cellular Multi-Objective Genetic Algorithm For Optimal Broadcasting Strategy In Metropolitan MANETs,” *Computer Communications*, vol. 30, no. 4, pp. 685–697, 2007.

[4] A. Alhindi and Q. Zhang, “MOEA/D with Tabu Search for Multiobjective Permutation Flow Shop Scheduling Problems,” in *IEEE Congress on Evolutionary Computation (CEC’14)*, 2014, pp. 1155–1164.

[5] M. Asafuddoula, T. Ray, and R. Sarker, “A Decomposition-Based Evolutionary Algorithm for Many Objective Optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 3, pp. 445–460, 2015.

[6] ——, “A Decomposition Based Evolutionary Algorithm for Many Objective Optimization with Systematic Sampling and Adaptive Epsilon Control,” in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, R. Purshouse, P. Fleming, C. Fonseca, S. Greco, and J. Shaw, Eds. Springer Berlin Heidelberg, 2013, vol. 7811, pp. 413–427.

[7] A. Auger, J. Bader, and D. Brockhoff, “Theoretically Investigating Optimal μ-Distributions for the Hypervolume Indicator: First Results For Three Objectives,” in *Conference on Parallel Problem Solving from Nature (PPSN XI)*, ser. LNCS, R. Schaefer *et al.*, Eds., vol. 6238. Springer, 2010, pp. 586–596.

[8] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, “Articulating User Preferences in Many-Objective Problems by Sampling the Weighted Hypervolume,” in *Genetic and Evolutionary Computation Conference (GECCO 2009)*, G. Raidl *et al.*, Eds. New York, NY, USA: ACM, 2009, pp. 555–562.

[9] ——, “Investigating and Exploiting the Bias of the Weighted Hypervolume to Articulate User Preferences,” in *Genetic and Evolutionary Computation Conference (GECCO 2009)*, G. Raidl *et al.*, Eds. New York, NY, USA: ACM, 2009, pp. 563–570.

[10] ——, “Hypervolume-based Multiobjective Optimization: Theoretical Foundations and Practical Implications,” *Theoretical Computer Science*, vol. 425, pp. 75–103, 2012.

[11] ——, “Hypervolume-Based Multiobjective Optimization: Theoretical Foundations and Practical Implications,” *Theoretical Computer Science*, vol. 425, pp. 75–103, 2012.

[12] J. Bader, “Hypervolume-Based Search for Multiobjective Optimization: Theory and Methods,” Ph.D. dissertation, ETH Zurich, Switzerland, 2010.

[13] ——, “Hypervolume-Based Search for Multiobjective Optimization: Theory and Methods,” Ph.D. dissertation, ETH Zurich, Switzerland, 2010.

[14] J. Bader, K. Deb, and E. Zitzler, “Faster Hypervolume-based Search using Monte Carlo Sampling,” in *Conference on Multiple Criteria Decision Making (MCDM'08)*, ser. LNEMS, M. Ehrgott *et al.*, Eds., vol. 634, Heidelberg, Germany, 2010, pp. 313–326.

[15] J. Bader and E. Zitzler, “HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization,” Computer Engineering and Networks Laboratory (TIK), ETH Zurich, TIK Report 286, November 2008.

[16] ——, “HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization,” Computer Engineering and Networks Laboratory (TIK), ETH Zurich, TIK Report 286, 2008.

[17] ——, “Robustness in Hypervolume-based Multiobjective Search,” Computer Engineering and Networks Laboratory (TIK), ETH Zurich, TIK Report 317, 2010.

[18] ——, “Hype: An algorithm for fast hypervolume-based many-objective optimization,” *Evolutionary Computing*, vol. 19, no. 1, pp. 45–76, Mar. 2011.

[19] ——, “HypE: An Algorithm for Fast Hypervolume-Based Many-Objective Optimization,” *Evolutionary Computation*, vol. 19, no. 1, pp. 45–76, 2011.

[20] N. Beume, B. Naujoks, and M. Emmerich, “SMS-EMOA: Multiobjective Selection based on Dominated hypervolume,” *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.

[21] S. Bhagavatula, S. Sanjeevi, D. Kumar, and C. Yadav, “Multi-objective indicator based evolutionary algorithm for portfolio optimization,” in *Advance Computing Conference (IACC), 2014 IEEE International*, 2014, pp. 1206–1210.

[22] L. Bianchi, M. Dorigo, L. Gambardella, and W. Gutjahr, “A Survey on Metaheuristics for Stochastic Combinatorial Optimization,” *Natural Computing*, vol. 8, no. 2, pp. 239–287, 2009.

[23] P. Boonma and J. Suzuki, “PIBEA: Prospect Indicator Based Evolutionary Algorithm for Multiobjective Optimization Problems.”

[24] L. Bradstreet, L. While, and L. Barone, “A fast incremental hypervolume algorithm,” *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 6, pp. 714–723, 2008.

[25] J. Branke, K. Deb, K. Miettinen, and R. Slowinski, Eds., *Multiobjective Optimization: Interactive and Evolutionary Approaches*. Springer, 2008.

[26] D. Brockhoff, T. Friedrich, and F. Neumann, “Analyzing Hypervolume Indicator Based Algorithms,” in *Conference on Parallel Problem Solving From Nature (PPSN X)*, ser. LNCS, G. Rudolph *et al.*, Eds., vol. 5199. Springer, 2008, pp. 651–660.

[27] D. Brockhoff, T. Wagner, and H. Trautmann, “R2 Indicator Based Multiobjective Search,” *Evolutionary Computation Journal*, vol. 23, pp. 369–395, 2015.

[28] D. Brockhoff and E. Zitzler, “Improving Hypervolume-based Multiobjective Evolutionary Algorithms by Using Objective Reduction Methods,” in *Congress on Evolutionary Computation (CEC 2007)*. IEEE Press, 2007, pp. 2086–2093.

[29] D. Brockhoff, J. Bader, L. Thiele, and E. Zitzler, “Directed Multiobjective Optimization Based on the Weighted Hypervolume Indicator,” *Journal of Multi-Criteria Decision Analysis*, vol. 20, no. 5-6, pp. 291–317, 2013.

[30] D. Brockhoff, T. Friedrich, and F. Neumann, “Analyzing Hypervolume Indicator Based Algorithms,” in *Proceedings of the* 10^{th} *International Conference on Parallel Problem Solving from Nature: PPSN X*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 651–660.

[31] Carlos and R. L. Becerra, “Evolutionary Multi-Objective Optimization in Materials Science and Engineering,” *Materials and Manufacturing Processes*, vol. 24, no. 2, pp. 119–129, 2009.

[32] P. C. Chang, S. H. Chen, Q. Zhang, and J. L. Lin, “MOEA/D for Flowshop Scheduling Problems,” in *IEEE Congress on Evolutionary Computation*, 2008, pp. 1433–1438.

[33] C.-M. Chen, Y.-P. Chen, and Q. Zhang, “Enhancing MOEA/D with Guided Mutation and Priority Update for Multi-Objective Optimization,” in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009*, 2009, pp. 209–216.

[34] T. Chugh, K. Sindhya, J. Hakanen, and K. Miettinen, “An Interactive Simple Indicator-Based Evolutionary Algorithm (I-SIBEA) for Multiobjective Optimization Problems,” in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, A. Gaspar-Cunha, C. Henggeler Antunes, and C. C. Coello, Eds. Springer International Publishing, 2015, vol. 9018, pp. 277–291.

[35] C. Coello and G. Lamont, *Applications of Multi-objective Evolutionary Algorithms*, ser. Advances in natural computation. World Scientific, 2004.

[36] C. A. C. Coello, “A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques,” *Knowledge and Information Systems*, vol. 1, pp. 269–308, 1999.

[37] C. A. Coello Coello, G. B.Lamont, and D. A. Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, New York,, March 2002.

[38] Y. Collette and P. Siarry, *Multiobjective Optimization: Principles and Case Studies*. Springer Science & Business Media, 2003.

[39] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, “Scalable Multi-Objective Optimization Test Problems,” *In Congress on Evolutionary Computation (CEC2002), Piscataway, New Jersey: IEEE service Center*, vol. 1, pp. 825–830, MAy 2002.

[40] ——, “Scalable Test Problems for Evolutionary Multi-Objective Optimization,” in *Evolutionary Multiobjective Optimization: Theoretical Advances and Applications*, A. Abraham, R. Jain, and R. Goldberg, Eds. Springer, 2005, ch. 6, pp. 105–145.

[41] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, 2nd ed., S. Ross and R. Weber, Eds. John Wiley and Sons Ltd, 2002.

[42] K. Deb, A. Pratap, S. Agarwal, and T.Meyarivan, “A Fast and Elitist Multiobjective Genetic Algorithm:NSGA-II,” *IEEE Transsation On Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[43] D. Debels and M. Vanhoucke, “A Decomposition-Based Genetic Algorithm for the Resource-Constrained Project-Scheduling Problem,” *Operations Research*, vol. 55, no. 3, pp. 457–469, 2007.

[44] B. Derbel, A. Liefooghe, G. Marquet, and E.-G. Talbi, “A Fine Grained Message Passing MOEA/D,” in *IEEE Congress on Evolutionary Computation (CEC'15)*, 2015, pp. 1837–1844.

[45] J. D.Knowles, “Local-Search and Hybrid Evolutionary Algorithms for Pareto Optimization,” PhD Thesis, Department of Computer Science, University of Reading, Reading, RG6 6AY, UK., 2002.

[46] J. Dubois-Lacoste, M. Lpez-Ibez, and T. Sttzle, “Adaptive Anytime Two-Phase Local Search,” vol. 6073, pp. 52–67.

[47] ——, “Improving the Anytime Behavior of Two-Phase Local Search,” *Annals of Mathematics and Artificial Intelligence*, vol. 61, no. 2, pp. 125–154, 2011.

[48] J. Durillo, Q. Zhang, A. Nebro, and E. Alba, “Distribution of Computational Effort in Parallel MOEA/D,” in *Learning and Intelligent Optimization*, ser. Lecture Notes in Computer Science, C. Coello, Ed. Springer Berlin Heidelberg, 2011, vol. 6683, pp. 488–502.

[49] R. Eberhart and J.Kennedy, “A New Optimizer Using Particle Swarm Theory,” in *Proceedings of the Sixth International Symposium on Micro Machine and Human Science, MHS'95*, oct. 1995, pp. 39–43.

[50] M. Fleischer, “The Measure of Pareto Optima Applications to Multi-objective Metaheuristics,” in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, C. Fonseca, P. Fleming, E. Zitzler, L. Thiele, and K. Deb, Eds. Springer Berlin Heidelberg, 2003, vol. 2632, pp. 519–533.

[51] J. Fliege, L. M. G. Drummond, and B. F. Svaiter, “Newton's method for multiobjective optimization,” *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 602–626, 2009.

[52] C. Fonseca and P. Fleming, “An Overview of Evolutionary Algorithm in Multi-Objective Optimization,” *Evolutionary Computation*, vol. 3, no. 1, pp. 1–16, 1995.

[53] P. M. França, A. Mendes, and P. Moscato, "A Memetic Algorithm for the total Tardiness Single Machine Scheduling Problem," *European Journal of Operational Research*, vol. 132, no. 1, pp. 224–242, 2001.

[54] X. Gandibleux, N. Mezdaoui, and A. Frville, "A Tabu Search Procedure to Solve MultiObjective Combinatorial Optimization Problems," in *Advances in Multiple Objective and Goal Programming*, ser. Lecture Notes in Economics and Mathematical Systems, R. Caballero, F. Ruiz, and R. Steuer, Eds. Springer Berlin Heidelberg, 1997, vol. 455, pp. 291–300.

[55] M. Gong, F. Liu, W. Zhang, L. Jiao, and Q. Zhang, "Interactive MOEA/D for Multi-objective Decision Making," in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*. ACM, 2011, pp. 721–728.

[56] C. Grosan, "Multiobjective Adaptive Representation Evolutionary Algorithm (MAREA)- A new Evolutionary Algorithm for Multiobjective Optimization." in *WSC*, ser. Advances in Soft Computing, A. Abraham, B. D. Baets, M. Kppen, and B. Nickolay, Eds., vol. 34. Springer, 2004, pp. 113–121.

[57] Guerra-Gomez, E. Tlelo-Cuautle, T. McConaghy, LuisG.delaFraga, G. Gielen, G.Reyes-Salgado, and J.M.Munoz-Pacheco, "Sizing Mixed-mode Circuits by Multi-objective Evolutionary Algorithms," in *53rd IEEE International Midwest Symposium on Circuits and Systems*, 2010, pp. 813–816.

[58] C. Guo, J. Zhibin, H. Zhang, and N. Li, "Decomposition-based classified Ant Colony Optimization Algorithm for Scheduling Semiconductor Wafer Fabrication System," *Computers & Industrial Engineering*, vol. 62, no. 1, pp. 141–151, 2012.

[59] M. P. Hansen and A. Jaszkiewicz, "Evaluating the Quality of Approximations to the Non-Dominated Set," Technical University of Denmark, Technical Report MM-REP-1998-7, 1998.

[60] S. Harris and E. Ifeachor, "Automatic Fesign of Frequency Sampling Filters by Hybrid Genetic Algorithm Techniques," *IEEE Transactions on Signal Processing*, vol. 46, no. 12, pp. 3304–3314, 1998.

[61] W. E. Hart, N. Krasnogor, and J. E. Smith, *Recent advances in Memetic Algorithms*. Springer Science & Business Media, 2005, vol. 166.

[62] R. Hedayatzadeh, B. Hasanizadeh, R. Akbari, and K. Ziarati, "A Multi-Objective Artificial Bee Colony for Optimizing Multi-Objective Problems," in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, vol. 5. IEEE, 2010, pp. V5–277.

[63] T. Y. Hisao Ishibuchi and T. Murata, "Balance between Genetic Search and Local Search in Memetic Algorithms for Multiobjective Permutation Flowshop Scheduling," *IEEE Transiction on Evolutionary Computation*, vol. 7, no. 2, pp. 204–223, April 2003.

[64] R. Hooke and T. A. Jeeves, "Direct Search Solution of Numerical and Statistical Problems," *Journal of ACM*, vol. 8, pp. 212–229, April 1961.

[65] J. Horn, N. Nafpliotis, and D. E. Goldberg., "A Niched Pareto Genetic Algorithm for Multiobjective Optimization," in *Proceedings of the First IEEE Conference on Evolutionary Computation, CEC'94*, 1994.

[66] H. Ishibuchi, N. Akedo, and Y. Nojima, "Behavior of Multiobjective Evolutionary Algorithms on Many-Objective Knapsack Problems," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 264–283, 2015.

[67] H. Ishibuchi and T. Murata, "Multi-Objective Genetic Local Search Algorithm and Its Application to Flowshop Scheduling," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, no. 3, pp. 392–403, 1998.

[68] ——, "Multi-objective Genetic Local Search for Minimizing the number of Fuzzy Rules for Pattern Classification Problems," in *The 1998 IEEE International Conference on Fuzzy Systems Proceedings*, vol. 2, 1998, pp. 1100–1105.

[69] H. Ishibuchi, N. Tsukamoto, and Y. Nojima, "Iterative Approach to Indicator Based Multiobjective Optimization," in *Evolutionary Computation (CEC'07)*, 2007, pp. 3967–3974.

[70] H. Ishibuchi and T. Murata., "Multi-Objective Genetic Local Search Algorithm." in *Proceedings of the Third IEEE International Conference on Evolutionary Computation*, I. T. Fukuda and T. Furuhashi, Eds., Nagoya, Japan, 1996, pp. 119–124.

[71] H. Ishibuchi, Y. Sakane, N. Tsukamoto, and Y. Nojima, "Adaptation of Scalarizing Functions in MOEA/D: An Adaptive Scalarizing Function-Based Multiobjective Evolutionary Algorithm," in *Proceedings of Evolutionary Multi-Criterion Optimization, 5th International Conference EMO'09, Nantes, France, April 7-10, 2009.*, 2009, pp. 438–452.

[72] ——, "Effects of using two Neighborhood Structures on the Performance of Cellular Evolutionary Algorithms for Many-objective Optimization," in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC'09,Trondheim, Norway, 18-21 May, 2009*, 2009, pp. 2508–2515.

[73] ——, "Evolutionary Many-Objective Optimization by NSGA-II and MOEA/D with Large Populations," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, San Antonio, TX, USA, 11-14 October 2009*, 2009, pp. 1758–1763.

[74] ——, "Simultaneous use of Different Scalarizing Functions in MOEA/D," in *Genetic and Evolutionary Computation Conference, GECCO'10, Proceedings, Portland, Oregon, USA, July 7-11, 2010*, 2010, pp. 519–526.

[75] H. Ishibuchi, N. Tsukamoto, and Y. Nojima, "Evolutionary Many-objective Optimization: A Short Review," in *Proceedings of the IEEE Congress on Evolutionary Computation(CEC'08), Hong Kong, China*, 2008, pp. 2419–2426.

[76] H. Ishibuchi, N. Tsukamoto, Y. Sakane, and Y. Nojima, "Indicator-based Evolutionary Algorithm with Hypervolume Approximation by Achievement Scalarizing Functions," in *2010, Proceedings of Genetic and Evolutionary Computation Conference GECCO Portland, Oregon, USA, July 7-11, 2010*, 2010, pp. 527–534.

[77] A. Jaszkiewicz, "On the Computational Efficiency of Multiple Objective Metaheuristics. The Knapsack Problem Case Study," *European Journal of Operational Research*, vol. 158, no. 2, pp. 418–433, 2004.

[78] A. Jaszkiewicz, M. Hapke, and P. Kominek, "Performance of Multiple Objective Evolutionary Algorithms on a Distribution System Design Problem-Computational Experiment," in *Proceedings of First International Conference on Evolutionary Multi-Criterion Optimization (EMO), Zurich, Switzerland*, March 7-9, 2001, pp. 241–255.

[79] S. Jiang, J. Zhang, Y.-S. Ong, A. Zhang, and P. S. Tan, "A Simple and Fast Hypervolume Indicator-Based Multiobjective Evolutionary Algorithm," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2202–2213, 2015.

[80] K.Deb, "Multiobjective Genetic Algorithms: Problems Difficulities and Construction of Test Problems," *Evolutionary Computation*, vol. 7, no. 3, pp. 205–230, 1999.

[81] K.Deb and R. Agrawal, "Simulated Binary Crossover for Continuous Search Space," *Complex System*, vol. 9, pp. 115–148, 1995.

[82] L. Ke, Q. Zhang, and R. Battiti, "MOEA/D-ACO: A Multiobjective Evolutionary Algorithm using Decomposition and AntColony Optimization," *IEEE T. Cybernetics*, vol. 43, no. 6, pp. 1845–1859, 2013.

[83] L. Ke, Q. Zhang, and R.Battiti, "Hybridization of Decomposition and Local Search for Multiobjective Optimization," *IEEE Transactions on Cybernetics*, vol. 44, no. 10, pp. 1808–1820, 2014.

[84] W. Khan, "Hybrid multiobjective evolutionary algorithm based on decomposition," PhD, Department of Mathematical Sciences, University of Essex, Wivenhoe Park, CO4 3SQ, Colchester, UK, January 2012.

[85] W. Khan, A. Salhi, M. A. Jan, and R. Khanum, "Enhanced Version of Gentically Adaptive Multi-Algorithm for Multiobjective Optimization," *International Journal of Advanced Computer Science and Applicat (IJACSA)*, vol. 12, no. 6, pp. 279–287, 2015.

[86] W. Khan and Q. Zhang, " MOEA/D-DRA with Two Crossover Operators," in *Proceedings of the UK Workshop on Computational Intelligence (UKCI 2010)*, 8th–10th September 2010, pp. 1–6.

[87] V. Khare, X. Yao, and K. Deb, "Performance scaling of multi-objective evolutionary algorithms," in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, C. Fonseca, P. Fleming, E. Zitzler, L. Thiele, and K. Deb, Eds. Springer Berlin Heidelberg, 2003, vol. 2632, pp. 376–390.

[88] J. Knowles and D. Corne, "The Pareto Archived Evolution Strategy: A new Baseline Algorithm for Pareto Multiobjective Optimization," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC' 99)*, Piscatay, NJ, JULY 1999, pp. 98–105.

[89] P. Koch, O. Kramer, G. Rudolph, and N. Beume, "On the hybridization of sms-emoa and local search for continuous multiobjective optimiza-

tion," in *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '09. New York, NY, USA: ACM, 2009, pp. 603–610.

[90] J. Kocijan, R. Murray-Smith, C. Rasmussen, and A. Girard, "Gaussian Process Model based Predictive Control," in *Proceedings of theAmerican Control Conference*, vol. 3, 2004, pp. 2214–2219.

[91] A. Konstantinidis, C. Charalambous, A. Zhou, and Q. Zhang, "Multi-objective Mobile Agent-based Sensor Network Routing using MOEA/D," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*. IEEE, 2010, pp. 1–8.

[92] A. Konstantinidis, K. Yang, and Q. Zhang, "Problem-specific Encoding and Genetic Operation for a Multi-Objective Deployment and Power Assignment Problem in Wireless Sensor Networks," in *IEEE International Conference on Communications*. IEEE, 2009, pp. 1–6.

[93] A. Konstantinidis, K. Yang, Q. Zhang, and F. Gordejuela-Sanchez, "Multiobjective K-connected Deployment and Power Assignment in WSNs using Constraint Handling," in *IEEE Global Telecommunications Conference (GLOBECOM'09)*. IEEE, 2009, pp. 1–6.

[94] N. Krasnogor, A. Aragn, and J. Pacheco, "Memetic Algorithms," in *Metaheuristic Procedures for Training Neutral Networks*, ser. Operations Research/Computer Science Interfaces Series, E. Alba and R. Mart, Eds. Springer US, 2006, vol. 36, pp. 225–248.

[95] M. Laumanns, E. Zitzler, and L. Thiele, "A Unified Model For Multi-Objective Evolutionary Algorithms with Elitism," in *Proceedings of the 2000 Congress on Evolutionary Computation*, vol. 1, 2000, pp. 46–53 vol.1.

[96] C. Lcken, B. Barn, and C. Brizuela, "A Survey on Multi-Objective Evolutionary Algorithms for Many-objective Problems," *Computational Optimization and Applications*, vol. 58, no. 3, pp. 707–756, 2014.

[97] H. Li and D. Landa-Silva, "An Adaptive Evolutionary Multi-objective Approach Based on Simulated Annealing," *Evolutionary Computing*, vol. 19, no. 4, pp. 561–595, Dec. 2011.

[98] H. Li, "Combination of Evolutionary Algorithms with Decomposition Techniques for Multiobjective Optimization," PhD, Department of Computer Science, University of Essex, Wevehoe Park, Colchester, Essex, CO4 3SQ, UK, 2007.

[99] H. Li, M. Ding, J. Deng, and Q. Zhang, "On the use of Random Weights in MOEA/D," in *IEEE Congress on Evolutionary Computation (CEC'15)*, 2015, pp. 978–985.

[100] H. Li, X. Su, Z. Xu, and Q. Zhang, "MOEA/D with Iterative Thresholding Algorithm for Sparse Optimization Problems," in *Parallel Problem Solving from Nature - PPSN XII*, ser. Lecture Notes in Computer Science, C. Coello, V. Cutello, K. Deb, S. Forrest, G. Nicosia, and M. Pavone, Eds. Springer Berlin Heidelberg, 2012, vol. 7492, pp. 93–101.

[101] H. Li and Q. Zhang, "Multiobjective Optimization Problems With Complicated Pareto Sets: MOEA/D and NSGA-II," *IEEE Transsation On Evolutionary Computation*, vol. 13, no. 2, pp. 284–302, April 2009.

[102] ——, "A Decomposition-based Evolutionary Strategy for Bi-objective LOTZ Problem," in *Adaptation in Artificial and Biological Systems*. University of Essex, UK., April 2006, pp. 1–5.

[103] ——, "A Multiobjective Differential Evolution Based on Decomposition for Multiobjective Optimization with Variable Linkages," in *Parallel Problem Solving from Nature - PPSN IX*, ser. Lecture Notes in Computer Science, T. Runarsson, H.-G. Beyer, E. Burke, J. Merelo-Guervs, L. Whitley, and X. Yao, Eds. Springer Berlin Heidelberg, 2006, vol. 4193, pp. 583–592.

[104] K. Li, K. Deb, and Q. Zhang, "Evolutionary Multiobjective Optimization with Hybrid Selection Principles," in *IEEE Congress on Evolutionary Computation (CEC'15)*, 2015, pp. 900–907.

[105] K. Li, K. Deb, Q. Zhang, and S. Kwong, "An Evolutionary Many-Objective Optimization Algorithm Based on Dominance and Decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 5, pp. 694–716, October 2015.

[106] K. Li, Á. Fialho, S. Kwong, and Q. Zhang, "Adaptive Operator Selection With Bandits for a Multiobjective Evolutionary Algorithm Based on Decomposition," *IEEE Trans. Evolutionary Computation*, vol. 18, no. 1, pp. 114–130, 2014.

[107] K. Li, S. Kwong, Q. Zhang, and K. Deb, "Interrelationship-Based Selection for Decomposition Multiobjective Optimization," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2076–2088, 2015.

[108] K. Li, Q. Zhang, S. Kwong, M. Li, and R. Wang, "Stable Matching-Based Selection in Evolutionary Multiobjective Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 6, pp. 909–923, 2014.

[109] X. Liao, Q. Li, X. Yang, W. Zhang, and W. Li, "Multiobjective Optimization for Crash Safety Design of Vehicles using Stepwise Regression Model," *Structural and Multidisciplinary Optimization*, vol. 35, no. 6, pp. 561–569, 2008.

[110] H. lin Liu, F. Gu, and Y. ming Cheung, "T-MOEA/D: MOEA/D with Objective Transform in Multi-objective Problems," in *International Conference of Information Science and Management Engineering (ISME'10)*, vol. 2, 2010, pp. 282–285.

[111] Y. Liu and B. Niu, "A Multi-objective Particle Swarm Optimization Based on Decomposition," in *Emerging Intelligent Computing Technology and Applications*, ser. Communications in Computer and Information Science, D.-S. Huang, P. Gupta, L. Wang, and M. Gromiha, Eds. Springer Berlin Heidelberg, 2013, vol. 375, pp. 200–205.

[112] F. Lootsma and K. Ragsdell, "State-of-the-art in Parallel Nonlinear Optimization," *Parallel Computing*, vol. 6, no. 2, pp. 133–155, 1988.

[113] T. Lust and A. Jaszkiewicz, "Speed-up Techniques for solving large-scale biobjective TSP," *Computers & Operations Research*, vol. 37, no. 3, pp. 521–533, 2010.

[114] P. H. Lyndon While and S. Huband, "A Faster Algorithm for Calculating Hypervolume," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 1, pp. 29–38, February 2006.

[115] X. Ma, F. Liu, Y. Qi, M. Gong, M. Yin, L. J. Lingling Li, and J. Wu, "MOEA/D with Opposition-Based Learning for Multiobjective Optimization Problem," *Neurocomputing*, vol. 146, pp. 48–64, 2014.

[116] X. Ma, Y. Qi, L. Li, F. Liu, L. Jiao, and J. Wu, "MOEA/D with Uniform Decomposition Measurement for Many-Objective Problems," *Soft Computing*, vol. 18, no. 12, pp. 2541–2564, 2014.

[117] A. Mambrini and D. Izzo, "PaDe: A Parallel Algorithm Based on the MOEA/D Framework and the Island Model," in *Proceedings OF 13th International Conference Parallel Problem Solving from Nature-PPSN XIII, Ljubljana, Slovenia, September 13-17, 2014.*, 2014, pp. 711–720.

[118] I. B. Mansour and I. Alaya, "Indicator Based Ant Colony Optimization for Multi-objective Knapsack Problem," *Procedia Computer Science*, vol. 60, pp. 448–457, 2015.

[119] G. Marquet, B. Derbel, A. Liefooghe, and E. Talbi, "Shake Them All! - Rethinking Selection and Replacement in MOEA/D," in *Proceedings of Parallel Problem Solving from Nature-PPSN XIII-13th International Conference, Ljubljana, Slovenia, September 13-17*, 2014, pp. 641–651.

[120] S. Z. Martnez and C. A. C. Coello, "A Hybridization of MOEA/D with the Nonlinear Simplex Search Algorithm," in *IEEE Congress on Evolutionary Computation (CEC'13)*, 2013, pp. 48–55.

[121] W. K. Mashwan, "Enhanced versions of Differential Evolution: State-of-the-art Survey," *International Journal Computing Sciences and Mathematics(IJCSM)*, vol. 5, no. 2, pp. 107–126, 2014.

[122] W. K. Mashwani, "A Multimethod Search Approach Based on Adaptive Generations Level," in *Seventh International Conference on Natural Computation(ICNC'11), Shanghai, China, 26-28 July*, 2011, pp. 23–27.

[123] ——, "Hybrid Multiobjective Evolutionary Algorithms: A Survey of the State-of-the-art," *International Journal of Computer Science Issues*, vol. 8, no. 6, pp. 374–392, 2011.

[124] ——, "Integration of NSGA-II and MOEA/D in Multimethod Search Approach: Algorithms," in *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*. ACM, 2011, pp. 75–76.

[125] ——, "MOEA/D with DE and PSO: MOEA/D-DE+PSO," in *The Thirty-first SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, December*, 2011, pp. 217–221.

[126] ——, "Comparison of Evolutionary Algorithm over Multiobjective Optimization Problems," in *Proceeding of International Conference on Modeling and Simulation (ICOMS), Air University Islamabad, Pakistan*, 2013.

[127] ——, "Comprehensive Survey of the Hybrid Evolutionary Algorithms," *International Journal of Applied Evolutionary Computation (IJAEC)*, vol. 4, no. 2, pp. 1–19, July 2013.

[128] ——, "Performanance of AMALGAM over CEC'09 Test Instances," in *Proceeding Third International Conference on Aerospace Science and Engineering (ICASE'13)*, 2013.

[129] W. K. Mashwani and A. Salhi, "A Decomposition-Based Hybrid Multiobjective Evolutionary Algorithm with Dynamic Resource Allocation," *Applied Soft Computing*, vol. 12, no. 9, pp. 2765–2780, 2012.

[130] ——, "Multiobjective Evolutionary Algorithm Based on Multimethod with Dynamic Resources Allocation," *Applied Soft Computing*, vol. 39, pp. 292–309, 2016.

[131] W. K. Mashwani, A. Salhi, M. A. Jan, R.A.Khanum, and M. Sulaiman, "Impact Analysis of Crossovers in Multiobjective Evolutionary Algorithm," *Science International Journal*, vol. 27, no. 6, pp. 4943–4956, December 2015.

[132] W. K. Mashwani and A. Salhi, "Multiobjective Memetic Algorithm Based on Decomposition," *Applied Soft Computing*, vol. 21, pp. 221–243, 2014.

[133] M. Medina, S. Das, C. Coello Coello, and J. Ramirez, "Two decomposition-based modem metaheuristic algorithms for multiobjective optimization- A comparative study," in *Computational Intelligence in Multi-Criteria Decision-Making (MCDM), 2013 IEEE Symposium on*, 2013, pp. 9–16.

[134] Y. Mei, K. Tang, and X. Yao, "Decomposition-Based Memetic Algorithm for Multiobjective Capacitated Arc Routing Problem," *IEEE Trans. Evolutionary Computation*, vol. 15, no. 2, pp. 151–165, 2011.

[135] ——, "Decomposition-based memetic algorithm for multiobjective capacitated arc routing problem," *Evolutionary Computation, IEEE Transactions on*, vol. 15, no. 2, pp. 151–165, 2011.

[136] K. M. Miettinien, *Nonlinear Multiobjective Optimization*, ser. Kluwer's International Series. Norwell, MA: Academic Publishers Kluwer, 1999.

[137] P. Moscato, A. Mendes, and C. Cotta, *Memetic Algorithms. New Optimization Techniques in Engineering*. Berlin Heidelberg: Springer, 2004.

[138] P. Moscato, "On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts-Towards Memetic Algorithms," *Caltech Concurrent Computation Program, C3P Report*, vol. 826, 1989.

[139] ——, "Memetic Algorithms: A Short Introduction," in *New Ideas in Optimization*. McGraw-Hill Ltd., UK, 1999, pp. 219–234.

[140] P. Moscato and M. G. Norman, "A Memetic Approach for the Traveling Salesman Problem Implementation of a Computational Ecology for Combinatorial Optimization On Message-passing Systems," *Parallel Computing and Transputer Applications*, vol. 1, pp. 177–186, 1992.

[141] T. Murata and M. Gen, "Cellular Genetic Algorithm for Multi-Objective Optimization," in *In Proc. of the 4th Asian Fuzzy System Symposium*, 2002, pp. 538–542.

[142] T. Murata, H. Ishibuchi, and M. Gen, "Specification of Genetic Search Direction in Cellular Multiobjective Genetic Algorithm," in *Evolutionary Multicriterion Optimization*. LNCS, Springer-Verlag, pp. 82–95.

[143] M. Nasir, A. K. Mondal, S. Sengupta, S. Das, and A. Abraham, "An Improved Multiobjective Evolutionary Algorithm based on Decomposition with Fuzzy Dominance," in *Proceedings of IEEE Congress on Evolutionary Computation (CEC,11)*. New Orleans, US: IEEE Press, June 5-8 2011, pp. 1–8.

[144] A. J. Nebro, J. J. Durillo, F. Luna, B. Dorronsoro, and E. Alba, "MOCell: A Cellular Genetic Algorithm for Multiobjective Optimization," *International Journal of Intelligent Systems*, pp. 25–36, 2007.

[145] A. Nebro and J. Durillo, "A Study of the Parallelization of the Multi-Objective Metaheuristic MOEA/D," in *Learning and Intelligent Optimization*, ser. Lecture Notes in Computer Science, C. Blum and R. Battiti, Eds. Springer Berlin Heidelberg, 2010, vol. 6073, pp. 303–317.

[146] F. Neri and C. Cotta, "Memetic Algorithms and Memetic Computing Optimization: A literature Review," *Swarm and Evolutionary Computation*, vol. 2, pp. 1–14, 2012.

[147] F. Neri, C. Cotta, and P. Moscato, *Handbook of Memetic Algorithms*. Springer Heidelberg, 2012, vol. 379.

[148] T. Niknam, F. Golestaneh, and M. Sadeghi, "Multiobjective Teaching Learning-Based Optimization for Dynamic Economic Emission Dispatch," *Systems Journal, IEEE*, vol. 6, no. 2, pp. 341–352, 2012.

[149] N.Srinivas and K.Deb, "A Multiobjective Optimization using Nondominated Sorting in Genetic Algorithms," *J Evol Comput*, vol. 2, no. 3, pp. 221–248, 1994.

[150] Y. S. Ong and A. Keane, "Meta-Lamarckian learning in Lemetic Algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 2, pp. 99–110, 2004.

[151] Y.-S. Ong, M. H. Lim, and X. Chen, "Research frontier-memetic computationpast, present & future," *IEEE Computational Intelligence Magazine*, vol. 5, no. 2, p. 24, 2010.

[152] S. Pal, B.-Y.Qu, S. Das, and P.N.Suganthan, "Optimal Synthesis of Linear Antenna Arrays with Multi-Objective Differential Evolution," *Progress In Electromagnetics Research B*, vol. 21, pp. 87–111, 2010.

[153] S. Pal, S. Das, A. Basak, and P. Suganthan, "Synthesis of Difference Patterns for Monopulse Antennas with Optimal Combination of array-size and number of subarrays-A Multi-Objective Optimization Approach," *Progress In Electromagnetics Research B*, vol. 21, pp. 257–280, 2010.

[154] P. Palmers, T. McConnaghy, M. Steyaert, and G. Gielen, "Massively Multi-topology Sizing of analog Integrated Circuits," in *Proceedings of the Conference on Design, Automation and Test in Europe (DATE'09)*. 3001 Leuven, Belgium: European Design and Automation Association, 2009, pp. 706–711.

[155] L. Paquete and T. Stützle, "A Two-Phase Local Search for the Biobjective Traveling Salesman Problem," in *Proceedings of the Evolutionary Multi-Criterion Optimization, Second International Conference, EMO'3, Faro, Portugal, April 8-11*, 2003, pp. 479–493.

[156] W. Peng and Q. Zhang, "A Decomposition-based Multi-objective Particle Swarm Optimization Algorithm for Continuous Optimization Problems," in *IEEE International Conference on Granular Computing*, 2008, pp. 534–537.

[157] D. Phan and J. Suzuki, "R2-IBEA: R2 indicator based evolutionary algorithm for multiobjective optimization," in *IEEE Congress on Evolutionary Computation (CEC'13)*, 2013, pp. 1836–1845.

[158] M. Pilát and R. Neruda, "Hypervolume-based Local Search in Multiobjective Evolutionary Optimization," in *Proceedings of the Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '14. New York, NY, USA: ACM, 2014, pp. 637–644.

[159] R. C. Purshouse and P. J. Fleming, "Evolutionary Many-Objective Optimisation: An Exploratory Analysis," in *IEEE Congress on Evolutionary Computation (CEC'03)*, vol. 3. IEEE, 2003, pp. 2066–2073.

[160] Y. Qi, Z. Hou, H. Li, J. Huang, and X. Li, "A decomposition based memetic algorithm for multi-objective vehicle routing problem with time windows," *Computers and Operations Research*, vol. 62, no. C, pp. 61–77, October 2015.

[161] Y. Qi, X. Ma, L. J. Fang Liu, J. Sun, and J. Wu, "MOEA/D with Adaptive Weight Adjustment," *Evolutionary Computation*, vol. 22, no. 2, pp. 231–264, May 2014.

[162] R.Storn and K.V.Price, "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces," ICSI, Technical Report TR-95-012, 1995.

[163] l. Rubio-Largo, Q. Zhang, and M. Vega-Rodrguez, "Multiobjective Evolutionary Algorithm Based On Decomposition for 3-objective Optimization Problems with Objectives in Different Scales," *Soft Computing*, vol. 19, no. 1, pp. 157–166, 2015.

[164] H. Sato, M. Miyakawa, and E. Pérez-Cortés, "A Parallel MOEA/D Generating Solutions in Minimum Overlapped Update Ranges of Solutions," in *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO Companion '15. New York, NY, USA: ACM, 2015, pp. 775–776.

[165] A. Schuster and Würzburg, "About Travelling Salesmen and Telephone Network-Combinatirial Optimization at High School," *ZDM international Reviwer on Mathematical Education*, vol. 36, no. 2, pp. 77–81, 2004.

[166] R. Shang, J. Wang, L. Jiao, and Y. Wang, "An Improved Decomposition-Based Memetic Algorithm for Multi-Objective Capac-

itated Arc Routing Problem," *Applied Soft Computing*, vol. 19, pp. 343–361, 2014.

[167] D. Simon, *Evolutionary Optimization Algorithms: Biologically Inspired and Population-Based Approches to Computer Intelligence*. John Wiley & Sons, 2013.

[168] V. A. Sosa-Hernandez, O. Schütze, G. Rudoph, and H. Trautmann, "Directed search method for indicator-based multi-objective evolutionary algorithms," in *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*, ser. GECCO '13 Companion. New York, NY, USA: ACM, 2013, pp. 1699–1702.

[169] K. Srensen, "Metaheuristicsthe Metaphor Exposed," *International Transactions in Operational Research*, vol. 22, no. 1, pp. 3–18, 2015.

[170] M. G. C. Tapia and C. A. C. Coello, "Applications of Multi-Objective Evolutionary Algorithms in Economics and Finance: A Survey." *IEEE Congress on Evolutionary Computation*, vol. 7, pp. 532–539, 2007.

[171] L. Thiele, K. Miettinen, P. J. Korhonen, and J. Molina, "A Preference-Based Evolutionary Algorithm for Multi-Objective Optimization," *Evolutionary Computation*, vol. 17, no. 3, pp. 411–436, 2009.

[172] H. Tizhoosh, "Opposition-Based Learning: A New Scheme for Machine Intelligence," in *Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, vol. 1, 2005, pp. 695–701.

[173] H. Trautmann, T. Wagner, D. Biermann, and C. Weihs, "Indicator-based Selection in Evolutionary Multiobjective Optimization Algorithms Based On the Desirability Index," *Journal of Multi-Criteria Decision Analysis*, vol. 20, no. 5-6, pp. 319–337, 2013.

[174] H. Trautmann, T. Wagner, and D. Brockhoff, "R2-EMOA: Focused Multiobjective Search Using R2-Indicator-Based Selection," in *LION*, ser. Lecture Notes in Computer Science, G. Nicosia and P. M. Pardalos, Eds., vol. 7997. Springer, 2013, pp. 70–74.

[175] D. Tuyttens, J. Teghem, P. Fortemps, and K. Nieuwenhuyze, "Performance of the MOSA Method for the Bicriteria Assignment Problem," *Journal of Heuristics*, vol. 6, no. 3, pp. 295–310, 2000.

[176] T. Ulrich, J. Bader, and L. Thiele, "Defining and optimizing indicator-based diversity measures in multiobjective search," in *Proceedings of the 11th international conference on Parallel problem solving from nature: Part I*, ser. PPSN'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 707–717.

[177] J. A. Vrugt and B. A. Robinson, "Improved Evolutionary Optimization from Genetically Adaptive Mutimethod Search," *Proceedings of the National Academy of Sciences of the United States of America: PNAS (USA)*, vol. 104, no. 3, pp. 708–701, 16th Jaunuary 2007.

[178] M. Wagner, K. Bringmann, T. Friedrich, and F. Neumann, "Efficient Optimization of Many Objectives by Approximation-Guided Evolution," *European Journal of Operational Research*, vol. 243, no. 2, pp. 465–479, 2015.

[179] T. Wagner, N. Beume, and B. Naujoks, "Pareto-, Aggregation-, and Indicator-Based Methods in Many-Objective Optimization," in *Evolutionary Multi-Criterion Optimization*, ser. Lecture Notes in Computer Science, S. Obayashi, K. Deb, C. Poloni, T. Hiroyasu, and T. Murata, Eds. Springer Berlin Heidelberg, 2007, vol. 4403, pp. 742–756.

[180] G. Wang, J. Chen, T. Cai, and B. Xin, "Decomposition-Based Multi-Objective Differential Evolution Particle Swarm Optimization for The Design of a Tubular Permanent Magnet Linear Synchronous Motor," *Engineering Optimization*, vol. 45, no. 9, pp. 1107–1127, 2013.

[181] M. Woehrle, D. Brockhoff, T. Hohm, and S. Bleuler, "Investigating Coverage and Connectivity Trade-offs in Wireless Sensor Networks: The Benefits of MOEAs," in *Multiple Criteria Decision Making for Sustainable Energy and Transportation Systems (MCDM 2008)*, ser. LNEMS, M. Ehrgott *et al.*, Eds., vol. 634. Heidelberg, Germany: Springer, 2010, pp. 211–221.

[182] S. Zapotecas-Martínez, B. Derbel, A. Liefooghe, D. Brockhoff, H. E. Aguirre, and K. Tanaka, "Injecting CMA-ES into MOEA/D," in *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '15. New York, NY, USA: ACM, 2015, pp. 783–790.

[183] Q. Zhang and H. Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition ," *IEEE transaction on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, December 2007.

[184] Q. Zhang, H. Li, D. Maringer, and E. P. K. Tsang, "MOEA/D with NBI-style Tchebycheff Approach for Portfolio Management," in *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2010, Barcelona, Spain, 18-23 July 2010*, 2010, pp. 1–8.

[185] Q. Zhang, W. Liu, and H. Li, "The Performance of a New Version of MOEA/D on CEC'09 Unconstrained MOP Test Instances," *IEEE Congress On Evolutionary Computation (IEEE CEC 2009), Trondheim, Norway*, pp. 203–208, May, 18–21 2009.

[186] Q. Zhang, W. Liu, E. Tsang, and B. Virginas, "Expensive multiobjective optimization by MOEA/D with Gaussian Process Model," *Trans. Evol. Comp*, vol. 14, pp. 456–474, June 2010.

[187] Q. Zhang, A. Zhou, S. Zhaoy, P. N. Suganthany, W. Liu, and S. Tiwariz, "Multiobjective Optimization Test Instances for the CEC 2009 Special Session and Competition," Technical Report CES-487, 2009.

[188] S.-Z. Zhao, P. N. Suganthan, and Q. Zhang, "Decomposition-Based Multiobjective Evolutionary Algorithm With an Ensemble of Neighborhood Sizes," *IEEE Trans. Evolutionary Computation*, vol. 16, no. 3, pp. 442–446, 2012.

[189] A. Zhou, B.-Y. Qu, H. Li, S.-Z. Zhao, P. N. Suganthan, and Q. Zhang., "Multiobjective evolutionary algorithms: A survey of the state-of-the-art," *Swarm and Evolutionary Computation*, vol. 1, pp. 32–49, 16 March 2011, online publised.

[190] A. Zhou, Q. Zhang, and G. Zhang, "A Multiobjective Evolutionary Algorithm Based on Decomposition and Probability Model," in *IEEE Congress on Evolutionary Computation (CEC'12)*. IEEE, 2012, pp. 1–8.

[191] E. Zitzler, "Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications," Ph.D. dissertation, ETH Zurich, Switzerland, 1999.

[192] E. Zitzler, M. Laumanns, and L. Thiele, "SPEA2: Improving the Strength Pareto Evolutionary Algorithm," Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Zurich, Switzerland, TIK Report 103, 2001.

[193] E. Zitzler and L. Thiele, "An Evolutionary Approach for Multiobjective Optimization: The Strength Pareto Approach," Computer Engineering and Networks Laboratory (TIK), ETH Zurich, TIK Report 43, May 1998.

[194] ——, "An Evolutionary Approach for Multiobjective Optimization: The Strength Pareto Approach," Computer Engineering and Networks Laboratory (TIK), ETH Zurich, TIK Report 43, May 1998.

[195] ——, "Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach," *IEEE Transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, November 1999.

[196] E. Zitzler, L. Thiele, and J. Bader, "SPAM: Set Preference Algorithm for Multiobjective Optimization," in *Conference on Parallel Problem Solving From Nature (PPSN X)*, ser. LNCS, G. Rudolph *et al.*, Eds., vol. 5199. Springer, 2008, pp. 847–858.

[197] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. Grunert da Fonseca, "Performance Assessment of Multiobjective Optimizers: An Analysis and Review," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 2, pp. 117–132, 2003.

[198] E. Zitzler, K. Deb, and L. Thiele, "Comparsion of Multiobjective Evolutionary Algorithms: Emperical Results," *Evolutionary Computation*, vol. 8, no. 2, pp. 173–195, 2000.

[199] E. Zitzler and S. Knzli, "Indicator-based Selection in Multiobjective Search," in *in Proceeding of* 8^{th} *International Conference on Parallel Problem Solving from Nature (PPSN VIII*. Springer, 2004, pp. 832–842.

[200] ——, "Indicator-Based Selection in Multiobjective Search," in *Parallel Problem Solving from Nature - PPSN VIII*, ser. Lecture Notes in Computer Science, X. Yao, E. Burke, J. Lozano, J. Smith, J. Merelo-Guervs, J. Bullinaria, J. Rowe, P. Tino, A. Kabn, and H.-P. Schwefel, Eds. Springer Berlin Heidelberg, 2004, vol. 3242, pp. 832–842.

Implementation and Evaluation of a Secure and Efficient Web Authentication Scheme using Mozilla Firefox and WAMP

Yassine SADQI, Ahmed ASIMI, Younes ASIMI, Zakaria TBATOU,Azidine GUEZZAZ
Departments of Mathematics and Computer Science, Information Systems and Vision Laboratory (LabSIV), Faculty of Sciences, Ibn Zohr University B.P 8106, City Dakhla, Agadir, Morocco

***Abstract*— User authentication and session management are two of the most critical aspects of computer security and privacy on the web. However, despite their importance, in practice, authentication and session management are implemented through the use of vulnerable techniques. To solve this complex problem, we proposed new authentication architecture, called StrongAuth. Later, we presented an improved version of StrongAuth that includes a secure session management mechanism based on public key cryptography and other cryptographic primitives. In this paper, we present an experimental implementation and evaluation of the proposed scheme to demonstrate its feasibility in real-world scenarios. Specifically, we realize a prototype consisting of two modules: (1) a registration module that implements the registration, and (2) an authentication module integrating both the mutual authentication and the session management phases of the proposed scheme. The experimental results show that in comparison to traditional authentication and session management mechanisms, the proposed prototype presents the lowest total runtime.**

Keywords*— authentication; session management; web security; cryptographic primitives; computer security and privacy; security implementation; authentication scheme; Mozilla Firefox; WAMP

I. INTRODUCTION (*Heading 1*)

Despite its widely studied security problems [1]–[13], password authentication through an HTML form is the dominant mechanism for authenticating users in modern web applications [4], [14], [15]. More specifically, the lack of authentication standard HTML form and the limited security background of webmasters have created a set of unique design and implementation choices that contain multiple security vulnerabilities [3], [12]. While authentication experts proposed a wide range of secure alternatives [16][17]–[20], Bonneau et al. [4] showed that the majority of these schemes offer more security than passwords, but they are difficult to use and / or expensive to deploy [4].

In [13] we have proposed a new authentication architecture, called StrongAuth, to enhance security without sacrificing usability and deployability. Specifically, our proposal does not require any additional equipment except a modern web browser. Later, we presented an improved version of StrongAuth including a secure session management mechanism [21]. This version covers the complete cycle of authentication in the context of web applications, consisting of mutual authentication phases and the subsequent HTTP requests authentication [21].

In this paper, we realize a prototype consisting of two modules: A registration module that implements the registration phase of the proposed scheme [21], and another authentication module that integrates both the mutual authentication and the session management phases. On the one hand, we integrate the client architecture component as an extension of Mozilla Firefox that can easily install it using Mozilla Firefox Add-ons Manager and the server component as a service of PHP applications within the WAMP platform, which allows us to avoid recompiling the source code of Mozilla Firefox and have an independent server component of the application code; which facilitates the deployment. On the other hand, we evaluate the performance of registration and authentication modules to evaluate the theoretical study presented in [21].

The rest of the paper is organized as follows: in section 2 we briefly review the registration, the authentication and session management phases of the proposed scheme [21]. In Section 3 we are particularly interested in the implementation of our prototype to show the feasibility of the proposed scheme. Section 4 presents our results and discussions of the experimental evaluation of the proposed prototype. Section 5 concluded the paper.

In the rest of this paper, we denoted by:

U_i ith User.
ID_i Unique identifier of user U_i.
P_i Password of user U_i.
$Salt_i$ Cryptographic Salt of user U_i .
d Web application domain name.
RW_i Random value used at most once within the scope of a given session generated by the web application for U_i.
RB_i Random value used at most once within the scope of a given session generated by the browser for U_i ..
USK_i,UPK_i Asymmetric key pair for user U_i generated by

	the browser using a secure asymmetric key generation algorithm.
SSK	Web application Private Key.
SPK	Web application Public Key certificate.
PSK_i	Pre-Session key shared between U_i and the web application using HTTPS.
SK_i	Session Key.
X_i^{new}	Renewal of the parameter X_i.
PBKDF( )	Password-Based Key Derivation Function.
MK_i	The master key that is output from an execution of *PBKDF* by the browser.
SE(k,b)	Encryption of b by k using a secure symmetric encryption algorithm.
AE(k,b)	Encryption of b by k using a secure asymmetric encryption algorithm.
SD(k,b)	Decryption of b by k using a secure symmetric decryption algorithm.
AD(k,b)	Decryption of b by k using a secure asymmetric decryption algorithm.
H( )	Cryptographic one way hash function.
HMAC(K,m)	Compute the keyed-Hash Message Authentication Code of message *m* using the secret key *K*.
A || B	The concatenation of binary strings A and B.
$\oplus$	XOR operation.
==	Comparison.
CCS	Client Cryptographic Services.
UACM	User Authentication Credentials Manager.
CS	Client Storage.
SCS	Server Cryptographic Services.
RD	Registration Database.
url	URL of the resource requested by U_i browser.

II. Review of the proposed Scheme

In this section, we briefly review the registration, the authentication and session management phases of the proposed scheme. For all details see [21].

A. Registration Phase

The registration phase is invoked whenever a new user U_i wants to register within the web application. This registration process does not ask from U_i more than choosing an ID_i and P_i. It is know that password-based authentication presents several security problems. For this, the browser transparently from U_i integrates other cryptographic parameters that are used to strengthen users' authentication. Also in this phase, the proposed scheme relies on HTTPS to protect $UPK_{i.}$ confidentiality and integrity. This phase proceeds as follows:

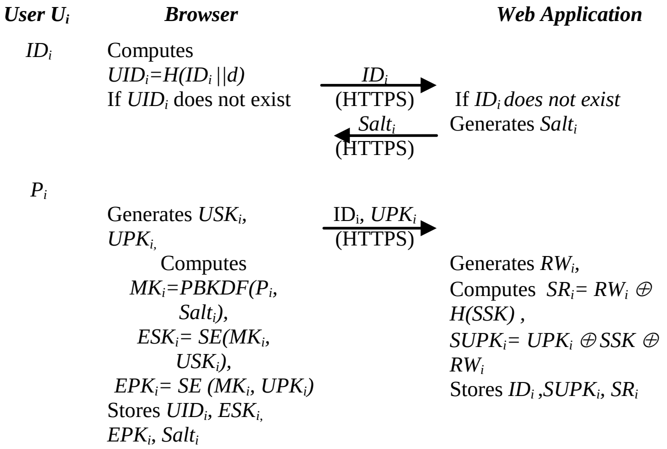

User U_i
Browser
Web Application
ID_i
Computes
$UID_i=H(ID_i \|\|d)$
If UID_i does not exist
ID_i
(HTTPS)
If ID_i does not exist
Generates $Salt_i$
$Salt_i$
(HTTPS)
P_i
Generates USK_i, UPK_i,
ID_i, UPK_i
(HTTPS)
Computes
$MK_i=PBKDF(P_i, Salt_i)$,
$ESK_i= SE(MK_i, USK_i)$,
$EPK_i= SE (MK_i, UPK_i)$
Stores UID_i, ESK_i, EPK_i, $Salt_i$
Generates RW_i,
Computes $SR_i= RW_i \oplus H(SSK)$,
$SUPK_i= UPK_i \oplus SSK \oplus RW_i$
Stores ID_i ,$SUPK_i$, SR_i

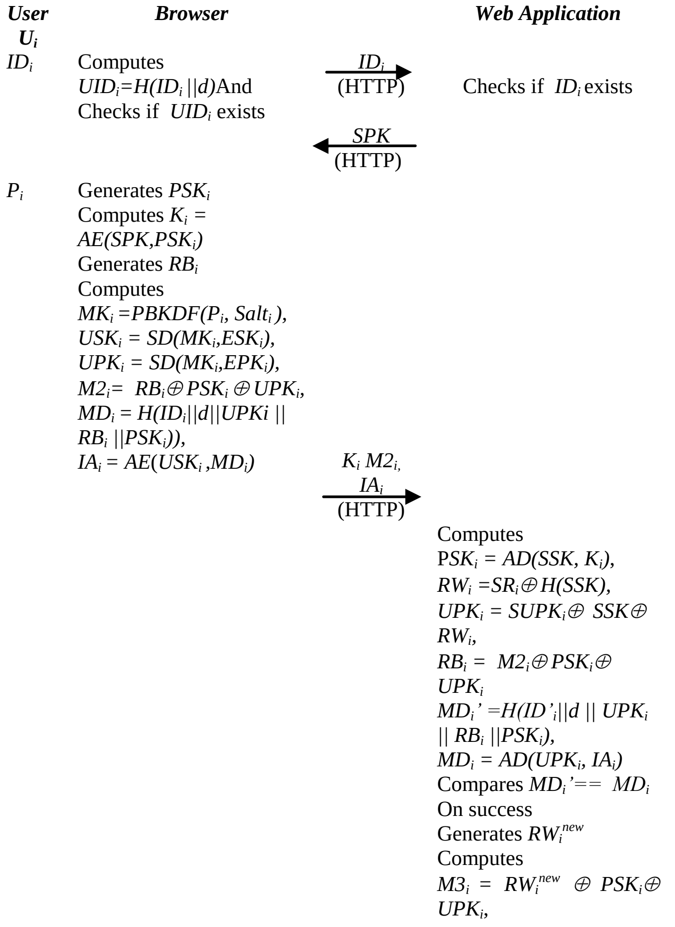
Fig. 1. Registration phase

B. Mutual authentication and session management phase

Primarily, this phase aims to provide:

- A strong mutual authentication between the communicating entities without disclosure of the original authentication settings.
- An agreement on a session key SK_i.
- An HMAC signature in each HTTP request from the browser to the web application using SK_i.

Figure 2 describes the protocol steps


User U_i
Browser
Web Application
ID_i
Computes
$UID_i=H(ID_i \|\|d)$And
Checks if UID_i exists
ID_i
(HTTP)
Checks if ID_i exists
SPK
(HTTP)
P_i
Generates PSK_i
Computes $K_i = AE(SPK,PSK_i)$
Generates RB_i
Computes
$MK_i =PBKDF(P_i, Salt_i)$,
$USK_i = SD(MK_i,ESK_i)$,
$UPK_i = SD(MK_i,EPK_i)$,
$M2_i= RB_i \oplus PSK_i \oplus UPK_i$,
$MD_i = H(ID_i\|\|d\|\|UPKi \|\| RB_i \|\|PSK_i))$,
$IA_i = AE(USK_i ,MD_i)$
K_i $M2_i$, IA_i
(HTTP)
Computes
$PSK_i = AD(SSK, K_i)$,
$RW_i =SR_i \oplus H(SSK)$,
$UPK_i = SUPK_i \oplus SSK \oplus RW_i$,
$RB_i = M2_i \oplus PSK_i \oplus UPK_i$
$MD_i' =H(ID'_i\|\|d \|\| UPK_i \|\| RB_i \|\|PSK_i)$,
$MD_i = AD(UPK_i, IA_i)$
Compares $MD_i'== MD_i$
On success
Generates RW_i^{new}
Computes
$M3_i = RW_i^{new} \oplus PSK_i \oplus UPK_i$,

Computes
$RW_i^{new} = M3_i \oplus PSK_i \oplus UPK_i$,
$M4_i' = H(RB_i \,||\, RW_i^{new} \,||\, PSK_i \,||\, UPK_i)$
Compares $M4_i == M4_i'$
On success
$SK_i = H(ID_i \,||\, d \,||\, UPK_i \,||\, RB_i \,||\, RW_i^{new} \,||\, PSK_i)$

$\xleftarrow[\text{HTTP}]{M3_i\text{ , }M4_i}$

$M4_i = H(RB_i \,||\, RW_i^{new} \,||\, PSK_i \,||\, UPK_i)$,

$SK_i = H(ID_i \,||\, d \,||\, UPK_i \,||\, RB_i \,||\, RW_i^{new} \,||\, PSK_i)$

Generates RB_i^{new}
Computes
$M5_i = RB_i^{new} \oplus PSK_i \oplus UPK_i$,
$M6_i = HMAC(SK_i, url \,||\, RB_i^{new} \,||\, RW_i^{new})$

$\xrightarrow[\text{HTTP}]{M5_i,M6_i}$

Computes $RB_i^{new} = M5_i \oplus PSK_i \oplus UPK_i$,
$M6'_i = HMAC(SK_i, url \,||\, RB_i^{new} \,||\, RW_i^{new})$
Compares $M6'_i == M6_i$

Fig. 2. Mutual authentication and session phases

III. Implementation of the Proposed Prototype

To show the feasibility and effectiveness of the proposed scheme [21], we realize a prototype. Mainly it aims to experience the registration phase, mutual authentication and session management phases. Figure 3 shows the different elements of our prototype, as well as technologies that we used. Tables I and II show the software solutions used in our prototype implementation.

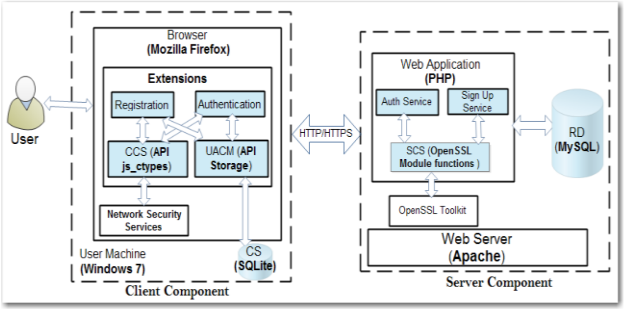


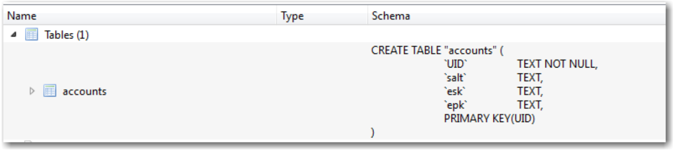
Fig. 3. General architecture of the proposed prototype

TABLE I. The technologies used in the implementation of the client component

Entities	Technology
Browser	*Mozilla Firefox*: The implementation of the client component of our prototype is based on the Mozilla Firefox browser.
Client Cryptographic Services (CCS)	*API js-ctypes* [22]**:** Since we are using Firefox, we choose to use the NSS cryptographic services. To interact with NSS [23] we use the js-ctypes API.
User Authentication Credentials Manager **(UACM)**	*API Storage* [24]: This interface allows the manipulation of the SQLite database from extensions or internal component of Firefox.
Client Storage (CS)	*SQLite* [25]: The main objective of using the multiplatform engine database SQLite is to overcome any installation or administration, which facilitates deployment. Figure 4 shows the client-side authentication settings.

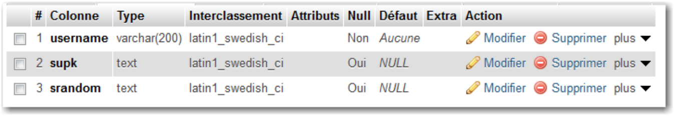
TABLE II. The technologies used in the implementation of the server component

Entities	Technology
Web Server	*Apache* [26]: Apache HTTP Server with the support of the module Mod_SSL [27] that allows the implementation of HTTPS.
Web Application	Application based on PHP 5.
Server Cryptographic Services (SCS)	*Hash* and *OpenSSL* : PHP platform provides a set of cryptographic extensions either as a part of the core PHP functionalities (without the use of a third-party program), or relies on other cryptographic libraries. In our prototype implementation we use two extensions: • *Hash* [28]**:** For the calculation of hash and message authentication code (HMAC). Hash is a digital hash engine, part of the core of PHP. That means we can use these functions in the web application without installing a third-party library. • *OpenSSL* [29]**:** As opposed to Hash, the use of this module requires the presence of an equal or higher version 0.9.6 of OpenSSL cryptographic library. The purpose of this extension is to present a set of cryptographic functions that can be used easily in a PHP script(e.g. asymmetric/symmetric encryption, generation and verification of digital signatures, etc.).
Registration Database (RD)	The relational database management system MySQL [30]. Figure 5 illustrates the authentication settings on the web application side.



```
CREATE TABLE "accounts" (
          `UID`        TEXT NOT NULL,
          `salt`       TEXT,
          `esk`        TEXT,
          `epk`        TEXT,
          PRIMARY KEY(UID)
)
```

Fig. 4. Client-side authentication settings



	#	Colonne	Type	Interclassement	Attributs	Null	Défaut	Extra	Action
□	1	**username**	varchar(200)	latin1_swedish_ci		Non	*Aucune*		Modifier Supprimer plus ▼
□	2	**supk**	text	latin1_swedish_ci		Oui	*NULL*		Modifier Supprimer plus ▼
□	3	**srandom**	text	latin1_swedish_ci		Oui	*NULL*		Modifier Supprimer plus ▼

Fig. 5. Web application-side authentication settings

To simplify the implementation and the experimental evaluation of our prototype, we divide the implementation into two modules: (1) registration module, (2) and an authentication module. The following two subsections provide details on each of these two modules.

A. Registration Module

The purpose of this module is to implement the registration phase of the proposed scheme [21]. We develop the client component as an extension of Mozilla Firefox to avoid recompiling the browser source code (no changes were required to the browser source code). Note that instead of HTML password form field (<input type = "password"), the proposed prototype presents the user with a private window to choose securely passwords (Figure 8). On one hand, in the proposed scheme the browser does not send the password to the web application. The password is used only within the client component to generate a symmetric encryption key (the application does not need to know the user's password). On the

other, Sandler and Wallach [7] discussed in detail that the use of password field is a serious problem, facilitating password theft. Our prototype is not the only one using the standard notification API to create a trusted path to the private password window. Menalis et al. [14] also used this concept.

In detail, the implementation of our Mozilla Firefox registration extension required:

- 1020 lines of chrome / privileged JavaScript code: About 600 lines for client cryptographic service implementation based on the js_ctypes API, which provides access to the cryptographic library of Firefox (NSS), and 420 lines to integrate other client component features and the interaction with the server component.
- 85 lines of XUL code: 70 lines of XUL code to define the interface of the private window for password selection and confirmation (Figure 9). Also, we use an overlay file of about 15 lines of XUL code to integrate the extension components into the Firefox user interface.
- Other extension configuration files (chrome.mnifest, install.rdf ....).

On the other hand, web application cryptographic operations side and the communication with the client component have required 130 lines of custom PHP code. Also, to support HTTPS, we add 20 lines of code to the configuration file of the Apache web server. More importantly, our server component is completely independent of the application code.

The steps of the registration module proceed as follows:

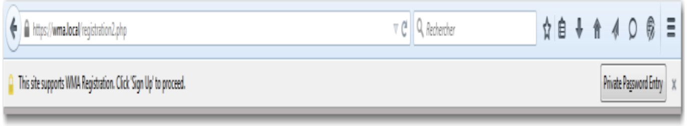
1) The application presents a registration form based on HTML and CSS to the user. This preserves the same user experience. Since users are used to complete such information (name, email ...) in the registration phase of current web application. Once the user filled out the form and click on the "Sign Up" button, the extensions sends the information to the web application.

```
{ "salt":
"MTQxMDUwLjYwOTk0OTAwIDE0MzAyNTkyODWJIE+EEuJ7aaJTCqE7uwm",
"username": "yassine",
"sessionRegID" :"NjAwMzMzNTE4ODAxNTYwNg==",
"regURL" :"https:\/\/wma.local\/wma_reg2.php",
"sitename" :"Test of the Registration Phase",
"description":"Welcome to the test of the Registration prototype. Please
choose a password to finish the registration.",
"imgURL" :"https:\/\/wma.local\/logo.png",
"passwordLabel" :"Password",
"password2Label" :"Confirm your Password",
"failURL" :"https:\/\/wma.local\/registration\/error.php"
}
```

Fig. 6. JSON response sent by the registration service of the PHP application

2) The web application registration service responds with a JSON object containing the salt generated using the Random Generator of a Safe cryptographic Salt per session (RGSCS [31]) and other settings used thereafter (Figure 6).

3) The extension displays a notification bar that tells the user that the application support the registration protocol (Figure 7).



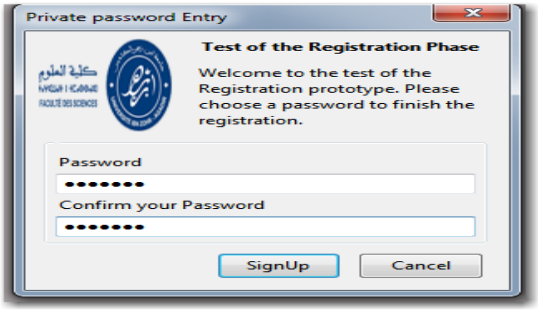
https://wma.local/registration2.php

This site supports WMA Registration. Click 'Sign Up' to proceed.

Fig. 7. Notification bar used to create a trusted path to the private password window

4) The user clicks the button bar "Private Password Entry" to display the password private window (Figure 8). The extension uses the settings (site name, description, imgURL, passwordLabel, password2Label, failURL) containing in the JSON object (Figure 6) to customize the information displayed in this window.

5) In this step, the user chooses a password, confirm it, and click on the "OK" button. If there is an error (e.g., the passwords are different, the user clicks the button without entering a password, etc.), the notification bar displays the corresponding error message and a button "Try again "to try again (Figure 9). Otherwise, the protocol proceeds.



Private password Entry

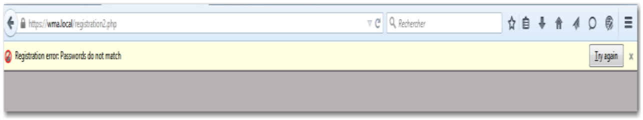
Test of the Registration Phase

Welcome to the test of the Registration prototype. Please choose a password to finish the registration.

Password: •••••••

Confirm your Password: •••••••

Fig. 8. Private window to choose password safely



https://wma.local/registration2.php

Registration error: Passwords do not match

Fig. 9. Error message displayed by our extension using the Mozilla Firefox Notification standard API

6) The CCS extension generates a pair of RSA keys and sends the public key to the web application (Figure 10).

```
https://wma.local/wma_reg2.php
POST /wma_reg2.php HTTP/1.1
Host: wma.local
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:37.0)
Gecko/20100101 Firefox/37.0
….
msg=wmaClientRegistrationExchange&version=wma_v1&sessionRegID=Nj
AwMzg4MTY0ODY2ODc3MQ%3D%3D&username=ahmed &
clientPublicKey=-----BEGIN%20CERTIFICATE-----
```

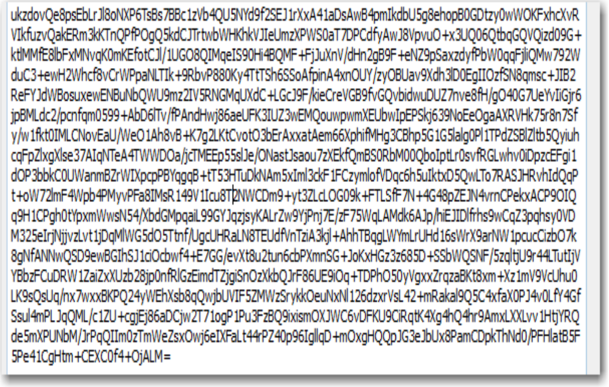
```
%0D%0AMIICOTCCASGgAwIBAgIBFjANBgkqhkiG9w0BAQUFADAi M
QswCQYDVQQGEwJVUzET%0D%0AMBEGA1UECgwKSlMtVEVTVC1
DQTAeFw0xMzA1MDEyMzU5NTlaFw0yMzA1MDEyMzU5%0D%0ANTl
aMCIxCzAJBgNVBAYTAlVTMRMwEQYDVQQDDApUZXN0IFVzZXIx
MIGfMA0GCSqG%0D%0ASIb3DQEBAQUAA4GNADCBiQKBgQDDiEY
DFtsE196LQRrTnjbaOzw7MsPHwHOK9Rh9%0D%0APgMbswFg3Y4eOr4
P0kDsdzG1X7S1M4guAO6BWGYL32ic4q8wl%2BqEOouMgVacSdah%0D
%0An08bAMpWWik7UjOUfaB6T2JTwL6lsA%2BMA86MPO11764d94%2
BLZaF%2BSs1Pf%2Br7WhrU%0D%0ArUT7QQIDAQABMA0GCSqGSIb3
DQEBBQUAA4IBAQBf3rUVlzLf6huxeNQw8Ju%2BCvpH%0D%0AupdUa
266veguVybpFO5vA1sHeIpCi1W0ew75Mh6kLZz6RUUWcBUePoKAQAcF
RVCI%0D%0A%2BM0tzqUksL7C6NR15UhzNpg8nadyjHx7SxRYkH8v9m
NT%2Fse9WtMQvmSlsZcd1b4k%0D%0AumD9kSictyu%2F2ueM4%2F6nG
xuYb2XDj1iC3MP%2FVNn%2FhYqT%2FJelVPGn%2FVQT1INAqIp4%0
D%0Am0%2Fxkjtmmhoy%2FlUtbO%2FeEuI86MuLchT7V0JgkYoSiLvi8Ss
m0xcZEXScojDCpHsx%0D%0AErMcCoA6DHVX0fgtGzSCQrsuUUekLxm
hA6hdCu7ot6YBTpkDbYihzrbqzV0B%0D%0A-----
END%20CERTIFICATE-----%0D%0A
```

Fig. 10. Message sent by the registration of extension containing the public key and other parameters related to the current user's session

7) To improve performance, our prototype performs these two operations in parallel:

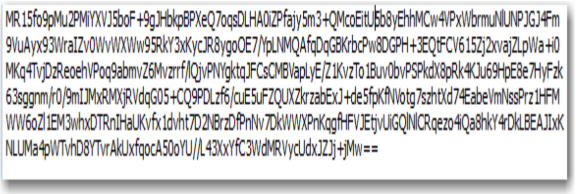
- On the one hand, CCS uses the "salt" sent by the application (Step 2) and the password chosen in the previous step, to generate a secret symmetric key via PBKDF. This key is used by the CCS to encrypt RSA keys with AES algorithm (Figures 11 and 12). Subsequently, based on the API of Mozilla Firefox Storage, *UCAM* recorded in the *CS* authentication settings associated with that specific application (Figure 13). The code of the insertion function UCAM can create the file from the embedded SQLite database. This facilitates the use and deployment.

- On the other, the registration service calculates and stores in the application database the identification information associated with that user (Figure 14).



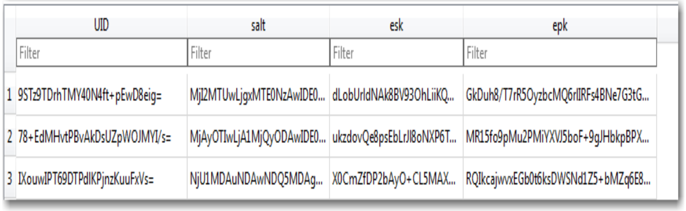
```
ukzdovQe8psEbLrJl8oNXP6TsBs7BBc1zVb4QU5NYd9f2SEJ1rXxA41aDsAwB4pmIkdbU5g8ehopB0GDtzy0wWOKFxhcXvR
VIkfuzvQakERm3kKTnQPfPOgQ5kdCJTrtwbWHKhkVJIeUmzXPWS0aT7DPCdfyAwJ8VpvuO+x3UQ06QtbqGQVQizd09G+
ktlMMfE8lbFxMNvqK0mKEfotCJl/1UGO8QIMqeIS90Hi4BQMF+FjJuXnV/dHn2gB9F+eNZ9pSaxzdyfPbW0qqFjliQMw792W
duC3+ewH2Whcf8vCrWPpaNLTIk+9RbvP880Ky4TtTSh6SSoAfpinA4xnOUY/zyOBUav9Xdh3lD0EgIIOzfSN8qmsc+JIB2
ReFYJdWBosuxewENBuNbQWU9mz2IV5RNGMqUXdC+LGcJ9F/kieCreVGB9fvGQvbidwuDUZ7nve8fH/gO40G7UeYvIiGjr6
jpBMLdc2/pcnfqm0599+AbD6lTv/fPAndHwj86aeUFK3IUZ3wEMQouwpwmXEUbwIpEPSkj639NoEeOgaAXRVHk75r8n7Sf
y/w1fkt0IMLCNovEaU/WeO1Ah8vB+K7g2LKtCvotO3bErAxxatAem66XphifMHg3CBhp5G1G5lalg0Pl1TPdZSBlZltb5Qyiuh
cqFpZlxgXlse37AIqNTeA4TWWDOa/jcTMEEp55slJe/ONastJsaou7zXEkfQmBS0RbM00QboIptLr0svfRGLwhv0iDpzcEFgi1
dOP3bbkC0UWanmBZrWIXpcpPBYqgqB+tT53HTuDkNAm5xIml3ckF1FCzymlofVDqc6h5uIktxD5QwLTo7RASJHRvhIdQqP
t+oW72lmF4Wpb4PMyvPFa8IMsR149V1Icu8T2NWCDm9+yt3ZLcLOG09k+FTLSfF7N+4G48pZEJN4vrnCPekxACP9OIQ
q9H1CPgh0tYpxmWwsN54/XbdGMpqaiL99GYJqzjsyKALrZw9YjPnj7E/zF75WqLAMdk6AJp/hiEJIDlfrhs9wCqZ3pqhsy0VD
M325eIrjNjjvzLvt1jDqMlWG5dO5Ttnf/UgcUHRaLN8TEUdfVnTziA3kjl+AhhTBqgLWYmLrUHd16sWrX9arNW1pcucCizbO7k
8gNfANNwQSD9ewBGIhSJ1ciOcbwf4+E7GG/evXt8u2tun6cbPXmnSG+JoKxHGz3z685D+SSbWQSNF/5zqltjU9r44LTutIjV
YBbzFCuDRW1ZaiZxXUzb28jp0nfRlGzEimdTZjgiSnOzXkbQJrF86UE9iOq+TDPhO50yVgxxZrqzaBKt8xm+Xz1mV9VcUhu0
LK9sQsUq/nx7wxxBKPQ24yWEhXsb8qQwjbUVIF5ZMWzSrykkOeuNxNl126dzxrVsL42+mRakal9Q5C4xfaX0PJ4v0LfY4Gf
Ssul4mPLJqQML/c1ZU+cgjEj86aDCjw2T71ogP1Pu3FzBQ9ixismOXJWC6vDFKU9CiRqtK4Xg4hQ4hr9AmxLXXLvv1HtjYRQ
de5mXPUNbM/JrPqQIIm0zTmWeZsxOwj6eIXFaLt44rPZ40p96IgllqD+mOxgHQQpJG3eJbUx8PamCDpkThNd0/PFHlatB5F
5Pe41CgHtm+CEXC0f4+OjALM=
```

Fig. 11. The encrypted private key of the user Ahmed



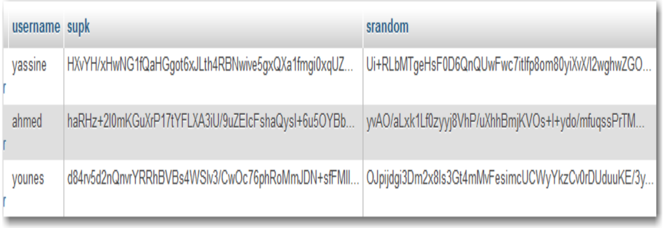
```
MR15fo9pMu2PMiYXVJ5boF+9gJHbkpBPXeQ7oqsDLHA0iZPfajy5m3+QMcoEitU5b8yEhhMCw4VPxWbrmuNlUNPJGJ4Fm
9VuAyx93WraIZv0WvWXWw95RkY3xKycJR8ygoOE7/YpLNMQAfqDqGBKrbcPw8DGPH+3EQtFCV615Zj2xvajZLpWa+i0
MKq4TvjDzReoehVPoq9abmvZ6Mvzrrf/lQjvPNYgktqJFCsCMBVapLyE/Z1KvzTo1Buv0bvPSPkdX8pRk4KJu69HpE8e7HyFzk
63sggnm/r0/9mIJMxRMXjRVdqG05+CQ9PDLzf6/cuE5uFZQUXZkrzabExJ+de5fpKfNVotg7szhtXd74EabeVmNssPrz1HFM
WW6oZl1EM3whxDTRnIHaUKvfx1dvht7D2NBrzDfPnNv7DkWWXPnKqgfHFVJEtjvUiGQlNlCRqezo4iQa8hkY4rDkLBEAJIxK
NLUMa4pWTvhD8YTvrAkUxfqocA50oYU//L43XxYfC3WdMRVycUdxJZJj+jMw==
```

Fig. 12. The encrypted public key of the user Ahmed



	UID	salt	esk	epk
	Filter	Filter	Filter	Filter
1	9STz9TDrhTMY40N4ft+pEwD8eig=	MjI2MTUwLjgxMTE0NzAwIDE0...	dLobUrldNAk8BV93OhLiiKQ...	GkDuh8/T7rR5OyzbcMQ6rlIRFs4BNe7G3tG...
2	78+EdMHvtPBvAkDsUZpWOJMYI/s=	MjAyOTIwLjA1MjQyODAwIDE0...	ukzdovQe8psEbLrJl8oNXP6T...	MR15fo9pMu2PMiYXVJ5boF+9gJHbkpBPX...
3	IXouwIPT69DTPdIKPjnzKuuFxVs=	NjU1MDAuNDAwNDQ5MDAg...	X0CmZfDP2bAyO+CL5MAX...	RQIkcajwvxEGb0t6ksDWSNd1Z5+bMZq6E8...

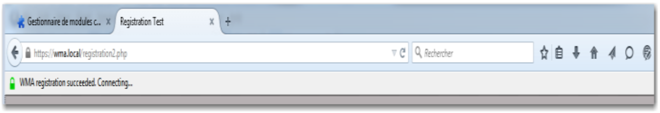
Fig. 13. Client-side authentication settings for three users



username	supk	srandom
yassine	HXvYH/xHwNG1fQaHGgot6xJLth4RBNwive5gxQXa1fmgi0xqUZ...	Ui+RLbMTgeHsF0D6QnQUwFwc7itlfp8om80yiXvX/I2wghwZGO...
ahmed	haRHz+2l0mKGuXrP17tYFLXA3iU/9uZEIcFshaQysl+6u5OYBb...	ywAO/aLxk1Lf0zyyj8VhP/uXhhBmjKVOs+l+ydo/mfuqssPrTM...
younes	d84rv5d2nQnvrYRRhBVBs4WSlv3/CwOc76phRoMmJDN+sfFMll...	OJpijdgi3Dm2x8ls3Gt4mMvFesimcUCWyYkzCv0rDUduuKE/3y...

Fig. 14. Server side authentication settings for three users

8) After the success of previous operations and before redirecting the user's private application, the extension uses the Mozilla Firefox notification bar to inform the user of the success of the procedure (Figure 15).



Gestionnaire de modules c...
Registration Test
https://wma.local/registration2.php
Rechercher

WMA registration succeeded. Connecting...

Fig. 15. Successful message displayed using the Mozilla Firefox Notification standard API

B. Authentication Module

The authentication module of your prototype integrates two related phases of our architecture: (1) initial connection and mutual authentication phase and (2) HTTP requests authentication or session management phase.

Similar to the registration module, we implement the client component of this module as an extension of Mozilla Firefox. For this, we need:

- 1270 lines of chrome / privileged JavaScript code: Approximately 750 lines for client cryptographic service implementation based on the js_ctypes API, which provides access to Firefox cryptographic library (NSS), and 520 lines to include other client component functionality, as well as interaction with the server component.

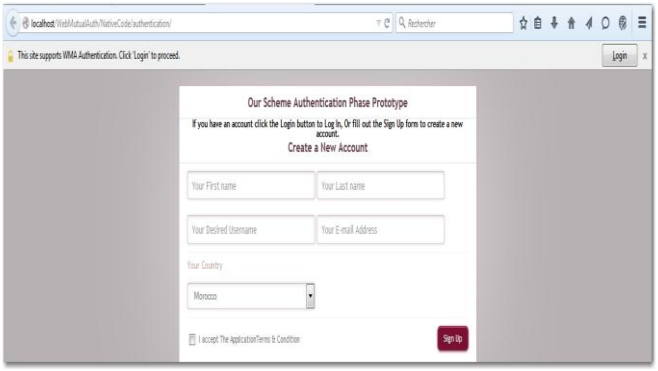
- 85 lines of XUL code: 70 lines of XUL code to define the interface of the private window (Figure 17). Also, we use an overlay file of about 15 lines of XUL code to integrate the components of the extension in the Firefox user interface.

- Other extension configuration files (chrome.mnifest, intall.rdf ....).

In addition, the implementation of server cryptographic services and interaction with the client component has required about 200 lines of PHP code.

In detail, the implementation of the proposed prototype contains the following steps:

1) When our extension detects the support of the proposed protocol. It displays a notification bar asking the user to click on the "Login" button (Figure 16). As shown in Figure 16 unregistered users should first create a new account.



localhost/WebMutualAuth/NativeCode/authentication/

This site supports WMA Authentication. Click 'Login' to proceed.

Our Scheme Authentication Phase Prototype

If you have an account click the Login button to Log In, Or fill out the Sign Up form to create a new account.

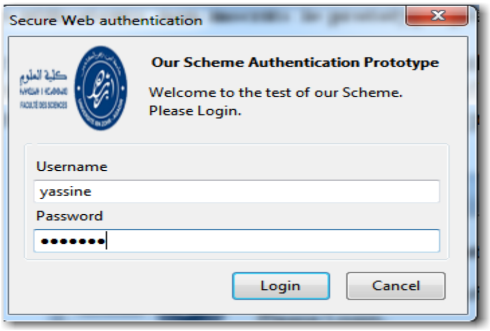
Create a New Account

Your Country: Morocco

☐ I accept The ApplicationTerms & Condition

Fig. 16. Notification bar indicating the support of the mutual authentication protocol

2) The user clicks the "Login" button. This displays a private login window (Figure 17) to enter his username and password. As we have explained in the registration module, each application is free to customize the information to display for their users (logo, site name ...).



Secure Web authentication

Our Scheme Authentication Prototype

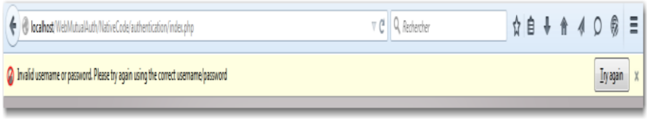
Welcome to the test of our Scheme.
Please Login.

Username: yassine

Password: •••••••

Fig. 17. Private window to safely enter the user's password

3) Once the user enters his/her username / password and click "Login", the UACM checks for authentication parameters associated with that user. In the normal case, the extension retrieves the corresponding user's authentication settings. Otherwise, the extension displays an error message (Figure 18).



localhost/WebMutualAuth/NativeCode/authentication/index.php

Invalid username or password. Please try again using the correct username/password

Fig. 18. Example of an error message asking the user to retry the authentication procedure by clicking the button "Try again"

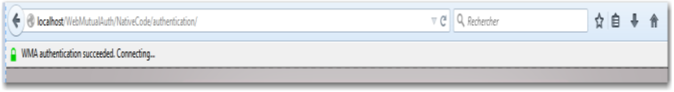
4) The *CCS* execute a series of cryptographic operations to identify the user with authentication service application via a digital signature based on its private key and established a pre-session key (PSK_i). Figure 19 shows a part of the request sent by the extension to the authentication service.

```
http://localhost/WebMutualAuth/NativeCode/authentication/wma.php
POST /WebMutualAuth/NativeCode/authentication/wma.php HTTP/1.1
Host: localhost
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:37.0) Gecko/20100101 Firefox/37.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
…
username=yassine &
encryptedPSK=
qZgN0nXc5xBYsjy5XQwMjCfu%2Bl%2FNp0sRT7qmcqa8Th9C9Wcog
wlaxoQdk1esMTMi9WLANdT67PM
Hs1yrV4Wa0BdUOWpxTfVRJSg2ewn4aBkpuOX52VuxF7IO%2BLdZo
5caJlDSZDJZOS86kP%2F6%2FOYKI6R0m00AeZwrgSk513ornQebd%
2BWY%2FmqJuVgZD4PQVYHa%2Bjr4E2RffLQ4ILZ%2BFxbOoBL%
2BtTkA3Gc6IsyxqL7I7TYzDwcJWejEaaVGHW%2BwGY%2Blxz2Kdh
sqXgcy8zZz78QTUiPH1maXM6oWbhu0PLBZ6%2FZo3k%2FtFsVJOk
CXzp118g0BO8xMaiZ1QnVqu5CVzwy5Q%3D%3D &
 IA = 616a5956765577306558475674 6c4c4d477139586b352b4b36342
f37566a 36796 97 6594865484b374e583438615
7493262693165486 64c53516238654f5a784652714d65697a766d6432504
76e55704a4246427a527855436f454e655131345a773149356c735761523
264496b45474f676d385135555952764c51386667485761 4968327a324f4
e6177422f4235716b395976676676384638707854427 36a7253774d536c4
8756d6d6b622f4134777133775673417 42f784b4c655275455a346751765
1385231667576583858482b52736a384e753946766a572b79433776534e6
76b4d4b3579415651546a397942304932654 94e59796f37612b57776e616
e4232724f5044724b734c536a75413333745a626c3672536a595969526f5
342564c7671446e3239594c5a52504e796d625657726c4f39504459726f5
236757546786c4865734c4b73514a34446c4f5237496848334f413d3d
….
```

Fig. 19. Part of the HTTP request sent by the authentication extension to the web application authentication's service

5) The *SCS* verifies the cryptographic settings of user authentication, and uses the user public key and other session parameters to authenticate mutually with the browser.

6) The extension checks the authenticity of the web application and displays a message indicating that mutual authentication is successful (Figure 20).



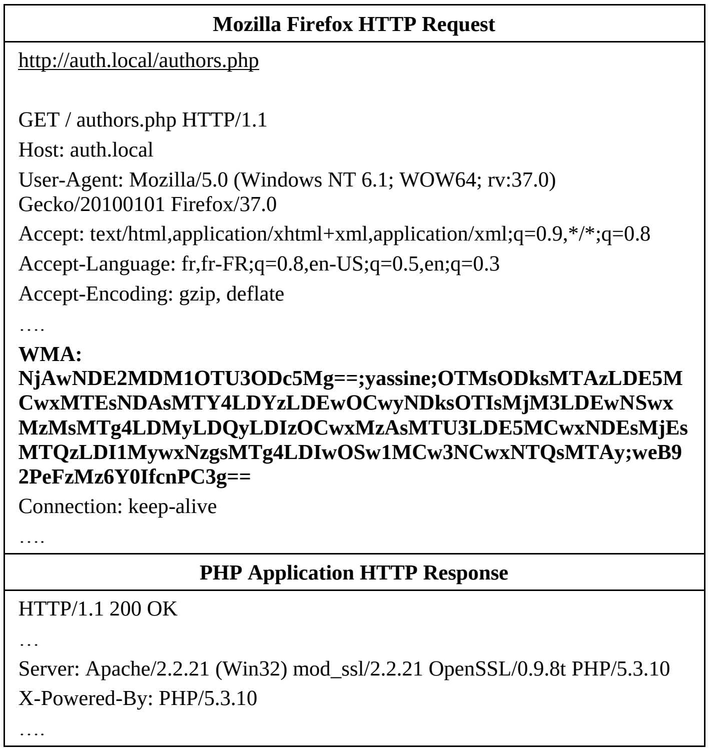
localhost/WebMutualAuth/NativeCode/authentication/

WMA authentication succeeded. Connecting...

Fig. 20. Message in Mozilla Firefox notification bar tells the user that the mutual authentication is executed successfully

After mutual authentication both parties establish a session key (*SK_i*). Since this shared key is known only by the authentication extension and the web application, it cannot be obtained by an attacker. This key is used in the session management phase of the proposed scheme [21]. Specifically, our authentication extension includes an HMAC signature [32], [33] in each HTTP request for the web application. Before sending the requested resources, the application must first validate the HMAC signature using *SK_i*.

Taking inspiration from cookies that use special HTTP headers [34], we decided to create a new HTTP header called "WMA" using the Mozilla Firefox setRequestHeader function. It is a part of the Mozilla Firefox nsIHttpChannel interface [35], which allows to modify the HTTP requests and responses. Figure 21 illustrates an example of a secure session management mechanism. In this example, the user Yassine wants to access the web page "authors.php". For each request that requires authentication, our extension attaches HMAC signature, as well as other parameters in our new HTTP header "WMA". The Web application checks the HMAC signature to authorize or deny access to the requested resources.



Mozilla Firefox HTTP Request

```
http://auth.local/authors.php

GET / authors.php HTTP/1.1
Host: auth.local
User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:37.0)
Gecko/20100101 Firefox/37.0
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Accept-Language: fr,fr-FR;q=0.8,en-US;q=0.5,en;q=0.3
Accept-Encoding: gzip, deflate
….
WMA:
NjAwNDE2MDM1OTU3ODc5Mg==;yassine;OTMsODksMTAzLDE5M
CwxMTEsNDAsMTY4LDYzLDEwOCwyNDksOTIsMjM3LDEwNSwx
MzMsMTg4LDMyLDQyLDIzOCwxMzAsMTU3LDE5MCwxNDEsMjEs
MTQzLDI1MywxNzgsMTg4LDIwOSw1MCw3NCwxNTQsMTAy;weB9
2PeFzMz6Y0IfcnPC3g==
Connection: keep-alive
….
```

PHP Application HTTP Response

```
HTTP/1.1 200 OK
…
Server: Apache/2.2.21 (Win32) mod_ssl/2.2.21 OpenSSL/0.9.8t PHP/5.3.10
X-Powered-By: PHP/5.3.10
….
```

Fig. 21. Successful HTTP request authentication

IV. Experimental Evaluation

In this section, we present the results of the experimental evaluation of our prototype. Indeed, in [21], the authors performed a comparative and theoretical evaluation of the proposed scheme regarding computational complexity. Specifically, they showed that in comparison of related schemes, the proposed scheme is efficient and present the lowest computational complexity. On the one hand, HTTPS is only required in the user registration phase. After that, the other phases are running on HTTP. On the other, as we discussed, even if the initial mutual authentication phase requires expensive cryptographic operations (especially asymmetric cryptography) increasing the computing time, the session management phase will need only a negligible overhead. Thus, the main objective of this evaluation is to confirm the results above by conducting performance tests on our prototype. This performance depends on several parameters; including the ability of the processor and memory, the cryptographic algorithms, and the bandwidth of the network.

A. Materials and Algorithms

The material used in our experiments consists of a server; HP Intel (R) Core (TM) i5-3230M CPU 2.60 GHz with 4 GB RAM running on Windows 7 and a client; laptop Accent Genuine Intel (R) CPU 1.3 GHz with 2 GB RAM Windows 7 Professional Version.

TABLE III. Summary of cryptographic algorithms used in our tests at the server, the PHP application, and Mozilla Firefox

	Apache web server (HTTPS)	**Application PHP**	**Mozilla Firefox**
Asymmetric Cryptography	RSA-2048 bits	RSA-2048 bits	RSA-1024 bits
Symmetric Cryptography	AES-128	AES-128	AES-128
Hash function	SHA-256	SHA-256	SHA-256
HMAC Function	HMAC_SHA256	HMAC_SHA1 HMAC_SHA256	HMAC_SHA1 HMAC_SHA256
Key derivation function	-	-	PBKDF 2 with 1000 iterations.

We use a default configuration for all server and client software (i.e., no performance optimizations). All tests are performed in the context of a Fast Ethernet local area network (flow rates up to 100 Mbit / s). The average ping time on the network was 35.25 ms with a standard deviation of ± 20.013 ms. The cost of the cryptographic processing is evaluated by considering an implementation of the algorithms listed in Table III.

- *RSA keys with 1024 bits and 2048 bits as asymmetric cryptography algorithm*: Since RSA is based on the difficult problem of factoring large numbers, RSA key size is often a very controversial subject. Officially the largest factored number is 768 bits. Therefore, the use of 1024-bit RSA key is considered sufficient to guarantee practical security [36]. Nevertheless, not to be placed just outside the known attack capabilities, security agencies such as NIST and ANSSI recommend in their latest reports using RSA with 2048 bit [36], [37]. Therefore, in our prototype, the PHP application, and the web server uses RSA-2048 bit. In return, we chose a RSA-1024 bit for the users. Indeed, in the proposed architecture, the RSA public key of the user is neither transmitted nor stored in clear during registration and user's authentication. In the client and server sides, the public key is stored encrypted, and its transmission to the web application requires a secure connection (HTTPS). This complicates brute force attacks which require the knowledge of the public part of the RSA key.

- *AES-128 as symmetric encryption algorithm*: Today, AES specified in FIPS 197 [38] is the standard used in most security protocols (TLS, IPsec, etc.). The ANSSI and NIST recommend at least a 100-bit key for data to be protected until 2020, but ANSSI indicates in their report [38] that the use of a 128-bit key is preferable.

- *SHA-256 as a hash function*: Following multiple attacks against the SHA1 algorithm, the majority of applications decided to move to SHA-256 [39].

- *HMAC_SHA1 and HMAC_SHA256 as HMAC function*: While hash functions such as MD5 and SHA-1 are not longer considered safe due to reported collision attacks [40, p. 1], [41]. They may be used in the HMAC functions. HMAC does not require a collision-resistant hash function for its formal security proof [42], [43]. The use of more robust functions like SHA-2 [39] give more security guaranteed, but at a price in the performance level.

- *PBKDF 2 with 1000 iterations as key derivation function*: The use of a key derivation function that requires N iterations to get key increases the calculation cost to perform a dictionary attack on a password with t bits entropy form 2^t operations to 2^t * N operations. Therefore, it makes dictionary attacks and brute force more difficult. However, the computation required for the key derivation by legitimate users also increases with the number of iterations. Thus, there is an obvious compromise: A large number of iterations makes attacks more expensive, but affects performance for the authorized user. PBKDF Version 2 is defined in RFC 2898 [44]. NIST recommends a minimum of 1000 iterations [45].

B. Results and Discussion of the Registration Module

Table IV summarizes the execution time of our registration module, compared with the traditional registration (based on a username and password on HTTPS). The time reported is the average of 10 timings. The total run time includes the time of round trips and networks latency. We calculated the time needed to perform cryptographic operations both client side (Mozilla Firefox registration extension) and PHP applications side. This allowed us to assess the impact of these calculations on performance. As we can see in Table 4, the average total time performance of our proposal is 544.347 ms (with a standard deviation of 113.16 ms). This time is calculated starting the moment the user clicks the Sign up button (Figure 8), after entering their password to the success of this phase (Figure 15). It is clear that our registration module requires more time compared to the traditional registration of users (172.680 ± 32.085 ms). Of course, this is due to the operations integrated into our solution to enhance the security of this phase. Specifically, client-side cryptographic operations require 303.303 ms (± 80.607).

TABLE IV. AVERAGE EXECUTION TIME IN MS (± STANDARD DEVIATION) OF OUR REGISTRATION MODULE, COMPARED WITH THE TRADITIONAL USER'S REGISTRATION

Operation	Our Registration Module	Traditional Registration (Passwords over HTTPS)
Client cryptographic computations	303,303 ± 80,607	-
PHP application cryptographic computations	0,185± 0,018	0,127 ± 0,012
Total runtime	544,347 ± 113,16	172,680 ± 32,085

Figure 22 shows that over 96% of the client time is spent for the RSA key pair generation. Accordingly, the more we increase RSA keys length and the numbers of iterations of PBKDF, performances are affected. At the application side, we obtained almost similar performance to traditional registration. The origin of this small difference is the addition of a generation of a random value and two concatenation operations in our proposal. These are used to avoid storing the unencrypted public key of the user in the application database.

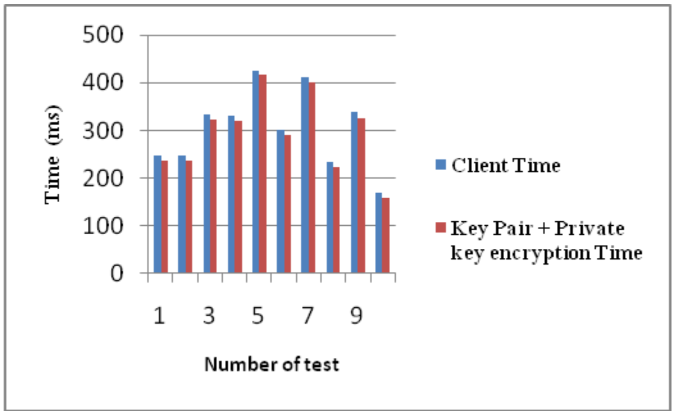


Fig. 22. Comparison of the generation time of 1024 bits RSA keys pair and the time required for other client-side operations

C. Results and Discussion of the Authentication Module

Table V summarizes the results of run time of our authentication module in comparison with traditional authentication (HTML form + HTTPS) and SSL / TLS client certificate authentication. From this table, we can clearly notice that compared to other mechanisms, our authentication module has the lowest total run time (about 148.415 ms 20.315 ms ±) and client certificate authentication SSL / TLS has the highest time (2923.5 ± 350.589). These results confirm those obtained during the theoretical performance analysis in [21].

Indeed, as we have already explained, the authentication phase of the proposed architecture does not require HTTPS, but relies on cryptographic parameters to enhance user authentication over an HTTP connection such as symmetric and asymmetric cryptography. The calculation of these parameters on the client side of our prototype takes 43.862 ms (2,376 ms ± standard deviation) and 15.268 ms (5,701 ms with standard deviation) on the web application side. While the proposed solution has an impact on performance that we think is acceptable the performance compared to password authentication that requires only 0.065 ms in the application-side and no client-side computing; but security always has a cost, and we believe that the price of insecurity is much higher. Also, these calculations are not likely to affect the user experience.

TABLE V. AVERAGE EXECUTION TIME IN MS (± STANDARD DEVIATION) OF OUR AUTHENTICATION MODULE COMPARED WITH AUTHENTICATION BASED ON AN HTML FORM AND A PASSWORD OVER HTTPS AND WITH THE CLIENT SSL / TLS CERTIFICATE AUTHENTICATION.

Operation	**Our Authentication Module**	**SSL/TLS client certificate authentication**
Client cryptographic computations	43,862 ± 2,376	-
PHP application cryptographic computations	15,268 ± 5,701	-
Total runtime	148,415 ± 20,315	2923,5±350,589

Also, as discussed in [21], to create a secure session management mechanism, the proposed scheme attaches an HMAC signature in each HTTP request from the browser to the web application. This ensures the integrity and authenticity of HTTP requests. To assess the impact of this mechanism, we measured the time required to generate an HTTP request signed and time to validate it by the application. Table 6 presents the average time and standard deviations of our prototype in ms when using an HMAC_SHA1 and HMAC_SHA256 functions, compared with traditional session management (the use of cookies sent over HTTPS).

Again, the results in Table VI reaffirm those in [21]. In particular, it is clear to see that the computation time added by generating and validating a HMAC_SHA1 or HMAC_SHA2 is negligible compared to the total execution time (page load time). In other words, the user experience is not affected. Also, despite the cookies by session management requires no cryptographic operations on both the client side and the application side, but the use of HTTPS increases the total execution time required to load a requested page in a user (about 175.680 ms).

TABLE VI. COMPARING THE RESULTS OF BOTH HMAC_SHA1 AND HMAC_SHA256 FUNCTIONS APPLIED DURING THE SESSION MANAGEMENT

	Our authentication Module: Session management phase		**HTTP cookies authentication over HTTPS**
Operation	**HMAC_SHA1**	**HMAC_SHA256**	
Client cryptographic computations	1,846±0,246	1,974±0,24	-
PHP application cryptographic computations	0,0744± 0,036	0,098±0,0283	-
Total runtime	61,64±19,832	65,3593 ± 33,04	175,680 ± 42,124

V. CONCLUSION

In this paper, we demonstrated the implementation feasibility and experimental evaluation of a secure and efficient authentication scheme. We first presented the details of our proposed prototype implementation. Specifically, we separated the prototype in two modules to simplify the implementation process: A registration module that implements the registration phase and an authentication module which incorporates both mutual authentication and session management phases. In each module, the client component of the proposed prototype is developed as an extension of Mozilla Firefox browser that can easily install and the server component as a service of a PHP web application. This allowed us to avoid recompiling the source code of Mozilla Firefox and have an independent server component of the application code; which also facilitated the deployment procedure. After that, we focused on the experimental evaluation of the proposed prototype. Our experimental results confirmed the proposed scheme-theoretical analysis. In fact, even if the registration phases and mutual authentication of our prototype require expensive cryptographic operations (especially asymmetric cryptography) increasing the computing time, the session management phase will need only a negligible overhead. Compared to the related scheme, we showed that the proposed scheme not only improves the usability and deployability, but also improves the user authentication performances.

REFERENCES

[1] K. Fu, E. Sit, K. Smith, and N. Feamster, “The Dos and Don’ts of Client Authentication on the Web.,” in *USENIX Security Symposium*, 2001, pp. 251–268.

[2] M. Weir, S. Aggarwal, M. Collins, and H. Stern, “Testing metrics for password creation policies by attacking large sets of revealed passwords,” 2010, pp. 162–175.

[3] J. Bonneau and S. Preibusch, “The Password Thicket: Technical and Market Failures in Human Authentication on the Web.,” presented at the The Ninth Workshop on the Economics of Information Security, 2010.

[4] J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano, “The quest to replace passwords: A framework for comparative evaluation of web authentication schemes,” presented at the 2012 IEEE Symposium on security and privacy, San Francisco, 2012, pp. 553–567.

[5] J. Bonneau, “The science of guessing: analyzing an anonymized corpus of 70 million passwords,” presented at the 2012 IEEE Symposium on Security and Privacy, San Francisco, 2012, pp. 538–552.

[6] B. Grawemeyer and H. Johnson, “Using and managing multiple passwords: A week to a view,” *Interact. Comput.*, vol. 23, no. 3, pp. 256–267, May 2011.

[7] D. Sandler and D. S. Wallach, “<input type=‘password’> must die,” presented at the Web 2.0 Security & Privacy, 2008.

[8] D. Florencio and C. Herley, “A large-scale study of web password habits,” presented at the the 16th international conference on World Wide Web, New York, 2007, pp. 657–666.

[9] E. Grosse and M. Upadhyay, “Authentication at Scale,” *Secur. Priv. IEEE*, vol. 11, no. 1, pp. 15 – 22, 2013.

[10] D. Florêncio, C. Herley, and B. Coskun, “Do strong web passwords accomplish anything,” presented at the 2nd USENIX Workshop on Hot Topics in Security, Boston, 2007.

[11] S. Grzonkowski, W. Zaremba, M. Zaremba, and B. McDaniel, “Extending web applications with a lightweight zero knowledge proof authentication,” New York, 2008, pp. 65–70.

[12] D. Stuttard and M. Pinto, *The web application hacker’s handbook finding and exploiting security flaws*. Indianapolis: Wiley, 2011.

[13] Y. Sadqi, A. Asimi, and Y. Asimi, “A Cryptographic Mutual Authentication Scheme for Web Applications,” *Int. J. Netw. Secur. Its Appl.*, vol. 6, pp. 1–15, 2014.

[14] M. Manulis, D. Stebila, and N. Denham, “Secure Modular Password Authentication for the Web Using Channel Bindings,” in *Security Standardisation Research*, vol. 8893, L. Chen and C. Mitchell, Eds. Springer International Publishing, 2014, pp. 167–189.

[15] Y. Sadqi, A. Asimi, and Y. Asimi, “A Lightweight and Secure Session Management Protocol,” *Lect. Notes Comput. Sci.*, vol. 8593, pp. 319–323, 2014.

[16] IETF, “RFC 5246 - The Transport Layer Security (TLS) Protocol Version 1.2,” 2008. [Online]. Available: http://tools.ietf.org/html/rfc5246.

[17] M. Wu, S. Garfinkel, and R. Miller, “Secure web authentication with mobile phones,” presented at the DIMACS workshop on usable privacy and security software, 2004, pp. 9–10.

[18] B. Parno, C. Kuo, and A. Perrig, “Phoolproof phishing prevention,” presented at the Financial Cryptography and Data Security (FC’06), Anguilla, British West Indies, 2006.

[19] Google, “About 2-step verification - Accounts Help.” [Online]. Available: https://support.google.com/accounts/answer/180744?hl=en&ref_topic=1099588.

[20] Mozilla, “Mozilla Persona — simple sign-in with email — mozilla.org.” [Online]. Available: http://www.mozilla.org/en-US/persona/.

[21] Y. Sadqi, A. Asimi, and Y. Asimi, “A Secure and Efficient User Authentication Scheme for the Web,” *Int. J. Internet Technol. Secur. Trans.*, vol. 5, no. 1, pp. 43–63, 2015.

[22] Mozilla, “js-ctypes - Mozilla | MDN.” [Online]. Available: https://developer.mozilla.org/en-US/docs/Mozilla/js-ctypes.

[23] Mozilla, “NSS - Mozilla | MDN.” [Online]. Available: https://developer.mozilla.org/en-US/docs/Mozilla/Projects/NSS.

[24] Mozilla, “Storage | MDN.” [Online]. Available: https://developer.mozilla.org/en-US/docs/Storage.

[25] SQLite, “SQLite.” [Online]. Available: https://www.sqlite.org/.

[26] Apache Foundation, “Apache HTTP Server Project.” [Online]. Available: http://httpd.apache.org/.

[27] Apache Foundation, “mod_ssl - Apache HTTP Server Version 2.2.” [Online]. Available: http://httpd.apache.org/docs/2.2/mod/mod_ssl.html.

[28] PHP, “Hash: Hash - Manual.” [Online]. Available: http://php.net/manual/en/book.hash.php.

[29] PHP, “OpenSSL: PHP OpenSSL - Manual.” [Online]. Available: http://php.net/manual/en/book.openssl.php.

[30] Oracle, “MySQL :The world’s most popular open source database.” [Online]. Available: https://www.mysql.com/.

[31] Y. Asimi, A. Amghar, A. Asimi, and Y. Sadqi, “New Random Generator of a Safe Cryptographic Salt per session (RGSCS),” *Int. J. Netw. Secur.*, vol. 18, no. 3, pp. 445–453, 2016.

[32] M. Bellare, R. Canetti, and H. Krawczyk, “Message authentication using hash functions: The HMAC construction,” *RSA Lab. CryptoBytes*, vol. 2, no. 1, pp. 12–15, 1996.

[33] RFC 2104, “HMAC: Keyed-Hashing for Message Authentication,” 1997. [Online]. Available: https://www.ietf.org/rfc/rfc2104.txt.

[34] IETF, “RFC 6265 - HTTP State Management Mechanism,” 2011. [Online]. Available: http://tools.ietf.org/html/rfc6265. [Accessed: 08-May-2015].

[35] Mozilla, “nsIHttpChannel - Mozilla | MDN.” [Online]. Available: https://developer.mozilla.org/en-US/docs/Mozilla/Tech/XPCOM/Reference/Interface/nsIHttpChannel. [Accessed: 08-May-2015].

[36] ANSSI, “Référentiel Général de Sécurité: Règles et recommandations concernant le choix et le dimensionnement des mécanismes cryptographiques,” 2014.

[37] E. Barker and A. Roginsky, “Transitions: Recommendation for transitioning the use of cryptographic algorithms and key lengths,” *NIST Spec. Publ.*, vol. 800, p. 131A, 2011.

[38] NIST, “FIPS 197: ADVANCED ENCRYPTION STANDARD (AES),” *Process. Stand. Publ.*, Nov. 2001.

[39] NIST, “FIPS 180-4: Secure Hash Standard (SHS),” *Process. Stand. Publ.*, 2012.

[40] X. Wang, Y. L. Yin, and H. Yu, “Finding collisions in the full SHA-1,” in *Advances in Cryptology–CRYPTO 2005*, 2005, pp. 17–36.

[41] X. Wang and H. Yu, “How to break MD5 and other hash functions,” in *Advances in Cryptology–EUROCRYPT 2005*, Springer, 2005, pp. 19–35.

[42] S. Turner and L. Chen, “RFC 6151 - Updated Security Considerations for the MD5 Message-Digest and the HMAC-MD5 Algorithms,” 2011. [Online]. Available: https://tools.ietf.org/html/rfc6151.

[43] M. Bellare, “New proofs for NMAC and HMAC: Security without collision-resistance,” in *Advances in Cryptology-CRYPTO 2006*, Springer, 2006, pp. 602–619.

[44] RSA Laboratories, “PKCS #5 v2.1: Password-Based Cryptography Standard,” 2000. [Online]. Available: https://www.ietf.org/rfc/rfc2898.txt.

[45] M. S. Turan, E. Barker, W. Burr, and L. Chen, “Recommendation for password-based key derivation,” *NIST Spec. Publ.*, vol. 800, p. 132, 2010.

Indoor Navigation System based on Passive RFID Transponder with Digital Compass for Visually Impaired People

A. M. Kassim, T. Yasuno and H. Suzuki
Graduate School of Tokushima University,
2-1 Minamijosanjima, Tokushima, 770-8506, JAPAN

H. I. Jaafar and M. S. M. Aras
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, Durian Tunggal, 76100 Melaka, MALAYSIA

***Abstract*— Conventionally, visually impaired people using white cane or guide dog for traveling to desired destination. However, they could not identify their surround easily. Hence, this paper describes the development of navigation system which is applied to guide the visually impaired people at an indoor environment. To provide an efficient and user-friendly navigation tools, a navigation device is developed by using passive radio frequency identification (RFID) transponders which are mounted on the floor such as on tactile paving to build such as RFID networks. The developed navigation system is equipped with a digital compass to facilitate the visually impaired people to walk properly at right direction especially when turning process. The idea of positioning and localization with the digital compass and direction guiding through voice commands is implemented in this system. Some experiments also done which is focused on the calibration of the digital compass and relocates the visually impaired people back to the right route if they are out of the direction. Besides, a comparison between two subjects which are human and a mobile robot is made to check the validity of the developed navigation system. As the result, the traveling speed of human and mobile robot is obtained from the experiment. This project is beneficial to visually impaired people because the navigation device designed with voice commands will help them to have a better experience, safer and comfortable travel.**

***Keywords*—navigation system, passive RFID, digital compass, visually impaired people**

I. Introduction

World Health Organization (WHO) has released the statistics in 2014 that shows 10% of the world's population have a disability, with 80% of them located in developing countries. This global data on visual impairments shows the people with visual impairment globally is around 285 million which 39 million are fully blind and 246 million are having low vision problem[1]. There are great numbers of people worldwide who have encountered vision loss including Malaysia. This total figure does not reflect to the real number of a people with disabilities in this world since the visually impaired people are cannot independently travel by themselves.

Moreover, a disability that may involve additional danger to the individual is blindness. Conventionally, they rely on a white cane or a guide dog to assist them in reaching the desired destination safely. It has been decades since visually impaired people use the white cane as their most common and affordable assistive tool to detect obstacles and path surrounding them. Difficulties still occur when using the white cane for visually impaired people as they are only able to detect path and obstacles from the front by swinging the white cane at the same time trying to feel the tip of the white cane that touches the ground.

However, this approach is useful if the passage to the desired place is already known to them. It also becomes troublesome once the destination is newly constructed and not implemented the universal design, especially on visually impaired people. They do not receive enough information with only the tip of the white cane as feedback. Lack of aid signs built for visually impaired people seems to be one of the difficulties for them. It is hard and nearly impossible to recognize the place by themselves and travel from one to another destination without proper navigation tools. Therefore, the advanced technology developed by researcher benefits the visually impaired people to move independently.

In these two decades, there are quite some equipments, tools or robots which have been developed by the researchers in this world to assist the visually impaired people. The assistive and rehabilitation technologies that have been researched and built are such as GuideCane[2], NavBelt[3], My 2nd Eye[4], SMART EYE [5], [6], and others. Besides that, one of the developed mobile robots for the blind which is designed to help the blind in shopping malls [7]. Without the state of the art of these technologies, the visually impaired people only counts on the conventional white cane to detect surrounding obstacles and sense the road in front of them.

The arrangement of this paper is done as follows where Section 1 presents an introduction on problem and challenge that have been faced by visually impaired people. Section 2 expresses previous studies related to the travel aid for visually impaired people. Besides, section 3 deliberates on the developed navigation system with the proposed control system for navigation purpose. Section 4 discloses on experimental setup involved in this study while Section 5 elaborates the results that obtained through the developed navigation system and the proposed control approach, and lastly the conclusion and future tasks of this study.

II. Related Works

The design challenges for the blind navigation device are real-time guidance, portability, power limitations, appropriate interface, and continuous availability, no dependence on infrastructure, low cost solution and minimal training [8]. There are several designs to help visually impaired people to achieve self-independence traveling at indoor environment. Location technology such as infrared data association (IrDA), RFID, Bluetooth or Wi-Fi has been developed to help them travel during indoors with contextual information or sound navigation [9] .

In addition, the usage of Global Positioning System (GPS) device also can help to guide the visually impaired people in an outdoor environment. Since GPS cannot function properly in indoor space, other researchers present the solution by using IrDA technology which works as detector to guide in indoor environment[10]. The Drishti system is the combination of GPS for outdoor navigation and ultrasonic sensor for indoor navigation[11]. One of the problems about GPS is the error with the measurement taken especially inside the tall buildings [12].

Furthermore, BLI-NAV a blind navigation system designed which consists of GPS receiver and path detector. Both devices used to detect user's location and determine the shortest route to destination. Voice command is given throughout the travel. Path algorithm is used to determine the shortest distance from start point to end point together with path detector. Moreover, user is able to avoid obstacles while traveling [13]. This system gave better results in real time performance and improves the efficiency of blind travel at indoor environment.

On the other hand, Pocket-PC based Electronic Travel Aid (ETA) proposed to help visually impaired people to travel at indoor environment. Pocket-PC will alert the user when near the obstacles through warning audio [14]. An ultrasonic navigation device for visually impaired people is designed. The micro-controller built in the device can guide the user in which route should be taken through speech output. Besides, the device helps to reduce navigation difficulties and an obstacles detection using ultrasounds and vibrators. Ultrasonic range sensor is used to detect surrounding obstacles and electronic compass is used for direction navigation purpose. Stereoscopic sonar system is also used to detect nearest obstacles and it feeds back to tell user about the current location [15] $\sim$ [17].

In, addition, Blind Assistant Navigation System that can help visually impaired people navigates independently at indoor environment also developed [18]. The system provides the localization by using wireless mesh network. The server will do the path planning which then communicates using wireless with the portable mobile unit. The visually impaired people can give commands and receives response from the server via audio signals using a headset with a microphone[19]. The proposed RFID technology in order to design the navigation system by providing information about their surroundings also developed. The system uses the RFID reader that mounted on end of the stick to read the transponder tags that are installed on the tactile paving[20], [21].

Besides, INSIGHT is the indoor navigation system to assist the visually impaired people to travel inside the buildings. The system used the RFID with Bluetooth technology to locate the user inside the buildings. The peopleal digital assistant (PDA) such as a mobile device used to interact with INSIGHT server and provide navigation information through voice commands. The zone that the user walked will be monitored by the system. The system will notify the user if the user travels the wrong direction [22]. The RFID network can help to determine the shortest distance from current location to the destination. Besides the system can help to find the way back if they lost their direction and recalculate the new path [23].

In this paper, the development of navigation system which is applied to guide the visually impaired people at indoor environment. In order to provide an efficient and user-friendly navigation tools, a navigation device is developed by using passive RFID transponders which are mounted on the floor such as on tactile paving to build such RFID networks. We also equip the navigation system with digital compass to travel properly at true direction especially when turning process. The idea of positioning and localization with a digital compass and direction guiding through voice commands is implemented in this system. Some experiments are conducted which is focused on the calibration of the digital compass and relocates the visually impaired people back to the normal route if they are out of the direction. Besides, a comparison between two subjects which are human and a mobile robot is done to check the validity of the developed navigation system. As the results, the traveling speed of human and mobile robot is obtained from the experiment. This project is beneficial to visually impaired people because the navigation device designed with voice commands will help them to have better experience, safer and comfortable travel.

III. Developed Navigation System for Visually Impaired people

A. System construction

In order to developed the navigation system which is benefit to the visually impaired people, the total system such as a path planning system, RFID detection system, obstacle detection system is needed. However, in this paper, the developed navigation system is only one part of total navigation system which not including path planning system, obstacle avoidance system, and localization system and etc. Here, the developed system is focused on the RFID detection system and the digital compass which is used to guide the right way for visually impaired people when travel alone. Figure 1 illustrates the system configuration of one part of navigation system including the RFID detection system and the digital compass. In the developed navigation system, there are some components, is used such as RFID reader/writer module, micro-controller, voice module, Braille keypad, digital compass and etc.

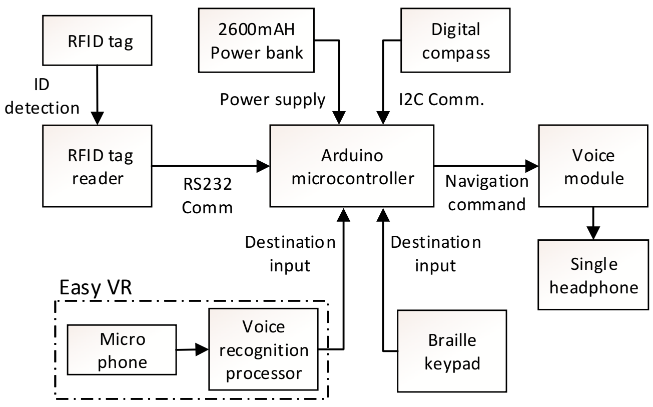


Fig. 1. System overview of developed navigation device for visually impaired people

As the main part in the developed navigation system, a micro-controller is installed which consists memory and program in order to communicate with other peripheral. The micro-controller type which is selected to be implemented in this project is an Arduino Promini. The reason is because the Arduino Promini has been developed in small size package, light on weight and has adequate I/0 port in order to construct the navigation device. In addition, the digital compass (HMC6352) is used because it can provide the high degree heading resolution and accurate in determining the direction. If the user travels out from the path, the navigation system will determine the direction heading and gives alert to user. Once the direction is correct, the user can continue their travel by the aid of audio navigation. On the other hand, the RFID reader/writer module manufactured by Parallax also plays an important part in the developed navigation system. The RFID reader/writer module is installed at the bottom of a retractable cane for the easy detection of passive RFID tags installed on tactile paving. The RFID reader/writer module could detect the RFID transponder tags at 125 kHz up to 3 inches distance.

Figure 2 shows the RFID detection system consists of the RFID reader/writer module. The RFID reader/writer module is directly connected to Arduino micro-controller which can be activated when the Arduino micro-controller is powered by the power supply/battery. It is also installed at the bottom of the electronic cane in order to detect the RFID tag easily which are installed on the tactile paving that can read the code of the tags and the code encryption will be done by the program inside the Arduino micro-controller. The information of the places which the RFID tags have been mounted will be prepared as a library inside the micro-controller. Each RFID tag contains pre-stored information such as the location and the surrounding environment including obstacles, place names, and building names in the library of the micro-controller with the micro SD card which has been installed with the voice module. The voice module (WTV020) is used in order to play the voice commands and inform the users which the direction that they should be taken and turn when the corner. It also played the voice command that followed by the measured value from digital compass.

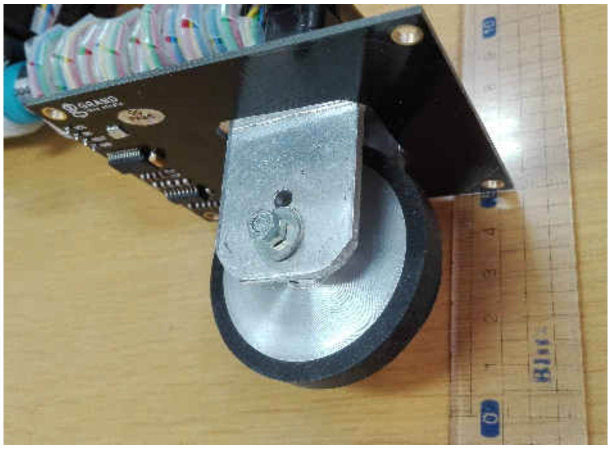


Fig. 2. Construction of RFID reader/writer module on developed electronic cane

Figure 3 shows the illustration of the electronic cane which is developed in this research. A conventional white cane is transformed to the electronic cane in order to attached all the developed system. At the bottom of the cane, a wheel is mounted in order for the RFID reader/writer module can easily detect RFID tags. The wheel size is about 4cm and the RFID reader/writer module is mounted at 6cm from the floor. Thus, the RFID tags can be detected respectively. At the same time, a wheel make the user does not need to raise and swing the cane while traveling which can makes the user tired to swing. If the user swing the retractable cane, the RFID tag could not be detected by the RFID reader/writer module. In addition, the retractable stick is used as the replacement of conventional white cane which is commonly used by visually impaired people for navigation purpose. The retractable stick can be shorten to 25cm in order to carry and mobilize when the developed navigation device unused. Besides, the retractable stick can be extended up to maximum 120cm which similar to conventional white cane. Hence, the retractable stick can be adjusted on the basis of the height and users comfortability.

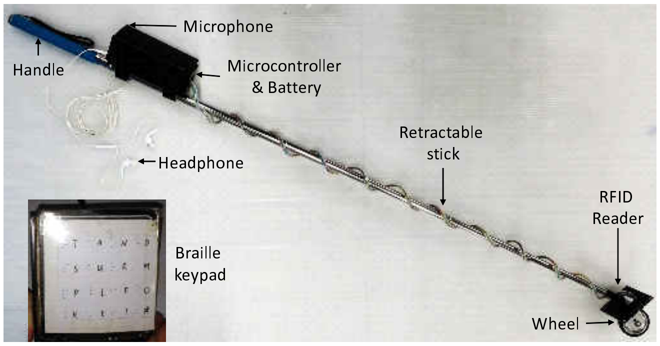


Fig. 3. Developed navigation device for visually impaired people

B. Control system flow chart for navigation system

Figure 4 shows the process flowchart of the designed and developed navigation device. In order to used the developed navigation device, the RFID reader/writer module needs to be initialized respectively. The user could set the desired destination at start point after detecting the nearest RFID tag which have been mounted on tactile paving or floor around them. In order to navigate to the desired location. The RFID reader/writer module is activated and will read the transponder tags. If the RFID reader/writer module is failed to read the tag even though the RFID tag is already in the detection range, the device will return to the tag reader activation process by initializing the RFID reader/writer module first. On the other hand, if the next RFID tags are successfully detected, the Arduino micro-controller will carry out the encryption of the tag identity. Each RFID tag has its own identification code and the code is transferred to Arduino micro-controller through RS232 for identification code encryption. After that, the micro-controller system proceeds to the navigation processing. The information extracted will be then reprocessing and converted into voice commands. Then, the users will receive the commands on how to travel their path and the information about surrounding environment.

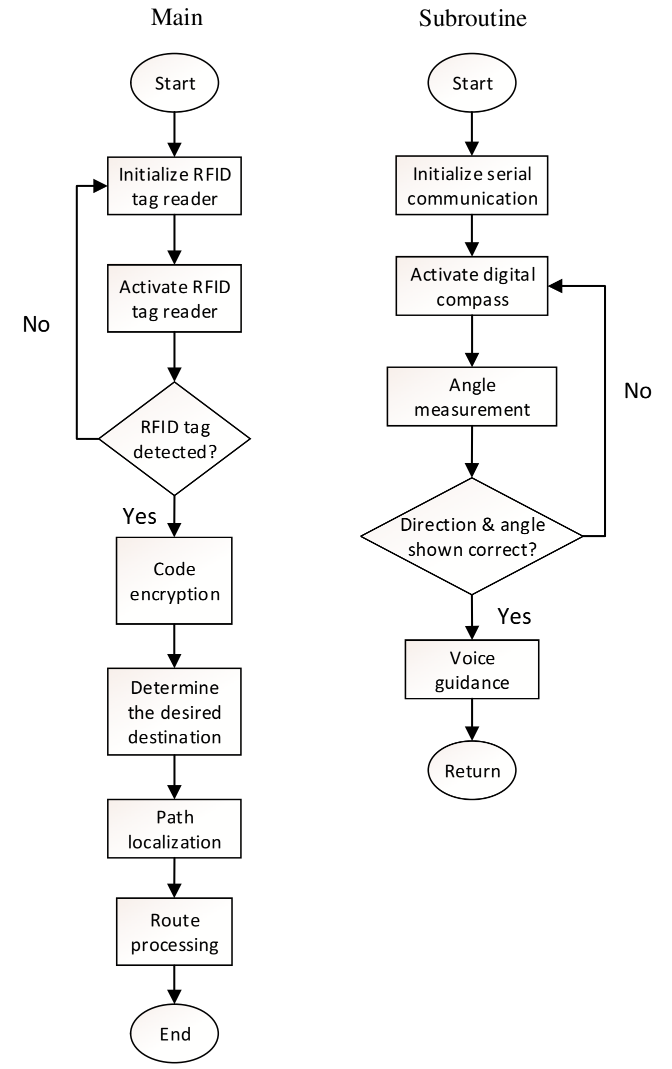


Fig. 4. Process flowchart of RFID navigation system

Once the RFID tag is detected at the starting point, the user can determine where he or she wants to go by inputting the desired location through the developed Braille keyboard or voice recognition device [24]. Figure 5 shows the modified keypad with Braille code. This Braille writing system is attached and connected on the top of the 4 x 4 numeric keypad. It is special modified for key in the destination with the combination alphabet " T, A, N, D, S, U, R, M, P, L, F, O, I, E, L, # " which can be inputted as desired destination such as TOILET, ATM, ROOM, RESTAURANT, STORE, PLATFORM and etc. The reason why these alphabets was selected in the prototype is because a simple map just need to be applied. All the desired destination have been pre-programmed in the library of the micro-controller. The users need to key in the destination on the modified keypad with the Braille code and the device will start to give the guidance to the visually impaired people through a headphone.

After the identity encryption process is succeed, the device process to path localization will lead to user's desired destination guided by voice commands which will be given through headphone. In case the user travels at the wrong path especially when the user turn at the junction, the process will proceed to route processing subroutine. In the route processing subroutine, the digital compass is activated by initializing the serial communication from the Arduino main board. The magneto sensor inside the digital compass will measure the angle deviation and recalculate the correct heading direction. If there is no more angle deviation, the user can continue his or her travel through voice commands until arrive to the destination and vice-versa.

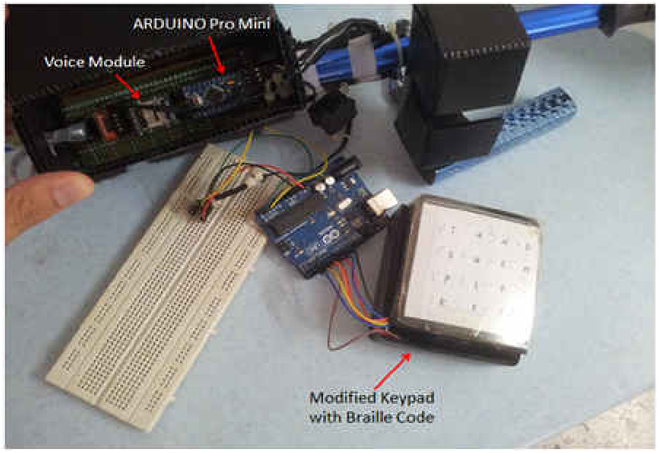


Fig. 5. The modified keypad using Braille characters

IV. Experimental Setup

Figure 6 illustrates the ZigBee wireless networks where the communication between the server/laptop and the developed navigation device. The ZigBee network is applied to connect and monitor the developed navigation device when the experiment is conducted. The ZigBee network acts as the center of data transferring between the navigation device and the server/ laptop. The movement of user will be shown to the map processing system on the server/laptop respectively such as in Fig. 6. Hence, the user current position to the desired position will be displayed on the map based on a generated route. The map system then identifies the address of the target. Concurrently, the RFID reader/writer module will read the RFID tags on the tactile paving or floor. The data of the RFID tags of the current position and the address is sent for the map processing.

Next, voice guidance commands also will be given based on the route which have been generated to the user through an earphone. The earphone connection is based on Bluetooth connection. The server/laptop will send the voice guidance, and user position also will be updated at the same time. Path recalculation also will be done again and produces the voice guidance if the user takes the wrong path from the recommended path. The benefits of the system is when user need to take the corner turning, the digital compass will compare the angle and ensure the user to take the corner effectively without hitting the nearby obstacles. The server receives data from ZigBee network and suggests to mount at fixed locations inside the buildings. The server must be updated and the information of the destinations and objects need to be stored inside the database with respect to the map system.

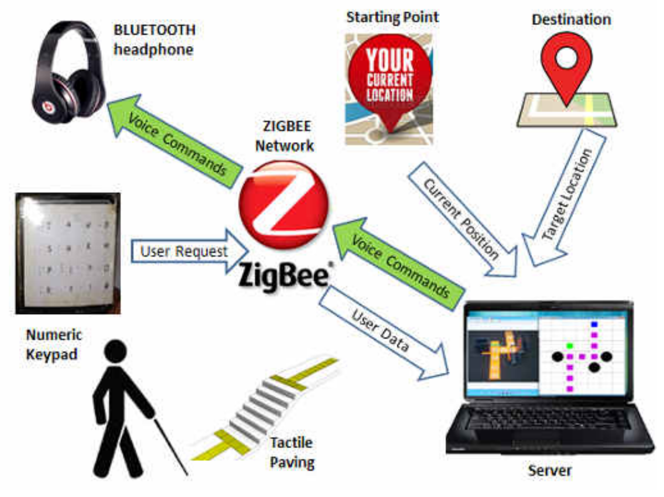


Fig. 6. System configuration including developed navigation system with server/laptop

In order to optimize the functionality of the developed navigation device for guiding the visually impaired people in the correct direction throughout the travel path, the experiment setup to evaluate the accuracy of the digital compass is set. The orientation or direction is attained by using digital compass which mounted on the developed electronic cane. Figure 7 shows the digital compass setup and the reference compass, respectively. The digital compass is connected to the Arduino Promini micro-controller to obtain the analog signal and convert it back to the digital signal by using the on-board analog digital converter (ADC). The digital signal will be displayed on the serial monitor of the Arduino Promini and the digital compass can be tuned accurately. The digital compass is fixed at the certain point where the RFID tag has been mounted to ensure the digital compass is always pointing at the north (N). The compass which inside the iPhone is used as the reference compass in order to make comparison when calibrating the digital compass.

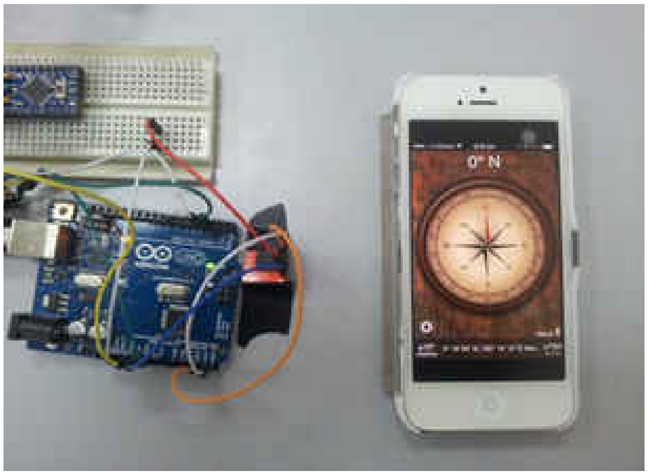


Fig. 7. Digital compass setup and the reference compass

Figure 8 shows the experimental field including tactile paving which is used for the experiment. Tactile pavings are numbered with 00, 01, 02, 03, 04, 05, 06 and 07 are the points paths while blue objects represent the obstacles which is unable to pass through. Voice module (WTV020) has been saved with five types of sound which are forward, turn left, turn right, stop and warning. At this stage, only five directions of sound are used as navigation after the shortest path is generated to follow the direction from start to target node. There are two destinations which can be set in this experiment which are ATM and toilet. There are some influence factors need to be considered during the blind navigation evaluation test such as systematic error and human error.

Systematic error where there is bias in measurement lead to the path completion time. The error occurred for the time response of the system where there is some time delay for the ZigBee during data transmission and sending voice commands. Human error is another error that is not intended and cannot avoided. For this case, the human error is response of the participants when they start to walk when they received the voice commands. There are some delay at the starting point and cornering. This will give the different results for the path completion time. However, the time that is needed to complete the path does not emphasize too much on how fast the people reaches the destination.

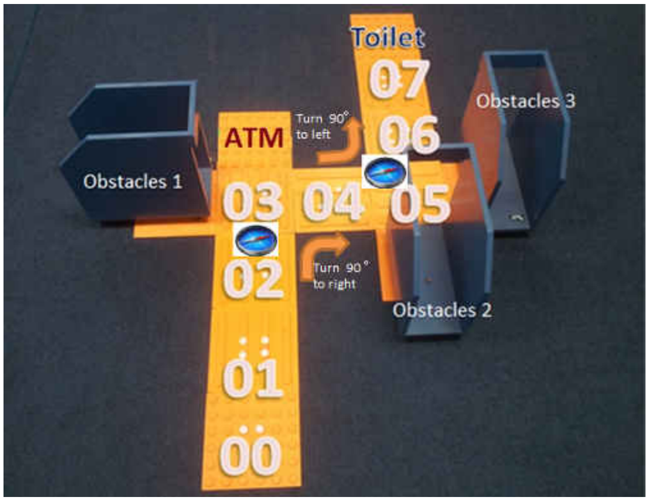


Fig. 8. Field setup which include RFID tags on tactile paving with some obstacles

V. Experimental Results

A. Performances of compass technology in navigation system

Table I shows the digital compass accuracy test results when the digital compass is pointing to north(N). From the table, the digital compass measurement clearly shows that the resolution of the digital compass is high and able to produce the digital compass reading in two significant values. The repeatability test is carried out about 20 times to proof that the results are valid to be used in the navigation device. Besides, the accuracy test for the digital compass have been done 20 times at different places and different time. The relative error is getting smaller and close to the north(N) direction which is shown in the result because the digital compass output are getting stable. This implies that the digital compass is suitable to use for blind navigation to provide accurate heading direction.

TABLE I. COMPASS ACCURACY TEST RESULT

Times	Pointing to North (N)	Degrees North(N)	Relative error	Percentage relative error
1	Yes	356.30	3.70	1.0278
2	Yes	356.40	3.60	1.0000
3	Yes	356.30	3.70	1.0278
4	Yes	355.70	4.30	1.1944
5	Yes	358.20	1.80	0.5000
6	Yes	357.80	2.20	0.6111
7	Yes	357.70	2.30	0.6389
8	Yes	357.80	2.20	0.6111
9	Yes	358.20	1.80	0.5000
10	Yes	357.90	2.10	0.5833
11	Yes	359.20	0.80	0.2222
12	Yes	359.30	0.70	0.1944
13	Yes	359.20	0.80	0.2222
14	Yes	359.30	0.70	0.1944
15	Yes	359.30	0.70	0.1944
16	Yes	359.20	0.80	0.2222
17	Yes	359.20	0.80	0.2222
18	Yes	359.60	0.40	0.1111
19	Yes	359.40	0.60	0.1666
20	Yes	359.40	0.60	0.1666

Note : North(N) direction point to 0 ° or 360 °
Mean of 20 times repeatability = 358.27 °
Mean of percent relative error = 0.4805%

Figure 9 shows the percent relative error % of the readings obtained from the digital compass when it is pointing to the north (N). The maximum peak of the percent relative error is 1.1944% . The average percent relative error is 0.4805% . The graph shows the percent relative error is decreasing gradually towards the end and becomes nearly constant between the 11 and 17 times of trials. This implies that the relative error is getting smaller and the readings are very close to 360° when the digital compass points to the north (N). The obtained data is said to have high reliability. Besides, the digital compass has very high sensitivity and able produces significant value at tenth decimal places.

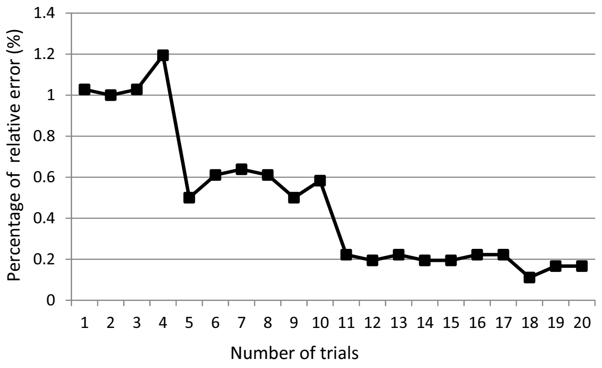


Fig. 9. Percentage of relative error for digital compass calibration when pointing to north

B. *Performance of developed navigation system in field test*

From the table which shown the validity of the digital compass which can be used inside the developed navigation system, the developed navigation system is evaluated for its performance at the real field by using the RFID tag that have been installed on tactile paving. The navigation device built with the digital compass for direction guidance and voice module able to inform user about the direction. The digital compass will be able to detect the error if the user travels out of the direction and the misdirection lead to the wrong path. Thus, the voice module will inform the user, " WARNING! " repeatedly and alert user from taking the wrong path. The navigation proceeds until the user turn over and travel on the right direction. Figure 10 shows the illustration of the developed navigation device which is conducted on different subjects such as human and mobile robot.

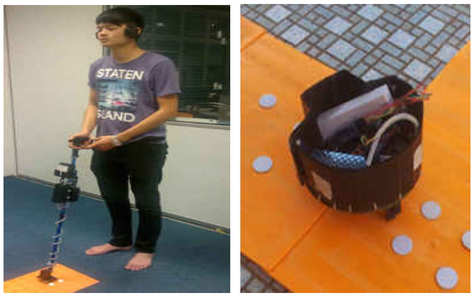

(a) Human (b) Mobile robot

Fig. 10. Experiment conducted for comparing the performance of navigation device based on different subjects such as human and mobile robot

Here, the traveling time which is consumed to finished the route is measured. The travel distance which is needed to be done is 240cm which each tactile paving about 30cm in length. There are two subjects which have been tested at the field which are human and mobile robot. The RFID reader/writer module is attached at the bottom of the mobile robot in order to easily detect the RFID tag. Then, RFID reader/writer module reads the tags first and at the same time the Arduino Promini micro-controller will synchronize with the digital compass for route processing. For experiment by using a mobile robot, the command are given to the DC motor in order to go straight or turn. However, for human, the voice module will inform the user how to turn at the corner, e.g. " 90 degree turn left " or " 90 degree turn right " . Through this experiment, the high accuracy of the digital compass is able to give direction fast and precise.

Table II shows the average traveling time which is recorded for mobile robot is 25s while for human is 33s. From these results, the difference which can be related is the size of the subject. The mobile robot which is used are wheeled robot which sized 20cm in diameter. However, the human subject which are tested in this experiment are 170cm in height and the position of human is 40cm behind the end of the electronic cane. Therefore, the human subject is quite difficult to turn when the voice command is given. Meanwhile, the mobile robot is easily turn because the mobile robot can turn at the same axis when the command which are given to the DC motor, respectively.

Figure 11 shows the picture from the video for human when the experiment is conducted. These pictures is taken at every second for all time elapsed which take 54s to complete the course that have been set. In order to set the desired destination which is TOILET by using the developed Braille keypad, the human subject took about 21s from the starting point. Then, the human subject took about 33s to travel from the starting point to the desired destination. The human subject travelled by using voice guidance from the developed navigation device such as forward, turn right, turn left, and etc.

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT SUBJECTS

Subject	Human with electronic cane	Mobile robot
Size	170cm (Height)	20cm (Diameter)
Average travelled time	33s	25s

VI. CONCLUSION

In this paper, the navigation device for visually impaired people has been developed and some experiments has been evaluated. The digital compass which is applied can provide accurate direction which help to guide the visually impaired while travel independently. The RFID detection system is beneficial to the visually impaired people since these people provided by feedback information about the current location and the surrounding obstacles. Besides, the performance evaluation for the human and robot subject also successfully done.

Since the developed navigation device was the earlier stage of development, the implementation of the shortest path algorithm will be applied to search the shortest route in the future. The blind navigation device with RFID technology supported by shortest path algorithm will be conducted to ensure them to have a better and comfortable travel at the indoor environment. Besides, the design of the navigation device also will be improved in terms of weight and sustainability design concept.

ACKNOWLEDGMENT

This research is a collaboration project between Tokushima University and Universiti Teknikal Malaysia Melaka. This research also was supported by research grant from Universiti Teknikal Malaysia Melaka award no. FRGS/1/2014/TK03/FKE/03/F00213.

REFERENCES

[1] World Health Organization, “ WHO — Visual Impairment and Blindness. ” Fact Sheet No. 282, 2014

[2] J. Borenstein and I. Ulrich, “ The GuideCane-a computerized travel aid for the active guidance of blind pedestrians, ” Proceedings of International Conference on Robotics and Automation, vol. 2, pp. 1283–1288, 1997.

[3] S. Shoval, I. Ulrich, and J. Borenstein, “ NavBelt and the Guide-Cane [obstacle-avoidance systems for the blind and visually impaired] ” IEEE Robotics and Automation Magazine, vol. 10, no. 1, pp. 9– 20, 2013

[4] A.M Kassim, Jamaluddin, M.H., Yaacob, M.R., Anwar, N.S.N., Sani, Z.M., Noordin, A., “ Design and development of MY 2nd EYE for visually impaired people ” 2011 IEEE Symposium on Industrial Electronics and Applications (ISIEA), pp. 700–703, 2011

[5] A.M Kassim, M.S Jamri, M.S.M Aras, M.Z.A Rashid, MR Yaacob, “ Design and development of obstacle detection and warning device for above abdomen level ” 2012 12th International Conference on Control, Automation and Systems (ICCAS), pp. 410–413, 2012

[6] A.M Kassim, A.Z Shukor, C.X Zhi, T Yasuno, “ Performance Study of Developed SMART EYE for Visually Impaired people ” Australian Journal of Basic and Applied Sciences, 7 (14) pp. 633–639, 2013

[7] V. Kulyukin, C. Gharpure, and J. Nicholson, “ RoboCart: toward robot-assisted navigation of grocery stores by the visually impaired, ” 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2845–2850, 2005.

[8] E. Wise et al., “ Indoor Navigation for the Blind and Vision Impaired: Where are we and where are we going?, ” 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–7, 2012

[9] U. Biader Ceipidor, C.M. Medaglia, F.Rizzo, A.Serbanati, “ Radio Virgilio/Sesamonet: an RFID-based Navigation system for visually impaired ” pp.1–6,

[10] Jaime Sanchez, “ Mobile Audio Navigation Interfaces for the Blind ” , Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments, Vol. 5615 of the series Lecture Notes in Computer Science, pp. 402–411, 2009

[11] Lisa Ran, Sumi Helal and Steve Moore, “ Drishti: An Integrated Indoor/Outdoor Blind Navigation System and Service “ Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications, PerCom 2004, pp. 23–30, March 2004

[12] M. Sarfraz and S. A. J. Rizvi, “ Indoor Navigational Aid System for the Visually Impaired, ” Geometric Modeling and Imaging (GMAI), pp. 127–132, 2007

[13] S. S. Santhosh, T. Sasiprabha, and R. Jeberson, “ BLI-NAV embedded navigation system for blind people, ” Recent Advances in Space Technology Services and Climate Change (RSTSCC) 2010, pp. 277–282, 2012

[14] M. H. Choudhury, D. Aguerrevere, and A. B. Barreto, “ A Pocket-PC Based Navigational Aid for Blind Individuals, ” 2004 IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS), pp. 43–48, 2004

[15] M. Bousbia-Salah, A. Redjati, M. Fezari, and M. Bettayeb, “ An ultrasonic navigation system for blind people, ” 2007 IEEE International Conference on Signal Processing and Communications ICSPC, pp. 1003–1006, 2007

[16] A.M Kassim and M.S Jamri and M.S.M Aras and M.Z.A Rashid, “ Design and Development of Vibration Method for Vehicle Reverse System (VRS) ” Journal of Procedia Engineering, vol. 41, pp. 1114–1120, 2012

[17] M.R Yaacob, N.S.N Anwar, A.M Kassim , “ Effect of Glittering and Reflective Objects of Different Colors to the Output Voltage-Distance Characteristics of Sharp GP2D120 IR ” ACEEE International Journal on Electrical and Power Engineering. 3 (2). pp. 6–10, 2012

[18] M. Shamsi, M. Al-Qutayri, and J. Jeedella, “ Blind assistant navigation system, ” 2011 1st Middle East Conference on Biomedical Engineering (MECBME), pp. 163–166, 2011

[19] S. Chumkamon, P. Tuvaphanthaphiphat, and P.Keeratiwintakorn, (2008) “ A blind navigation system using RFID for indoor environments, ” 5th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology, vol. 2, pp. 765–768.

[20] A.M. Kassim, H. I Jaafar, M.A. Azam, N. Abas, T.Yasuno, “ Performances study of distance measurement sensor with different object materials and properties ” 3rd IEEE International Conference on System Engineering and Technology (ICSET), pp. 281–284, 2013

[21] A.M. Kassim, H. I Jaafar, M.A. Azam, N. Abas, T.Yasuno, “Design and Development of Navigation System by using RFID Technology ” 3rd IEEE International Conference on System Engineering and Technology (ICSET), pp. 258–262, 2013

[22] A., Gandhi, S. R., Wilson, C., and Mullett, G., “ INSIGHT: RFID and Bluetooth enabled automated space for the blind and visually impaired ” IEEE International Conference of Engineering in Medicine and Biology Society, pp. 331–334, 2010

[23] A.M Kassim, A.Z Shukor, C.X Zhi, T Yasuno, “ Exploratory Study on Navigation System for Visually Impaired people ” Australian Journal of Basic and Applied Sciences, 7 (14) pp. 211–217, 2013

[24] Anuar bin Mohamed Kassim, Takashi Yasuno, Hazriq Izzuan Jaafar, Mohd Aras Mohd Shahrieel, “ Development and Evaluation of Voice Recognition Input Technology in Navigation System for Blind people ”, Journal of Signal Processing, Vol.19, No.4, pp.135-138, July 2015

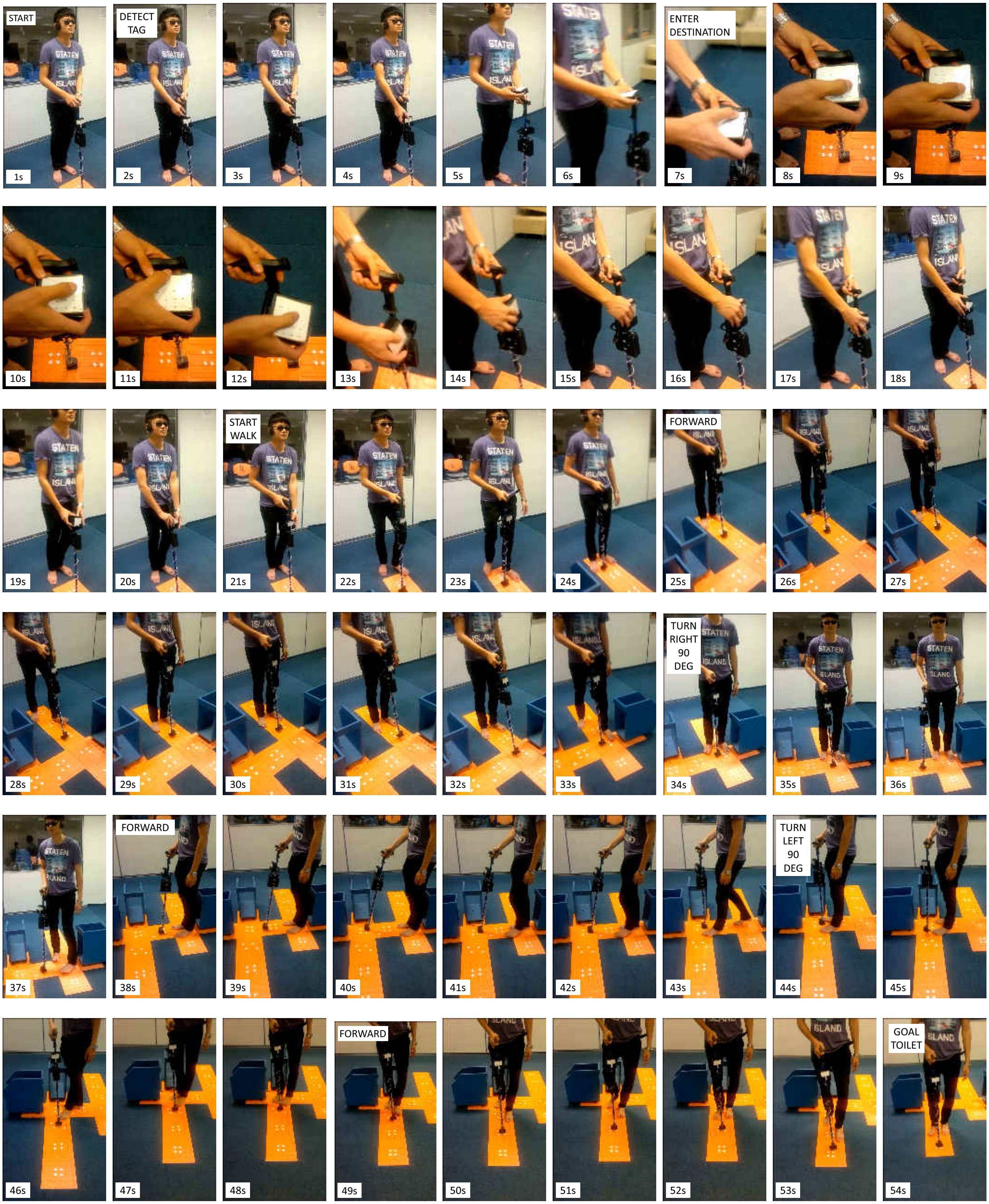


Fig. 11. Pictures for each second by human subject travelled using developed navigation device with voice guidance

Smart Cities: A Survey on Security Concerns

Sidra Ijaz, Munam Ali Shah, Abid Khan and Mansoor Ahmed
Department of Computer Science
COMSATS Institute of Information Technology(CIIT)
Islamabad

***Abstract*—A smart city is developed, deployed and maintained with the help of Internet of Things (IoT). The smart cities have become an emerging phenomena with rapid urban growth and boost in the field of information technology. However, the function and operation of a smart city is subject to the pivotal development of security architectures. The contribution made in this paper is twofold. Firstly, it aims to provide a detailed, categorized and comprehensive overview of the research on security problems and their existing solutions for smart cities. The categorization is based on several factors such as governance, socioeconomic and technological factors. This classification provides an easy and concise view of the security threats, vulnerabilities and available solutions for the respective technologies areas that are proposed over the period 2010-2015. Secondly, an IoT testbed for smart cities architecture, i.e., SmartSantander is also analyzed with respect to security threats and vulnerabilities to smart cities. The existing best practices regarding smart city security are discussed and analyzed with respect to their performance, which could be used by different stakeholders of the smart cities.**

***Index Terms*—Smart city, ICT, IoT, Information security, RFID, M2M, WSN, Smart grids, Biometrics**

I. Introduction

The concept of smart cities is very vast as its vision encompasses management and organization of the whole city through embedded technology. These are ideally the cities that monitor and integrate status of all their infrastructures, management, governance, people and communities, health, education, and natural environment through information and communication technologies (ICT). The smart city is designed, constructed, and maintained by using highly advanced integrated technologies, that include sensors, electronics, and networks which are linked with computerized systems comprised of databases, tracking, and decision-making algorithms [1]. With increasing boost in urbanization, the concerns about economic restructuring, environmental issues, governance issues and public sector problems need to be dealt in a smarter approach. The challenges of modern cities are becoming complex as the pace of change has become very gigantic. This requires organizational changes specially focusing on the latest technologies and communication through Internet.

The term global village seems very coherent with the smart city as urbanization is dependent on latest technologies and Internet. The concept is also influenced by the industries promoting and selling their products like GPS, ipad, smartphones and other technologies [2]. The smart city hence promises smarter growth. It is said that proper investments in developing the systems of a city through embedded technologies will help in immense growth in economic system as well [3]. There are certain pioneering cities that are considered as the next generation smart cities. Names of such cities include Barcelona, Amsterdam, Masder, Singapore and France[4]. The general idea of a smart city and it's major components is given in the Figure 1.

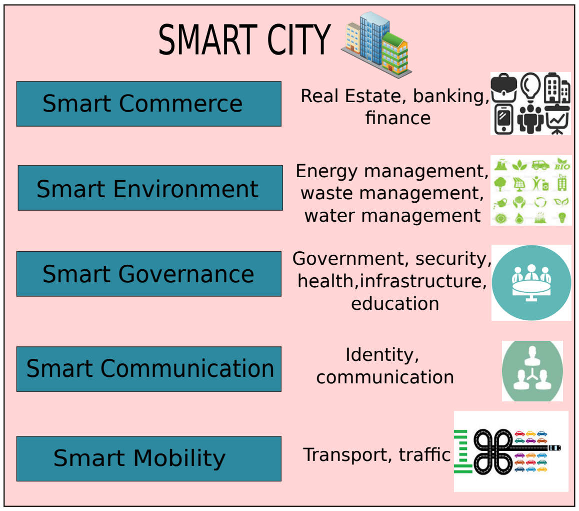

SMART CITY
Smart Commerce
Real Estate, banking, finance
Smart Environment
Energy management, waste management, water management
Smart Governance
Government, security, health,infrastructure, education
Smart Communication
Identity, communication
Smart Mobility
Transport, traffic

Fig. 1. Major components of a Smart City

It should be considered that the latest information and communication technologies (ICT) that are the core part of an efficient smart city are the Internet of things (IoT), smartphone technology, RFID(Radio Frequency Identification System), smart meters, semantic web, linked data, ontologies, artificial intelligence, cloud computing, collective intelligence, softwares, smart apps, and biometrics. The Internet of Things (IoT) is the network of physical/tangible objects integrated with computational devices, software, electronics, smart sensors and connectivity so that it can be used to achieve greater value and service by exchanging data with the maker, operator and other connected devices. Each thing is unambiguously distinctive through its embedded computing system but is able to inter-operate within the infrastructure of Internet. The concept of IoT play a vital role in development of ideal and secure smart city, as a smart city is solely dependent on the embedded technology. The IoT is considered as a major research and innovation idea that leads to a lot of opportunities for new services by interconnecting physical and virtual worlds with a huge amount of electronics distributed in different places including houses, vehicles, streets, buildings and many other environments [5].

The concept of smart cities is distinguished on the fact that it is solely dependent on embedded systems, smart technologies and the IoT. In general term, a smart city relies on information technology and the embedded infrastructure to facilitate it for a better living standard.

Though smart cities are facing many problems through their development, including socioeconomic and political issues, but the most important hurdle here is the technical issues. In technical problems, along with the other issues like system interoperability and cost efficient technology, the concern of security and privacy is very important [6]. The field of information security particularly deals with the issues of security and privacy of information. The goal of information security is to protect the information from attacks, viruses, frauds, and many other vicious activities that may cause harm either to the information, or the need of information in the technologically embedded smart cities. The security is very important in the infrastructures of smart cities because the networks will be prone to a large range of malicious attacks, and the internal and external parties are not trusted so security is a vital prerequisite to consumer acceptance [7]. As the concept of smart cities is still developing, the need to identify the core requirements of information security in various technologies is important.

In order to identify the correct requirements, and limitations, the aspiring smart cities should also be studied to identify the achievements and flaws in information security. As smart cities are exposed to malicious attacks which can alter or damage the whole infrastructure and communication systems, so the security of a smart city regarding the information point of view is very crucial. In other words, the key goals of a smart city would not be achieved if the information is not properly secured. Moreover, privacy should also be considered in a smart city environment. Privacy of systems that gather data and trigger emergency response when needed are also technological challenges that go hand in hand with the continuous security challenges [8].

The influence of information security is not in technical side only, but it also effects the economic concerns as well [9]. The issues of information security also needs to be addressed for a better economic development of a smart city. The requirements of of ideally secured and reliable smart cities need to be recognized considering most of the technologies, specially focusing on IoT, cloud computing, real world user interfaces, smart sensors, smartphones, semantic web etc. The factor of commercialization also needs to be considered as many IT companies have new solutions for the smart cities as well. The example is launch of Global Intelligent Urbanization by Cisco [10].

The objective of this paper is to survey pivotal problems and the solutions in practice regarding information security in a smart city in the light of key influencing factors. The rest of the paper is organized as follows: The Section II describes factors influencing a smart city. In Section III, the security issues of smart cities in perspective of governance, social and economic point of view are identified. The technologies that play part in making a smart city are considered in light of information security in Section IV. Section V gives comparative analysis of selected literature. Conclusions are discussed in Section VI.

II. Information security in a smart city

The security and privacy of information in a smart city has been interest of researchers. The reason behind it is that, in order to ensure the continuity of critical services like health care, governance and energy/utility issues in a smart city, the information security must be fool proof. The factors that are taken under consideration in order to identify the issues in information security in a smart city include governance factors, social/economic factors and most importantly economic factors. These factors are elaborated in the Figure (2). The researchers identify, explain and propose solutions to the information security issues by considering the mentioned factors. Most of the research work discuss the components and architecture of a smart city and then describe solutions for the security and privacy concerns.

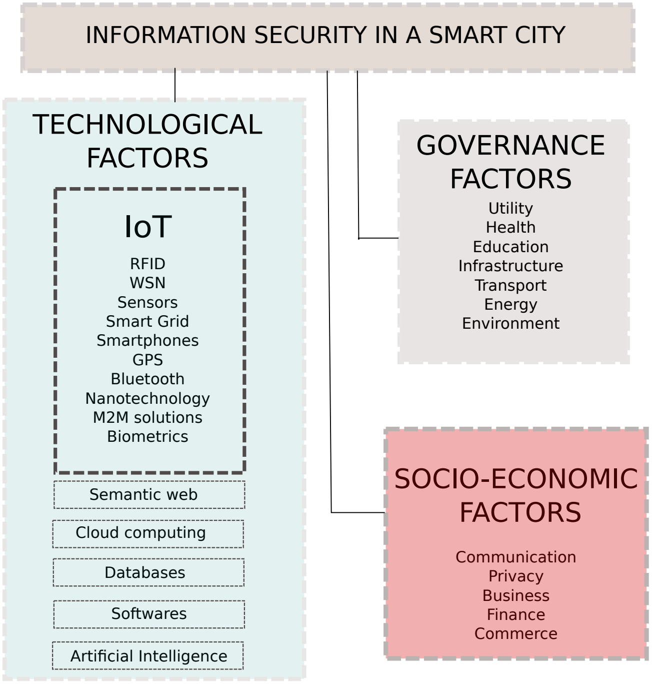


Fig. 2. Influencing factors on information security in a smart city

The IoT has been the key interest of the researchers as it is the core technology on which the smart cities are being developed and maintained [11]. For instance, in [7], the key hurdles and problems faced regarding security and privacy are discussed, keeping in the context of technological standards. the paper particularly focuses on Machine to Machine (M2M) standard solutions that are helpful in better implementation of IoT in a smart city.

A very important factor that plays key role in developing a smart city is ”big data”. The production of large data sets in a smart city is an inevitable phenomenon including national consensuses, government records, and other information about the citizens[12]. From such data, the smart cities can extract very important information helping real time analysis and ubiquitous computing. The author in paper [13] elaborates that though the big data provides various opportunities for smarter life, still it brings challenges of security and privacy. The challenges include lack of tools for management of big data, third party data sharing, threats in growing public databases, data leakage and concerns on digital security.

In another paper, the cyber security challenges are addressed [8]. Here the authors focus on two main challenges, security and privacy. They present a mathematical model depicting the interaction between people, IoT, and servers which are vulnerable to information security threats. Though the mathematical and graphical model for the IoT, people and servers is given stating that it will help in locating the problems in security and privacy, but the methodology to do so is not discussed. Moreover, Bohli et al [14] propose a distributed framework for IoT applications, which promises security, trust and privacy in information delivery. As IoT applications play a key role on building the smarter city, so some information security issues in a smart city can be addressed through the distributive framework.

The identification and classification of stakeholders of a smart city help in addressing the security problems in smart city in a better way. In this paper [15] all the concerned stakeholders regarding the security and privacy of information are identified by using onion model approach. The authors proposed that by identifying all the stakeholders, the security requirements and issues are identified in a better way. They also propose a comprehensive framework to deal with these issues.

In [16] the role of smart software is discussed in context of information security. The role of smart software in developing a smarter city is discussed specifically throwing light upon limitations regarding security issues. Certain security software models are also discussed. Another paper discusses the issue of security problems in sensing and querying in urbanization[17]. The authors propose an encryption scheme to deal with the issues of data integrity and privacy.

Smart grids are considered an important component of a smart city as they provide the services of very novel and efficient energy supply chain and information management[18]. In [19] the authors discuss the information security issues in a smart grid. The requirements of information security are identified by the authors and various models are discussed and compared regarding the concerned issues with methodologies.

Sanjay Goel in his paper [20] discuss the relationship between anonymity and security thus proving that there is a need of balance between anonymity and security in smart grids. The author then proposes a new idea for designing Internet in a way that security issues are dealt in a better way. The restructuring of Internet seems an interesting idea as author believes that redesigning and creation of multiple Internets with attempt to have a balance and anonymity can create a difference.

The paper [21] describes the main application systems for a smart city, and discuss various problem issues in constructing a smart city, particularly in China. Though the authors have discussed many hurdles in developing a smart city, but the information security issue has been overlooked and not been discussed by authors specifically.

Suciu et al [22] propose that by defining the platform of cloud computing and IoT properly for a smart city, better security can be achieved. For that, they have proposed a framework for the information that can be automatically be managed by the distributed cloud computing services. The privacy concerns of a common citizen of a smart city also need to be discussed. For example in [23] the privacy issues of citizens of a smart city are analyzed. This includes five dimensions according to the authors: identity privacy, query privacy, location privacy, footprint privacy and owner privacy. They propose a 5D model that addresses these five issues. The authors conclude that by following the model, privacy aware smart city is achievable.

The credibility of a smart city is questioned by Galdon-Clavell and Gemma [24] particularly on various problem areas in the implementation of smart cities, including security and privacy in context of individuals as well as institutions and governments. They describe the smart solutions and elaborate the issues regarding the implementation of such solutions. They propose is that by understanding the problem areas efficiently, smarter cities can be built. The issue of privacy is also discussed in [25] where the authors provide the issues and problem areas in trace analysis and mining for the smart cities. They conclude that though data mining and trace analysis play a key role in smart cities, still it is a challenging task to be done keeping in mind the privacy concerns and usage of limited and relevant data only.

III. Security concerns in governance, social and economic perspective

It is important to identify the core requirements in a smart city in context of information security, as this will help to develop a better understanding of the problem areas. Moreover it will also help to identify the correct and feasible solutions to those problems. As it was discussed in section II, the information security in smart city is mostly dependent on three factors: governance factors, socio-economic factors and the technological factors. These factors influence and identify the information security issues in a smart city. The ICT technologies work together to form a smart city, so not only these implement the whole infrastructure of a smart city and provide solutions to information security problems, but also trigger new concerns and problems regarding security, privacy, protection and resilience. The dependence of socio-economic and governance factors on the technological factors and the relationship of information security with them is shown in Figure 3.

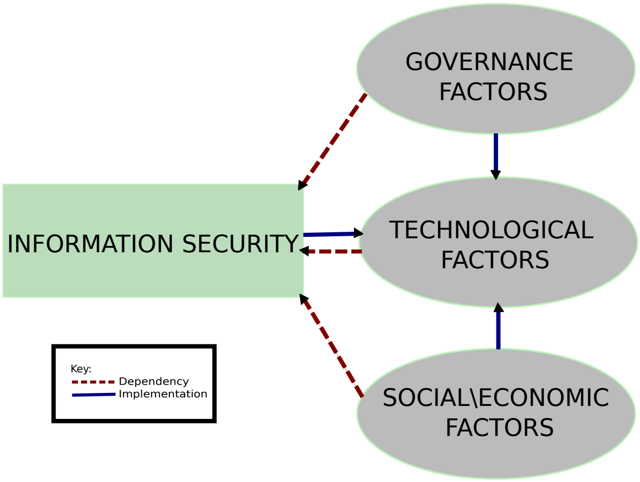


Fig. 3. Relationship between various factors influencing information security

Here it is illustrated that the governance factors and socio economic factors are dependent on the technological factors as these are implemented in a smart city through technology. These factors combine together to influence the information security issues in a smart city, which can again be managed through technology as it is a major driving force in this scenario. So the role of technology in security management and the issues of information security in implementation of all the technology requires the major focus. But in order to identify the core information security requirements, there is need to to study the governance, social and economic factors as well. In this section, these factors are analyzed in the light of information security.

A. Governance factors

As discussed in Figure 2, the governance factors that influence and trigger the security issues include utility, health sector, infrastructure, education, transport, etc. The biggest concern for the researchers is that a smart city though promises to provide all the ways to maintain whole infrastructure and management issues, but it's improper implementation can lead us to attacks and frauds. These malicious attacks and frauds can be very harmful to the core purpose of smart cities. In fact they could cause more damage than good they promise.

1) Need of security testing: Cesar Cerrudo, chief technology officer at security research firm IOActive Labs [26] have pointed on an important problem that the governance authorities that are the customers of technology firms don't bother to test the security of the systems they buy. Their priority is on testing the functionality of the technology, and they do not bother to focus on the security testing. So the awareness among the authorities to have a genuine concern over security issues is a key requirement.

2) Threats to critical infrastructures: The most important and crucial area is the critical infrastructure where changing a single process in a critical system can cause delay or loss of critical services [27].The main critical infrastructures include health care, industry and telecommunication. The implementation of critical infrastructures in smart cities is mainly on the IoT and smart grids. So the threats posing to these two technologies should be taken under consideration. Moreover the big data generated by critical systems can pose big problems regarding data integrity and resilience as it need to be properly stored, managed and protected. This is the responsibility of a smart city's critical infrastructure to maintain its security, resilience and data integrity [28]. Therefore, critical infrastructure need protection from malicious attacks that may cause crucial damage to smart cities and their promised services. The health sector is one of the most important type of critical infrastructure as if it is prone to security threats, it can not only cause privacy concerns of a patient [29] but may also pose threats to his life as the critical information can be changed by the attacker. So the health information systems in a smart city should have very secure encryption systems.

3) Smart mobility security and privacy requirements: Smart mobility may cause privacy concerns as personal information disclosure could happen in collecting, publishing, and utilizing trace data. Here, localization techniques include GPS, GSM, WiFi, Bluetooth, and RFID because centric servers do not need to know device IDs. [25] Some of the smartphone apps that provide services of smart mobility take mobile data and use trace analysis and data mining techniques. Moreover, The information sent and received from devices used in smart mobility infrastructure may subject to malicious attacks causing wrong traffic reports in satellite navigation systems [28]. Hence, it is clear by analyzing the problems in smart mobility that this domain requires optimized use of ICT technologies keeping in mind the security and privacy threats.

4) Energy and utility optimization: Energy and utility services are increasingly relying on smart grids that use bi-directional communication with the users in order to manage the distributed energy efficiently. Cloud computing also plays its role by providing features that are well suited for smart grid software platforms [30]. Data security and privacy remain top concerns for utilities and and the users that is playing a crucial setback in the adoption of smart grids [31]. Moreover the problem increases if it is implemented with clouds. In order to save energy and utilities from frauds and malicious attacks, a proper strategy should be made. This report by Semantic [28] suggests that public key infrastructure (PKI) or managed PKI can be used to tackle security issues in smart grids. The security problems and their solutions in a smart grid will be discussed in detail in section IV.

B. Social and economic factors

In a smart city, peoples' requests for assistance and personal management of social issues, are managed through technology that provides basic platform services for urban planning, emergency and community management, thus turning the smart city into a one-stop service system [21]. Moreover, the smart city promises smarter economic growth as it provides services to enhance the banking, finance and business activities more efficiently. The social and economic factors in a smart city include communication, individual identity, banking and finance. All of these are a critical part of a smart city and vulnerable to security and privacy issues.

1) Challenges in smart communication: The telecommunication sector is part of critical infrastructure of a smart city and is vulnerable to various malicious attacks, viruses, frauds and privacy attacks. As various financial and governance activities are also carried out through telecommunication and wireless networks so the need of security and authentication even increases. Moreover, Machine to machine (M2M) communications also help providing services offered to citizens of a smart city [32]. So the security threats relating to M2M communications should also be taken under consideration. The use of smartphones and tablets has bought new horizons in the communication between the citizens of a smart city. Moreover, it has also led to new threats to their privacy and information security. It is evident that more widely a technology is in use, it is more prone to attacks and viruses. As the smartphones have become very popular in recent years so they have become keen target of the hackers [33]. Wireless networking, bluetooth, cloud computing, IoT, in fact almost all ICT technologies play their part in smart communication and the security concerns relating to them should be considered while developing smart solutions.

2) Individual Privacy: The privacy of individuals is a fundamental right that should be guaranteed in a smart city. The individuals of a smart city use various services and communicate with each other through latest technology that is connected using heterogeneous networks and systems, which are the target for hackers who want to intrude in thier personal privacy thus depriving them from their personal right [23]. Here the role of social networking should also be considered regarding privacy and information security. The privacy concerns linked with the social networking depend on the level of identification of the provided information by the individual, the receivers and the way it may be used. Those social networking providers that promise not to expose their users identities openly still may provide the required enough data to identify the individual's profile [34].

3) Banking, finance and business: The banking, finance and business are all part of smart economy that is a fundamental component of a smart city. Though smart cities promise growth in economy, and better banking and business services, but this component of a smart city is most vulnerable to security threats as it can be attacked for personal financial use. The attackers also intend to sabotage the economy of certain organization, or a whole city.

IV. Technological Factors

Technology plays key role in making all the promises of a smart city functional. The smart city is solely dependent on technology in order to provide better services to the government and citizens. Smart city promises smarter economic growth, smarter governance and smarter services to the people through integrated and up to date technology. As this seems a very beautiful promise, but the concerns in security, privacy and data questions to it's credibility. In fact, smart cities are not so smart if the concerns in security and privacy are not properly catered. A summarized view of various security threats in a smart city are illustrated in Figure 4.

In this section, some of the core technologies used in developing and maintaining a smart city are discussed and analyzed under security.

A. IoT

The Internet of Things (IoT) incorporate a huge number of distinguished and heterogeneous devices, and gives free access to information for various on-line services for smart cities. IoT plays a gigantic role in developing and maintaining the services of a smart city, hence making the issue of secure information flow a huge task with respect to it. There are various IoT architectures such as urban IoTs [35] that are designed with respect to the needs of a smart city. There are numerous European projects that are made to tackle the research challenges in different aspects of the IoT. One of those is the SmartSantander, a real life experimental testbed for a smart city.

1) RFID tags: Radio frequency identification (RFID) tags are being used immensely in the various components of smart city including smart environment [36], industry [37] and mobility [38], etc. It has brought significant benefits in many other areas as well through improving real-time information visibility and traceability. This widespread technology is also prone to many threats and attacks thus making it vulnerable to security [39]. According to [40] the RFID tag is prone to give away sensitive information through unauthorized access, creating problem of data confidentiality and privacy. The problems to data integrity may also occur due to information leakage.

- Abuse of tags [40]
  The size of RFID tag is small making it cost efficient. As RFID tag can be embedded into various functions, so the margin for establish security protection system is very little due to its small size. The RFID tags are prone to illegal use by unauthorized users. Moreover the communication between RFID tag and RFID reader is done through a unique Electronic Product Code (EPC), which may be sabotaged by the attacker if they collect the EPC. Another problem is detaching the tag. In this case the transponders, that identify the tagged items in the RFID system, may be associated to some other thing and may be detached from its tag [41].

- Tag killing
  The tags are made useless through application of delete or kill commands by the attacker, or through physical destruction [41]. The reader in result cannot identify or read the tag. The DoS attacks are mainly used for this purpose. One important point to know that the tag killing process may also be used for enhancing the security of

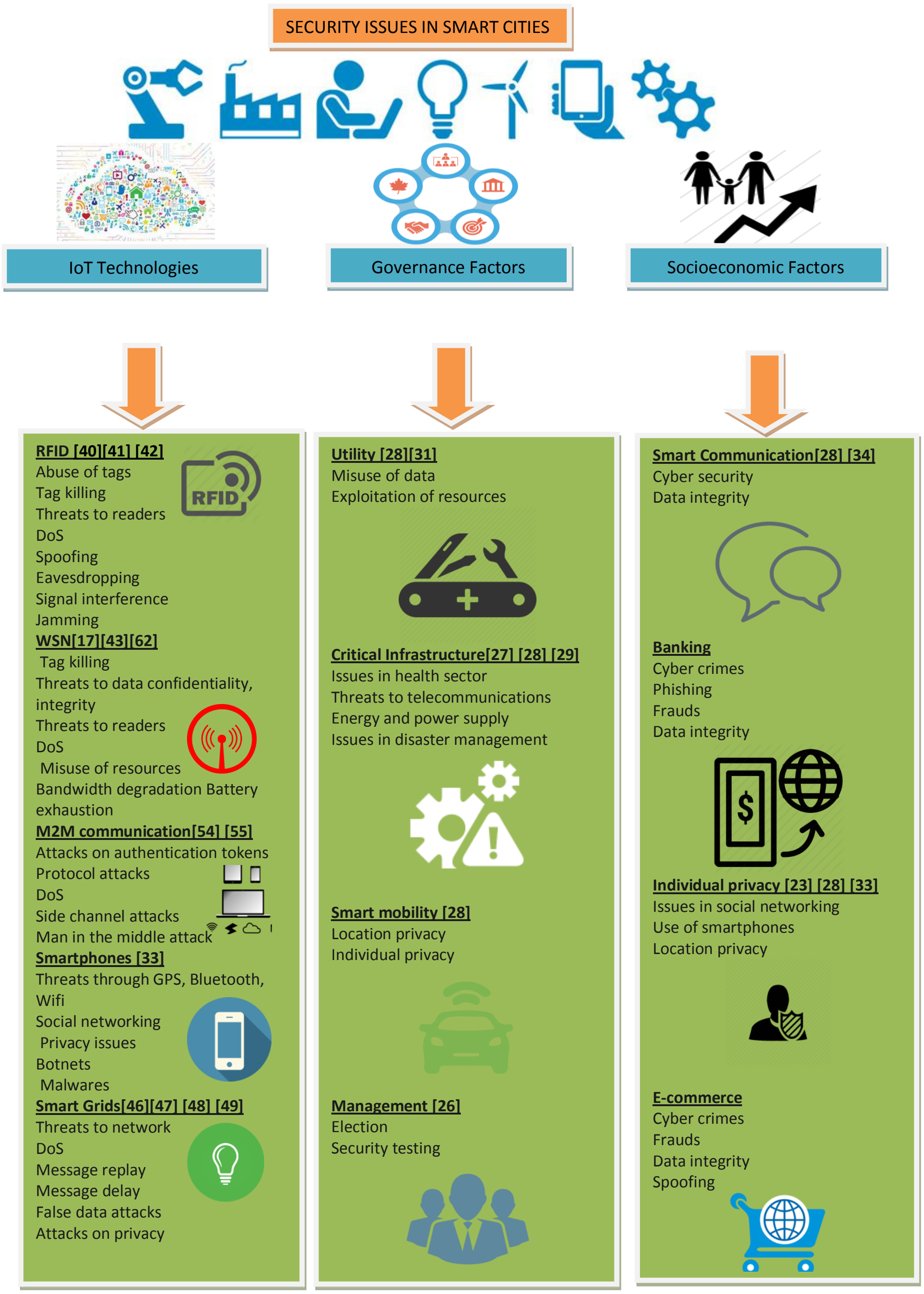


Fig. 4. Security issues in a smart city

the system by addressing privacy issues [40].

- Tag cloning
  Tag cloning is a process that gains the data from a original tag and makes an unauthorized copy of the captured data on a new tag. The copied data is transferred onto a tag of the attacker.

- Threats to readers
  One of the big security issue for RFID is sabotage of the reader. If the attacker gets control of the RFID reader, it can be sabotaged thus emitting some electromagnetic waves to destroy the data in the RFID tag [40].

- Threats to privacy
  The RFID tag can be tracked without the consent of users of a system [40]. Moreover, the uniqueness of tag's EPC makes it easy for hacker to track tags. Thus traceability and identification of the tags causes leakage of private information of the users of the system. So tag tracking is the main issue which harms the individual privacy [42]. Location privacy is also an issue that need to be addressed.

- Signal interference
  The RFID system adopts two frequency signals: low-frequency signal (125kHz, 225kHz, 13.65MHz) and high-frequency signal (433MHz, 915MHz,2.45GHz, 5.8GHz) [40], so there is signal interference between the two adjacent band. The attacker can induce signal interference that lead to issues of data integrity in the communication between the reader and tag.

- Jamming
  Jamming is an attempt to disturb the air interface disturbing the communication. This can effect the integrity of the system communication. This attack is done by powerful transmitters at a significant distance, and by passive means as well such as shielding [41].

- Threats to communication
  In RFID system the readers and tags communicate through wireless communication. The availability of the wireless signals, makes it easy for an attacker to search, manipulate, and jam wireless signals [40]. So encryption and authentication are crucial in order to protect the wireless transmission between the RFID readers and RFID tags. Attacks on wireless communications include active attacks and passive attacks [43].
  Wired communication between the RFID readers and the middle-ware system, through the Internet also has its security concerns. The need to ensure data confidentiality and integrity is very important.

- Denial of Service (DoS)
  An important type of attacks on RFID system are the Denial of Service (DoS) attacks. The purpose of the DoS is to disable the system making it useless. Device that broadcasts the radio signals can be used for malicious purposes and can disrupt or block the working of a RFID reader. There are many possibilities to perform DoS attacks including the possibility of placing the information in Faraday's cage [44].

- Spoofing
  Spoofing is a security threat in which tag data is duplicated and communicated to a reader. This occurs in case when the security protocol used in the RFID channel is revealed [41]. For instance, in case of an electronic seal, the e-seal information is transmitted to the reader from some other source that is the duplicate e-seal.

- Software attacks
  Software attacks are the most well known attacks including viruses, buffer overflows and worms injected in the RFID system to effect its functionality. These are the codded malicious programs aiming to infect and disturb the system making its performance slow or none.

- Cryptanalysis and eavesdropping
  Eavesdropping and cryptanalysis are the most frequent and talked about attacks on a RFID system. As cryptanalysis is getting stronger side by side with the cryptographic techniques so the need of better cryptography techniques is more evident than ever. The attacks include man in the middle attack, chosen plain text and chosen cipher-text attacks, and known plain-text and known cipher-text attacks. RFID tag emits data that is usually a unique identifier. When the data is communicated to RFID reader, it is prone to the risk of eavesdropping. In this particular case, eavesdropping is done by the attacker by catching data with a reader one for the correct tag family and frequency during a tag is being read by authorized reader [41].

- Suggestions and techniques for better security
  It is important to discuss various techniques and strategies that may play part in better security performance of RFID in a smart city. Privacy protection is an important need and right of the citizens and as discussed previously, that tag killing may play role in privacy protection but this also may lead to loss data permanently. A better option is tag sleeping. When the tag does not need to to be read, it is put to sleep temporarily [45]. There is another concept of tag blocking and selective blocking [40]. Other techniques include relabeling approach, re-encryption and minimalist cryptography [45].
  The interference problems in the RFID system can be dealt by data coding, multiple re transmission and data integrity check.
  Hash-Lock and Hash Link techniques are also important as they provide better authentication. There are other

authentication techniques based on hash such as ID exchange and distributed RFID challenge/answer authentication [39].
The authentication supported by hash requires the symmetric key distribution. So it is evident that protecting the shared key is essential in the procedure of authentication. The shared key is saved in the tag of the RFID.

2) Smart grids: Smart grids play core part in a smart city regarding energy deployment and management. These are actually communicating instruments including sensors and communication networks that help in communicating data in real time [18]. When the data is shared in real time scenario among power generator, distributed resources, the service provider and the users, any information that is prone to attacks, that would take the system to failure. This will unfortunately lead to user's uncertainty and discontentment with the system. Through literature review [46][47] [48], we classify main threats that should be kept under consideration while constructing and deploying a smart grid as follows:

- Threats to network availability
  The most vindictive attacks that target the network availability are the denial-of-service (DoS) attacks. These attacks attempt to delay, block or corrupt services by abusing information in the smart grid. It is evident [46] that mostly of the smart grid use IP based protocols. As TCP/IP is open to DoS attacks, so such attacks are becoming huge problem in a smart grid.

- Threats to data Integrity
  In case of smart grids particularly, data integrity is needed in case of data like sensor values and control commands. The main objective of data integrity includes defense mechanism for information modification through various means such as message injection, message replay, and message delay on the network. Threats to data integrity cause many issues like infrastructure or people of the smart city may be harmed. The main goal of the integrity attacks is either customers information or network operation information. These attacks tend to abuse critical data in a smart grid.
  False data [49] injection attacks are very powerful attacks against the state estimation in the power grid. In this case the hacker takes advantage of the configuration of a power system to launch malicious attacks by infusing wrong data to the monitoring center questioning data integrity.

- Threats to information privacy
  Privacy of smart grid communication systems is important as it is the main concern and right of the consumers. Smart grid communications should take care of the privacy during communication in real time.

- Threats to devices
  Smart meters are prone to physical attacks like battery change, removal, and modification. Moreover, functions including remote connect/disconnect meters and outage reporting may be used by unwarranted third parties.

- Proposed solutions for smart grids
  According to literature [48] [47] the possible solutions for threats to devices in a smart grid include ensuring the integrity of meter data and maintaining meter securely.
  Moreover for wireless networking, TCP/IP for smart grid networks is a better choice for Internet. Moreover the M2M solutions prosed by IEEE including 802.11i, 802.16e, and 3GPP LTE should be used. For sensor networks various encryption standards should be adopted for authentication. Public key infrastructure (PKI) and managed PKI are also a good choice for smart grids security.

3) Biometrics: Biometrics is an automated recognition of a person through unique behavioral and biological characteristics. There are two main types of biometric characteristics: physiological and behavioral. Both are acquired by applying proper sensors and distinctive features are taken in use to get a biometric template in authentication process [50]. In fact, it is generally thought that any other substitute to biometrics for identification in integrated security applications does not exists.
Biometrics is said to play a key role in information security issues in a smart city. According to Bill Maheu, who is senior director for Qualcomm Government Technologies, every year 3.7 trillion dollars are lost to global frauds, which can be solved sufficiently by implementing biometrics [51]. Biometrics in fact can make various components of a smart city secure with respect to frauds and malicious attacks[51]:

- Health
- Education
- Institution
- Utility
- Patrol and security

4) Smartphones: Smart phones are one of the core component of IoT infrastructure in a smart city as they give access to various services and smart applications that help in maintaining and developing a better smart city. These are also the main source of people's role in a smart city. Smartphones have become immensely popular in recent years, making them an attractive thing to be attacked by hackers and viruses. The main security threats in smartphones are illustrated as under [33]:

- Malicious smart applications
  In some cases, hackers upload vindictive applications to application marketplaces for iphone and android devices. Such applications may also be present in Internet. Such

smart applications can infect the smart-phone devices and may cause many security and information privacy issues.

- Botnets
  Botnet is formed by attackers by contaminating multiple devices with malwares that victimize broadly through th e-mail attachments or from smart applications or malicious websites.

- Spyware
  Attackers may misuse the available spyware to hijack a smart-phone, allowing them to locate and hear calls, check messages and e-mails, and track a users location through GPS updates. So the user's privacy is totally sabotaged in this way.

- Threats from bluetooth
  Wireless devices show their existence and permit unrequested connections and in case the end users do not know how to manage and configure their bluetooth settings properly.

- Location and GPS
  The location privacy of individuals can be sabotaged by the attackers by various attempts on the GPS feature provided in the smart-phone.

- Threats through WiFi
  Attacker on a smartphone can catch information during the communication between smartphones and Wi-Fi hotspots. The main problem is extreme vulnerability of the Wifi hotspot architecture where there is no encryption to protect transmitted data

- Threats in social networks
  As smartphone usage has gone through a major boost, so has mobile social networking flourished. People give a lot of personal information and time to social networks. Many links on social networking websites and applications may effectively spread malicious malware. Moreover individual privacy is also prone to major attacks on social networking websites.

Literature [52] proposes various solutions to threats to smartphones, illustrated as under:

- Anti Viruses and firewalls
  Anti-viruses for smartphones scan every data including files, memory, SMS, MMS, emails etc. These solutions can help in preventing the malicious malwares. Moreover the threat of access to phishing site is also controlled. The firewalls on the other hand block connections that are unauthorized preventing the network attacks.

- Secure API
  The secure APIs have the cryptography properties helping program and application developers for implementation of secure functionality.

- Authentication and access control
  The process of authentication process can prevent unauthorized use of smartphone devices. Moreover, access control is also important as it limits the access of malicious processes and attacker to resources and services.

- Filters
  Filters include SPAM filters that blocks SPAM MMS, SMS, emails and calls from unknown origins.

- Cellular M2M solutions
  Cellular Smart City M2M technology advancements are tacking momentum with time, and good number of organizations envision the future IoT applications to run over cellular networks. There are specific M2M solutions for smartphones [7]:
  - ETSI M2M
    It is made by different manufacturers and it provides the framework, requirements and architecture, for the technologies like 3GPP that can be used to populate the developed architecture.
  - 3GPP LTE-M
    It is OFDM based LTE making cellular M2M has suddenly become of interest for a significant target market.

5) M2M communication: Machine to machine (M2M) communications promise dramatic achievements in the applications and services offered to citizens, making smart city a reality [32]. Machine to machine protocols are used for communication fix the rules of engagement for at least two nodes of a network. Internet Protocol (IP) has become the standard for such communication purposes. Examples of protocols that can be used for communication are: ISA 100A, link Layer, Wireless HART, IPv6 and ZigBee [53]. IPv6 plays a gigantic role in the IoT. The plus of IPv6 is that it fulfills the demands of potability and helps variant systems to work together. According to [54] [55] the main security concerns in M2M communications include:

- Physical Attacks
  These attacks include using modified softwares for the purpose of fraud. The main breaches that occur due to these attacks are in integrity of data and M2M softwares.

- Attacks on authentication tokens
  The threats include physical attacks as discussed above and side-channel attacks. The authentication tokens can also be cloned for malicious purposes.

- Configuration Attacks
  The example of configuration attacks include malicious software updates configuration changes that lead to fraud. Moreover, mis-configuration by the user may also occur.

- Protocol Attacks
  The protocol attacks are mainly designed against the devices. For example: man-in-the-middle attacks, DoS attacks, and attacks on OAM and its traffic.

- Threats in network security
  These attacks mainly target mobile networks. The examples of such threats include impersonation of devices and traffic tunneling between them. Moreover, mis-configuration of the firewall in the devices is also a serious network security breach. The DoS attacks on the network also pose a major problem.

- Breaches in privacy
  Privacy is a huge concern of the individuals of a smart city as it is a basic human right. But it becomes very difficult to take care of the privacy of citizens through M2M communications as the ways in which data collection, mining, and provisioning is accomplished are totally different from those that we now know and there are a huge amount of occasions for personal data to be collected.
  Eavesdropping can cause major concerns over individual privacy and data integrity. Moreover, masquerading as other user's devices is also a gigantic security problem.

There are various M2M standard solutions available to establish a smart city with respect to security [7]:

- IEEE Standards Solutions
  The IEEE provides standard mechanisms for the physical (PHY) and medium access control (MAC) layers which are useful in implementation of a smart city. There exist three families that can provide low-power and short-range IoT operation for a smart city [7].
  IEEE 802.15.4:
  Important characteristics include real-time quality by guarantying time slots, collision dodging through CSMA/CA and merged assistance for secure communication. Moreover, the devices include power management functions for example, energy detection and link quality.
  IEEE 802.15.1 is used in Bluetooth.
  IEEE 802.15.11 This technology is provided from WiFi Alliance, a trade association in control of the certifying products if they adjust to particular standards of interpretability. The Wifi Protected Access (WPA) [56] is a security protocol that has become the regulation for providing security .11 networks. Here by using an already shared encryption key (PSK) or digital certificates, the WPA algorithm Temporal Key Integrity Protocol (TKIP) encrypts information providing authentication to the particular networks. he WPA algorithm (TKIP) further improved upon to the new WPA2 [57] [56] that utilize more securer encryption algorithm that is Advanced Encryption Standard (AES). Moreover this protocol also uses better and advanced key distribution techniques, which help in improved session security to avoid eavesdropping.

6) SmartSantander: An IoT testbed for a smart city: This facility is an IoT infrastructure deployment in Santander, Spain [58]. This unique arrangement contains more than 2000 IoT devices deployed in an urban scenario [59]. This project aims at developing an architecture of a smart city through of IoT by a twofold approach [60]:
i. Experimentation support:
It provides the platform for the research community to get results for their experimental research in a real life scenario. It is an amazing opportunity for the researchers that they can allocate and manage the required IoT resources to run their experiments.
ii. Service provision:
SmartSantander provides the services promised by a smart city in accordance with the needs and requirements of people described in form of use cases.
In SmartSantander a detailed network deployment is done on basis of the use cases oriented to define all the services and technological support needed over them. Moreover, many applications for smartphone operating systems are developed in order to add in the people's role in this environment of sensing and connecting. SmartSantander has a 3-tiered architecture: IoT nodes/End points, IoT nodes/Repeaters and Gateways [58]. Endpoints and repeaters are used for sensing various parameters, but the main difference is that endpoints don't forward the information, and the repeaters send the information to the required places. Gateway gathers all the information sent. The interface for the communication used in this scenario is IEEE 802.15.4 interface [61]. This architecture forms into wireless sensor networks (WSN) for the information sharing. There are various security concerns for a WSN, categorized as under [62]:

- Attack on data confidentiality
  There are various crypt-analysis attacks that cause threats to data confidentiality on a WSN during information sending and receiving.
- Threats to data integrity
  The data may be abused, changed and modified due to various attacks.
- Misuse of resources
  Another problem that arises in the scenario of a smart city is the misuse of the IoT devices for malicious purposes.
- Bandwidth degradation
  Bandwidth degradation may effect the information flow and prone to abuse of data.
- Battery or resource exhaustion
  The malicious attacks infect the IoT devices making their battery life and resources poor.

- Unauthorized Access
  An attacker can access to WSN resources to obtain the keys for malicious purposes.
- Threats to Authentication
  Authentication service ensures security of a system by restraining any attacker from entering the system. In WSN, the attacker may get hold of the user id and password hence getting over the authentication process. In this way attackers can get hold of all the services provided by the WSN. So it should be made sure that SmartSantander has a foolproof authentication system.
- DoS
  The denial of service attacks are a huge problem in WSNs as these attack suspend the services of whole system.

Following goals in SmartSantander are reached regarding security of information in IoT infrastructure:

- Data confidentiality
  Cryptographic techniques are used in order to ensure data confidentiality. The literature [63] compares various encryption techniques that makes it easy to choose the technique better for the system:
  PKI distribution mechanism provides the best security and on the other hand, symmetric cryptography mechanism requires significantly less computational complexity on the node but higher memory requirements. The ID-based encryption mechanisms provide asymmetric cryptography which is cost effective in terms of memory as well, but it is more complex to maintain. A qualitative analysis on the comparison of these schemes is given in [63] which is modified into quantitative analysis in Figure 5.

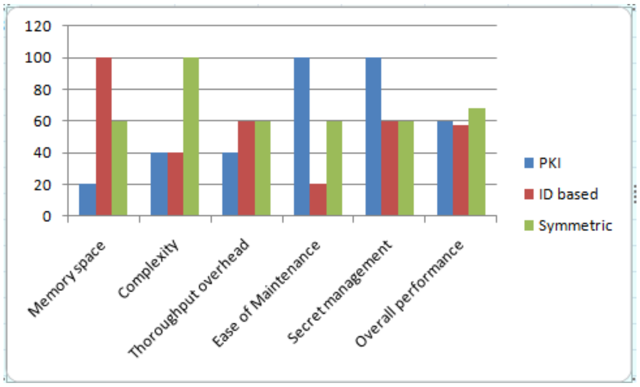


Fig. 5. PERFORMANCE ANALYSIS OF VARIOUS CRYPTOGRAPHY TECHNIQUES IN PERCENTAGE [63]

- Integrity
  For data integrity Cyclic Redundancy Code (CRC) or a Message Authentication Code (MAC) is used in this scenario.
- Authentication
  Here the message authentication code (MAC) is used for symmetric cryptography. A digital signature for asymmetric cryptography may also be used.
- No repudiation
  For no repudiation, secure protocol is used with acknowledgment.
- Freshness
  nonce is a random number that is used only once for a given time. It is used inside the secure protocol for achieving freshness.

The SmartSantander is setting position at the assemblage between providing different types of services and the deployment of a huge experimental testbed. This includes many stakeholders, including citizens, experimentalists, government and authorities. It is an ideal testbed to check and manage the security issues for a smart city.

V. Comparative Analysis

As the conception of smart cities is still under evolution, the need to identify the core threats of information security in various technologies is important. The security of information in a smart city has been interest of researchers because, in order to guarantee the provision of all the services in a smart city, the information security issues must be catered properly. Smart city involves various services in different components including mobility, communication and critical infrastructure. The need to achieve secure information sharing through the technology being used is crucial. The measures for secure information flow can be taken by identifying the problem areas and threats to security. So the purpose of research on information security in a smart city is that by understanding the problem areas and available solutions efficiently, smarter cities can be deployed and maintained. The IoT has been the key interest of the researchers as it is the core technology on which the smart cities are being developed and maintained.
Through literature review, various security breaches along with their existing solutions have been identified and illustrated in Table II. The security threats and solutions are illustrated with reference to the papers reviewed. This summarized review will help learning the core security threats and available solutions that are being discussed in latest research.

VI. Conclusions

The issue of information security in a smart city ranges over on a variety of aspects including social, economic, structural and governance factors. This paper provides a comprehensive overview on the threats, vulnerabilities and available solutions in order to facilitate much needed research in addressing the problem areas in smart city security. The technological factors are pivotal in deployment and maintenance of a smart city. In fact, technology is the driving force that establishes and maintains a smart city to deliver the promised services. Nonetheless, the significance of studying security of a smart city with regards to governance and socioeconomic factors help in identifying security concerns and requirements of the concerned stakeholders. Moreover, this practice also facilitates in identifying risks and vulnerabilities in plausible manner. It is evident that security is the weakest link in the implementation

IoT Technologies	Application in smart city	Security threats	Available solutions	Related literature
RFID	Industry, environment utility, mobility, infrastructure	Threats to readers, threats to privacy, abuse of tags, tag killing, signal interference, jamming, threats to communication, DoS, spoofing, software attacks, cryptanalysis and eavesdropping	Selective blocking, minimalist cryptography, tag sleeping, tag blocking re encryption, data coding, multiple retransmission, hash lock, hash link, SKD	[36] [37] [38] [41] [40] [39] [42] [45]
WSN	Environment, utility, Health, energy, infrastructure, governance and commerce	DoS attacks on data confidentiiality, threats to data integrity, Misuse of resources' bandwidh degradation, baterry exhaustion, unauthorized access	CRC, MAC, PKI, symmetric cryptography, and light weight asymmetric cryptography digital signatures, secure protocols	[17] [43] [62]
M2M communication	Smart communication, governance, health, critical infrastructure, education	Physical Attacks, Attacks on authentication tokens, Protocol Attacks, man-in-the-middle attack, DoS attacks, attacks on privacy, side-channel attacks, fraudulent software updates	IEEE standard solutions: 802.15.4, 802.15.1, 802.15.11	[54] [55] [32] [7]
Smart grids	Smart energy, power, utility critical infrastructure smart Appliances and smart homes	threats to network availability, DoS, breaches in data integrity, message replay, message delay, false data attacks, attacks on privacy,	Public key infrastructure (PKI) or managed PKI, AES for sensor networks, protected routing protocols, 802.11i, 802.16e, 3GPPLTE - M	[18] [28] [46] [47]
Smartphones	Smart communication, Smart mobility, Entertainment,	Malicious smart applications, bot-nets, spy-wares, threats from Bluetooth, Location privacy and GPS, threats through WiFi, threats in social networking, privacy issues	Antivirus, firewalls, secure APIs, authentication and access control filters, ETSI M2M, 3GPPLTE - M	[33] [52]
Biometrics	Health, atrol and security, education, institutions, coorporate sector, and utility	N.A	N.A	[50]

TABLE I
INFORMATION SECURITY ANALYSIS OF VARIOUS TECHNOLOGIES USED IN A SMART CITY

of a smart city. The serious repercussions of flawed security may undo the value of promised features and services of a smart city. The excellent functionality of smart solutions would have no value if the system has security loopholes. The smart solution manufacturers and decision making authorities, both are stakeholders and responsible for ensuring the security of a deployed system.

REFERENCES

[1] B. Bowerman, J. Braverman, J. Taylor, H. Todosow, and U. Von Wimmersperg, "The vision of a smart city," in *2nd International Life Extension Technology Workshop, Paris*, 2000.

[2] K. R. Kunzmann, "Smart cities: A new paradigm of urban development," *Crios*, vol. 4, no. 1, pp. 9–20, 2014.

[3] S. Dirks, C. Gurdgiev, and M. Keeling, "Smarter cities for smarter growth: How cities can optimize their systems for the talent-based economy," *IBM Institute for Business Value*, 2010.

[4] "Top five smart cities in the world," http://www.forbes.com/sites/peterhigh/2015/03/09/the-top-five-smart-cities-in-the-world/, accessed: 2015-04-03.

[5] N. Komninos, H. Schaffers, and M. Pallot, "Developing a policy roadmap for smart cities and the future internet," in *eChallenges e-2011 Conference Proceedings, IIMC International Information Management Corporation*. IMC International Information Management Corporation, 2011.

[6] M. Naphade, G. Banavar, C. Harrison, J. Paraszczak, and R. Morris, "Smarter cities and their innovation challenges," *Computer*, vol. 44, no. 6, pp. 32–39, 2011.

[7] A. Bartoli, J. Hernández-Serrano, M. Soriano, M. Dohler, A. Kountouris, and D. Barthel, "Security and privacy in your smart city," in *Proceedings of the Barcelona Smart Cities Congress*, 2011.

[8] A. S. Elmaghraby and M. M. Losavio, "Cyber security challenges in smart cities: Safety, security and privacy," *Journal of Advanced Research*, vol. 5, no. 4, pp. 491–497, 2014.

[9] R. Anderson, "Why information security is hard-an economic perspective," in *Computer Security Applications Conference, 2001. ACSAC 2001. Proceedings 17th Annual*. IEEE, 2001, pp. 358–365.

[10] "Cisco intelligent urbanisation," http://www.urenio.org/2009/03/13/cisco-intelligent-urbanisation/, accessed: 2015-04-22.

[11] M. Dohler, I. Vilajosana, X. Vilajosana, and J. LLosa, "Smart cities: An action plan," in *Barcelona Smart Cities Congress*, 2011.

[12] R. Kitchin, "The real-time city? big data and smart urbanism," *GeoJournal*, vol. 79, no. 1, pp. 1–14, 2014.

[13] C. Schmitt, "Security and privacy in the era of big data," 2014.

[14] J.-M. Bohli, P. Langendörfer, and A. F. Skarmeta, "Security and privacy challenge in data aggregation for the iot in smart cities," *River Publisher Series in Cmoounications*, p. 225, 2013.

[15] Z. Khan, Z. Pervez, and A. Ghafoor, "Towards cloud based smart cities data security and privacy management," 2014.

[16] M. Sen, A. Dutt, S. Agarwal, and A. Nath, "Issues of privacy and security in the role of software in smart cities," in *Communication Systems and Network Technologies (CSNT), 2013 International Conference on*. IEEE, 2013, pp. 518–523.

[17] M. Wen, J. Lei, and Z. Bi, "Sse: A secure searchable encryption scheme for urban sensing and querying," *International Journal of Distributed Sensor Networks*, vol. 2013, 2013.

[18] C. Clastres, "Smart grids: Another step towards competition, energy security and climate change objectives," *Energy Policy*, vol. 39, no. 9, pp. 5399–5408, 2011.

[19] A. P. A. Ling and M. Masao, "Selection of model in developing information security criteria on smart grid security system," in *Parallel and Distributed Processing with Applications Workshops (ISPAW), 2011 Ninth IEEE International Symposium on*. IEEE, 2011, pp. 91–98.

[20] S. Goel, "Anonymity vs. security: The right balance for the smart grid," *Communications of the Association for Information Systems*, vol. 36, no. 1, p. 2, 2015.

[21] K. Su, J. Li, and H. Fu, "Smart city and the applications," in *Electronics, Communications and Control (ICECC), 2011 International Conference on*. IEEE, 2011, pp. 1028–1031.

[22] G. Suciu, A. Vulpe, S. Halunga, O. Fratu, G. Todoran, and V. Suciu, "Smart cities built on resilient cloud computing and secure internet of things," in *Control Systems and Computer Science (CSCS), 2013 19th International Conference on*. IEEE, 2013, pp. 513–518.

[23] A. Martinez-Balleste, P. A. Pérez-Martínez, and A. Solanas, "The pursuit of citizens' privacy: a privacy-aware smart city is possible," *Communications Magazine, IEEE*, vol. 51, no. 6, pp. 136–141, 2013.

[24] G. Galdon-Clavell, "(not so) smart cities?: The drivers, impact and risks of surveillance-enabled smart environments," *Science and Public Policy*, vol. 40, no. 6, pp. 717–723, 2013.

[25] W. Z. S. L. Gang Pan, Guande Qi and Z. Wu, "Trace analysis and mining for smart cities: issues, methods, and applications," *IEEE Communications Magazine*, vol. 121, 2013.

[26] "Why smart cities need to get wise to security and fast," http://www.theguardian.com/technology/2015/may/13/smart-cities-internet-things-security-cesar-cerrudo-ioactive-labs, accessed: 2015-05-14.

[27] N. Abouzakhar, "Critical infrastructure cybersecurity: A review of recent threats and violations," 2013.

[28] Semantic, "Transformational smart cities: cyber security and resilience," 2010.

[29] A. Solanas, C. Patsakis, M. Conti, I. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. A. Pérez-Martínez, R. Di Pietro, D. N. Perrea *et al.*, "Smart health: a context-aware health paradigm within smart cities," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 74–81, 2014.

[30] Y. Simmhan, A. G. Kumbhare, B. Cao, and V. Prasanna, "An analysis of security and privacy issues in smart grid software architectures on clouds," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*. IEEE, 2011, pp. 582–589.

[31] J. Polonetsky and C. Wolf, "How privacy (or lack of it) could sabotage the grid," *Smart grid news*, 2009.

[32] J. Wan, D. Li, C. Zou, and K. Zhou, "M2m communications for smart city: An event-based architecture," in *Computer and Information Technology (CIT), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 895–900.

[33] N. Leavitt, "Mobile security: finally a serious problem?" *Computer*, vol. 44, no. 6, pp. 11–14, 2011.

[34] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 2005, pp. 71–80.

[35] A. Zanella, N. Bui, A. P. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, 2014.

[36] A. Luvisi and G. Lorenzini, "Rfid-plants in the smart city: Applications and outlook for urban green management," *Urban Forestry & Urban Greening*, vol. 13, no. 4, pp. 630–637, 2014.

[37] X. Zhu, S. K. Mukhopadhyay, and H. Kurata, "A review of rfid technology and its managerial applications in different industries," *Journal of Engineering and Technology Management*, vol. 29, no. 1, pp. 152–167, 2012.

[38] A. Ramos, A. Lazaro, and D. Girbau, "Multi-sensor uwb time-coded rfid tags for smart cities applications," in *European Microwave Conference (EuMC), 2014 44th*. IEEE, 2014, pp. 259–262.

[39] S. Xiwen, "Study on security issue of internet of things based on rfid," in *Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on*. IEEE, 2012, pp. 566–569.

[40] X. Nie and X. Zhong, "Security in the internet of things based on rfid: Issues and current countermeasures," in *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering*. Atlantis Press, 2013.

[41] S. Mohite, G. Kulkarni, and R. Sutar, "Rfid security issues," in *International Journal of Engineering Research and Technology*, vol. 2, no. 9 (September-2013). ESRSA Publications, 2013.

[42] R. Aggarwal and M. L. Das, "Rfid security in the context of internet of things," in *Proceedings of the First International Conference on Security of Internet of Things*. ACM, 2012, pp. 51–56.

[43] S. Babar, A. Stango, N. Prasad, J. Sen, and R. Prasad, "Proposed embedded security framework for internet of things (iot)," in *Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology (Wireless VITAE), 2011 2nd International Conference on*. IEEE, 2011, pp. 1–5.

[44] Q. Xiao, C. Boulet, and T. Gibbons, "Rfid security issues in military supply chains," in *Availability, Reliability and Security, 2007. ARES 2007. The Second International Conference on*. IEEE, 2007, pp. 599–605.

[45] R. Pateriya and S. Sharma, "The evolution of rfid security and privacy: a research survey," in *Communication Systems and Network Technologies (CSNT), 2011 International Conference on*. IEEE, 2011, pp. 115–119.

[46] Z. Lu, X. Lu, W. Wang, and C. Wang, "Review and evaluation of security threats on the communication networks in the smart grid," in *MILITARY COMMUNICATIONS CONFERENCE, 2010-MILCOM 2010*. IEEE, 2010, pp. 1830–1835.

[47] Y. Yan, Y. Qian, H. Sharif, and D. Tipper, "A survey on cyber security for smart grid communications," *Communications Surveys & Tutorials, IEEE*, vol. 14, no. 4, pp. 998–1010, 2012.

[48] J. Liu, Y. Xiao, S. Li, W. Liang, and C. Chen, "Cyber security and privacy issues in smart grids," *Communications Surveys & Tutorials, IEEE*, vol. 14, no. 4, pp. 981–997, 2012.

[49] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[50] C. Rathgeb and A. Uhl, "A survey on biometric cryptosystems and cancelable biometrics," *EURASIP Journal on Information Security*, vol. 2011, no. 1, pp. 1–25, 2011.

[51] "How connectivity and biometrics are making cities safer," http://smartcitiescouncil.com/article/how-connectivity-and-biometrics-are-making-cities-safer, accessed: 2015-05-31.

[52] W. Jeon, J. Kim, Y. Lee, and D. Won, "A practical analysis of smartphone security," in *Human Interface and the Management of Information. Interacting with Information*. Springer, 2011, pp. 311–320.

[53] N. Dlodlo, T. Foko, P. Mvelase, and S. Mathaba, "The state of affairs in internet of things research." Academic Conferences International Ltd, 2012.

[54] C. Hongsong, F. Zhongchuan, and Z. Dongyan, "Security and trust research in m2m system," in *Vehicular Electronics and Safety (ICVES), 2011 IEEE International Conference on*. IEEE, 2011, pp. 286–290.

[55] D. Jiang and C. ShiWei, "A study of information security for m2m of iot," in *Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, vol. 3. IEEE, 2010, pp. V3–576.

[56] S. Bai, Y. Wang, and Z. Xue, "Research on security of wpa/wpa2 protocol," *Information Security and Communications Privacy*, vol. 1, pp. 106–108, 2012.

[57] J.-C. Chen, M.-C. Jiang, and Y.-w. Liu, "Wireless lan security and ieee 802.11 i," *Wireless Communications, IEEE*, vol. 12, no. 1, pp. 27–36, 2005.

[58] "Smart santander," http://www.fed4fire.eu/smart-santander/, accessed: 2015-05-22.

[59] L. Sanchez, J. Galache, V. Gutierrez, J. Hernandez, J. Bernat, A. Gluhak, T. Garcia, P. Cunningham, and M. Cunningham, "Smartsantander: The meeting point between future internet research and experimentation and the smart cities ist future networks & mobile summit poland," in *Conference Proceedings Cunningham, P. and Cunningham, M.(Eds) IIMC International Information Management Corporation, Warsaw, Poland*, 2011.

[60] A. G. Jose, V. Gutiérrez, J. R. Santana, L. Sánchez, P. Sotres, J. Casanueva, and L. Muñoz, "Smartsantander: A joint service provision facility and experimentation-oriented testbed, within a smart city environment," 2013.

[61] L. Sanchez, L. Muñoz, J. A. Galache, P. Sotres, J. R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis *et al.*, "Smartsantander: Iot experimentation over a smart city testbed," *Computer Networks*, vol. 61, pp. 217–238, 2014.

[62] C. Hennebert, "Internet of things: Security management for large scale deployment in the city," 2013.

[63] C. Hennebert and V. Berg, "A framework of deployment strategy for hierarchical wsn security management," in *Data Privacy Management and Autonomous Spontaneus Security*. Springer, 2012, pp. 310–318.

Sorting Pairs of Points Based on Their Distances

Mohammad Farshi, Abolfazl Poureidi, Zorieh Soltani
Combinatorial and Geometric Algorithms Lab., Department of Computer Science
Yazd University, Yazd, P. O. Box 89195-741, Iran

***Abstract*—Sorting data is one of the main problems in computer science which studied vastly and used in several places. In several geometric problems, like problems on point sets or lines in the plane or Euclidean space with higher dimensions, the problem of sorting pairs of points based on the distance between them is used. Using general sorting algorithms, sorting $\binom{n}{2}$ distances between n points can be done in $\mathcal{O}(n^2 \log n)$ time. Ofcourse, sorting $\Theta(n^2)$ independent numbers does not have a faster solution, but since we have dependency between numbers in this case, finding a faster algorithm or showing that the problem in this case has $\Omega(n^2 \log n)$ time complexity is interesting. In this paper, we try to answer this question.**

Keywords*—*Sorting problem; Sorting distances

I. Introduction

Sorting problem is one of the fundamental problems in computer science which studied vastly in complexity theory and several efficient algorithms proposed for it. This problem has several applications in other problems and any result on this problem, directly affect the problems using it. Basic problems, like searching, finding the closest pair, constructing the minimum spanning tree, computing the convex hull, selecting kth smallest number which has vast applications fields like implementing search engines, making road maps, air-traffic control, computer graphics and robotics, all use sorting algorithms.

As we know, all (compare-based) sorting algorithms has $\Omega(n \log n)$ time complexity and therefore, other algorithms that use sorting, has the same lower bound on their complexity. The greedy algorithm for constructing geometric spanners is one of these algorithms [1], [2]. One point in some of these problems, like the greedy algorithm for computing geometric spanners, is that they need to sort some numbers that comes from distances between all pairs of input points. So the lower bound of sorting problem does not apply for sorting distances. So in this paper, we introduce a variant of sorting problem and try to solve it. However, we can not find the answer to the questions arise about this new problem, but we has a hope that one can use this technique to answer the questions.

Problem: Given a set $S = \{p_1, p_2, \ldots, p_n\} \subset \mathbb{R}^d$, sort all pairs (p_i, p_j) based on the Euclidean distance between them.

Using the general sorting algorithms, one can sort these $\binom{n}{2} = \Theta(n^2)$ pairs of points in $\mathcal{O}(n^2 \log n)$ time and the problem has $\Omega(n^2 \log n)$ lower bound, if the numbers are independent. But in this case, the numbers are not independent, they are distances between pairs of n input points. Now the question is that, is there an $o(n^2 \log n)$ algorithm to sort the distances or this problem has $\Omega(n^2 \log n)$ time complexity.

In this paper, we study this problem in its simplest case, when the point are from one-dimensional Euclidean space, i.e. real line. We tried to find an $o(n^2 \log n)$ time algorithm for this case, but we could not succeed. So, we try to show that this problem has $\Omega(n^2 \log n)$ time complexity.

II. Decision tree for sorting distances

To show a lower bound for the time complexity of the problem, we use a usual way as used in textbooks, see [3, Chapter 8]. In this way, we consider all permutations for sorting $\binom{n}{2}$ distances between n points on $\mathbb{R}$ and then we construct a decision tree on the permutations. The depth of the decision tree is the lower bound on the time complexity of the problem. The major issue in constructing such a decision problem is the following: because there is dependency between the distances between pairs of points, some of the permutations of the distances does not happen at all. This means that we have to remove them from the set of permutations. If the remaining permutation still is large enough, then we can show that the problem has $\Omega(n^2 \log n)$ time complexity, but if the size of remaining permutation is very low, there is a hope to find an algorithm with lower time complexity. So, in the following, we try to construct a decision tree such that, the leaves of the tree corresponds to the permutations that actually can occur in sorting distances.

A. Constructing the decision tree

In a decision tree for sorting n numbers, each leaf is corresponding to a permutation of the input. If the input numbers are independent, each permutation of the input should appear in the decision tree as a leaf. This is not the case for sorting distances, some of the permutations of the distances does not appear as a leaf of the decision tree. In this section, we try to remove these permutations from the list of all permutations.

To make a decision tree, we first consider the distance matrix and mention some of its properties. This matrix helps us to ignore unnecessary comparisons in sorting distances and only perform the comparisons which are necessary. Assume we have n points $p_1, \ldots p_n$ on the real line sorted from left to right (increasing order). Obviously, the distance between p_i and all the points after it appear in the sorted list of distances between pairs of points in order that we meet the points when we start at p_i and walk to the right. In other words, for each i,

$$|p_i p_{i+1}| < |p_i p_{i+2}| < \cdots < |p_i p_n|.$$

We define the distance matrix $D = (d_{i,j})_{n \times n}$ such that $d_{i,j} = |p_i p_j|$. Obviously the distance matrix D is symmetric. So the

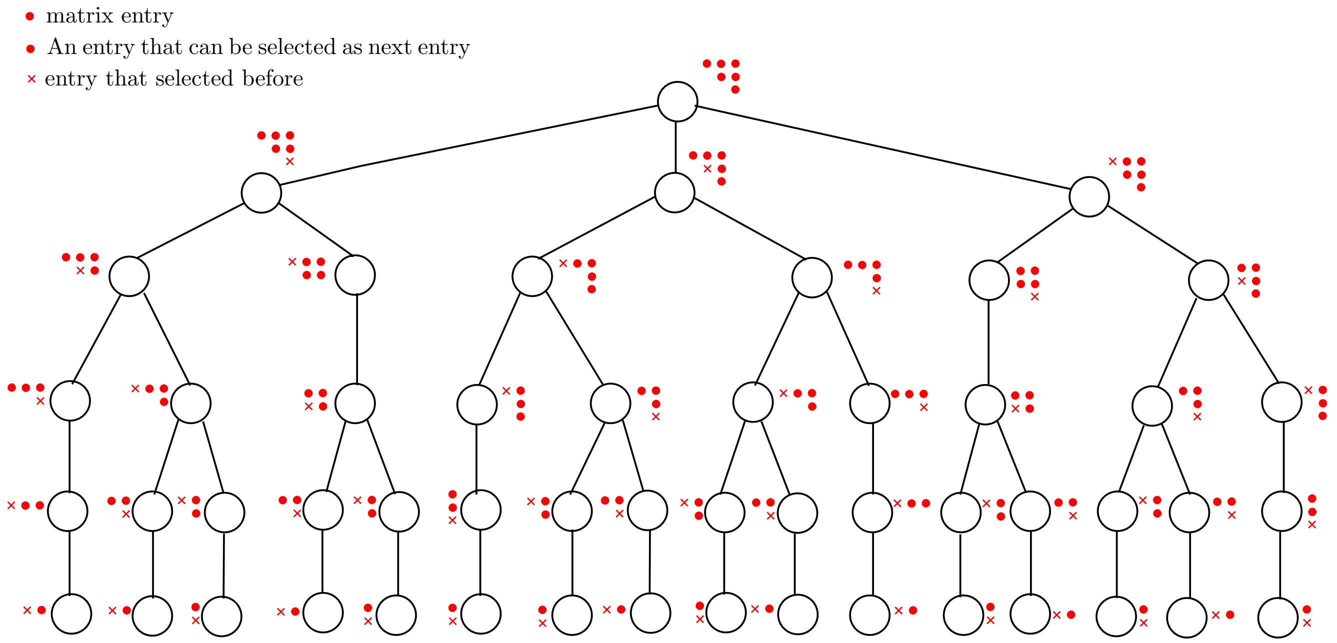

● matrix entry
● An entry that can be selected as next entry
× entry that selected before

Fig. 1: The Decision tree on a set of 4 points.

matrix D is as follows (the entries below the diagonal removed because of its symmetry):

$$D = \begin{array}{cc} & \begin{array}{cccccc} p_1 & p_2 & p_3 & p_4 & \cdots & p_n \end{array} \\ \begin{array}{c} p_1\\ p_2\\ p_3\\ \vdots\\ p_{n-1}\\ p_n \end{array} & \begin{pmatrix} 0 & d_{1,2} & d_{1,3} & d_{1,4} & \cdots & d_{1,n}\\ & 0 & d_{2,3} & d_{2,4} & \cdots & d_{2,n}\\ & & 0 & d_{3,4} & \cdots & d_{3,n}\\ & & & & \vdots & \\ & & & & 0 & d_{n-1,n}\\ & & & & & 0 \end{pmatrix}\end{array}$$

Since the points are on real line, the matrix D has the following properties:

- entries of each row increases as one one moves from left to right,
- entries of each column decreases as one one moves from top to bottom.

So in sorting distances, we do not need to compare entries that are in a row or a column, because we know their position in the sorted list based on their position in the matrix. As we will see in the next section, there are some other unnecessary comparison that does not have this property. Note that one can merge the rows of the matrix to get the sorted list of distances. This will gives us an $\mathcal{O}(n^2 \log n)$ algorithm because we have n lists and a total of $\Theta(n^2)$ numbers. In the rest of paper, we denote entities of the distance matrix by dots, because its value is clear from the position of the point in the matrix (see Fig. 2).

Because of the properties of the matrix D, the smallest (non-zero) entry of each row is the first element of the row after the entry on the diagonal of the matrix. So to find the smallest distance between pairs of points, it is sufficient to find the smallest element between these entries. We remove the smallest element from the list of candidates and report it as the first element of the sorted list and add the entry right after it to the candidate list. An important point here is that an entry of the matrix can be the next element of the sorted list, if there was no unselected entry on the left and below the entry.

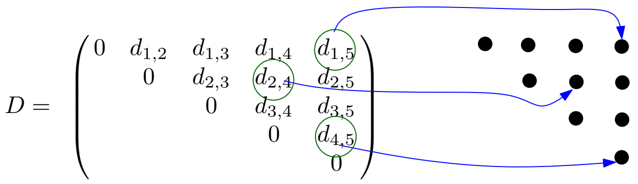

$D =$
0 $d_{1,2}$ $d_{1,3}$ $d_{1,4}$ $d_{1,5}$
0 $d_{2,3}$ $d_{2,4}$ $d_{2,5}$
0 $d_{3,4}$ $d_{3,5}$
0 $d_{4,5}$
0

Fig. 2: Representing the entries of matrix D.

As we mentioned before, leaves of a decision tree correspond to permutations of the (pairs of points) input such that this permutation of input can happen as a sorted list of the input. Obviously, all permutations of distances between the input points can not happen in the procedure of sorting. For example, any permutation that has $d_{1,3}$ before $d_{1,2}$ can not happen in sorting of distances, because $d_{1,3} > d_{1,2}$.

To bound the number permutation that may happen in sorting the distances, we construct a tree as follows. In the root of the tree, we do a comparison on the first entry of each row, which makes a list of $\mathcal{O}(n)$ elements. We choose the smallest element in the root and based on the element chooses in the root, we goes to the second level of the tree. Since any of the $n-1$ element of the first list can be the smallest one, we add $n-1$ children to the root. The second smallest element recognizes in the second level of the tree. The rest of the tree is constructed in a similar way. Fig. 1 shows the tree for a set of 4 points on the real line.

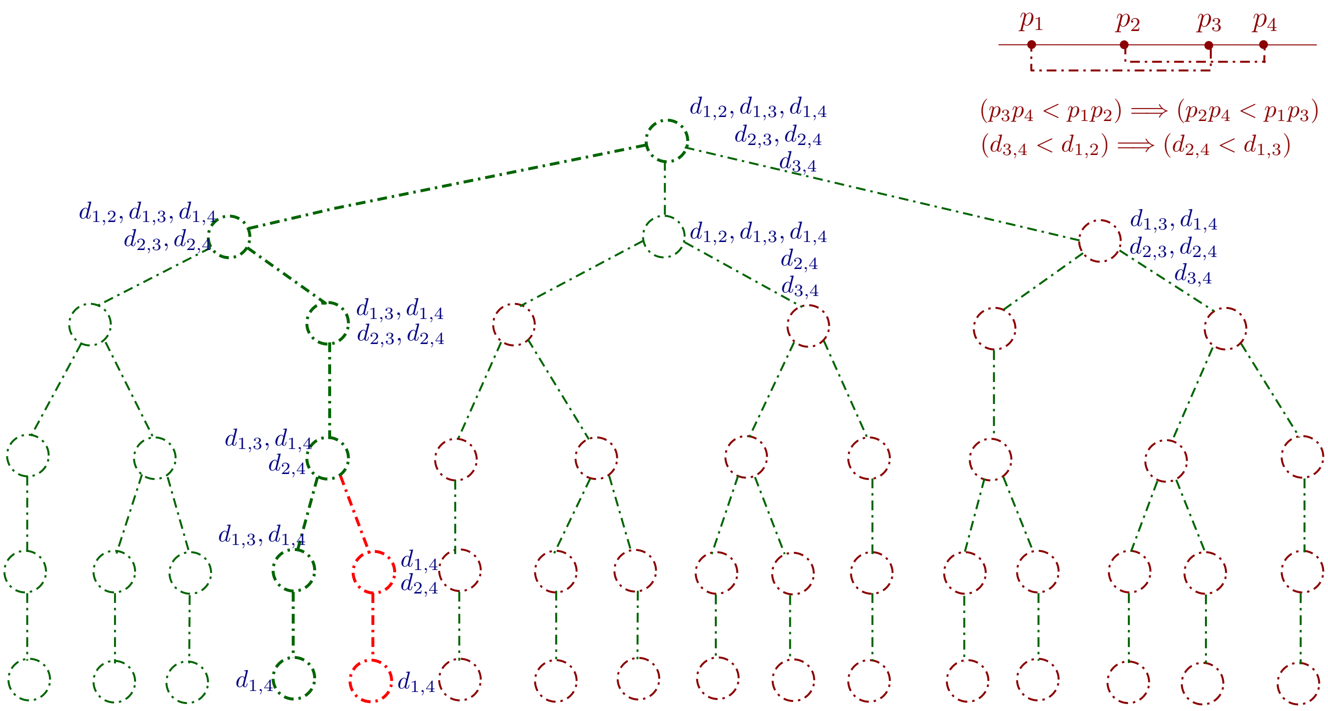

p_1
p_2
p_3
p_4
$(p_3p_4 < p_1p_2) \Longrightarrow (p_2p_4 < p_1p_3)$
$(d_{3,4} < d_{1,2}) \Longrightarrow (d_{2,4} < d_{1,3})$
$d_{1,2}, d_{1,3}, d_{1,4}$ $d_{2,3}, d_{2,4}$ $d_{3,4}$
$d_{1,2}, d_{1,3}, d_{1,4}$ $d_{2,3}, d_{2,4}$
$d_{1,2}, d_{1,3}, d_{1,4}$ $d_{2,4}$ $d_{3,4}$
$d_{1,3}, d_{1,4}$ $d_{2,3}, d_{2,4}$ $d_{3,4}$
$d_{1,3}, d_{1,4}$ $d_{2,3}, d_{2,4}$
$d_{1,3}, d_{1,4}$ $d_{2,4}$
$d_{1,3}, d_{1,4}$
$d_{1,4}$ $d_{2,4}$
$d_{1,4}$
$d_{1,4}$

Fig. 3: One unnecessary branch of the decision tree.

It is easy to see that we have all the possible permutations that can come-up at the end of sorting process in the decision tree. At the first view, one may claim that all the permutations that we have in this tree can happens in sorting distances, but this is not correct. As one can see in Fig. 3, for sorting distances between 4 points on the real line, after the comparison of $d_{3,4} < d_{1,2}$, we can conclude $d_{2,4} < d_{1,3}$ and so we do not need to compare them. So the red branch of the tree in Fig. 3 is not necessary. If we carefully check these red branches and then remove them from the tree, we can have only the necessary permutations.

In short, one should remove all unnecessary permutations of distances from the tree and find the number of remaining permutations. If this number is big enough, it gives the desired lower bound. Otherwise, one can have a hope that an $o(n^2 \log n)$ algorithm exists for sorting all distances between n points on the real line.

B. Computing the lower bound of the number of permutations

Based on the results in the previous section, all of the leaves of the tree are not necessary, but in this section we work on bounding the number of leaves in the tree.

Computing the exact number of permutations (or number of leaves of the tree) is difficult, because the degree of inner vertices of the tree are different (maximum degree is for the root which is $n - 1$ and minimum degree is 1), so we find a lower bound on the number of leaves of the tree. If this lower bound is of order of n^{n^2}, then we can conclude that the lower bound of the problem of sorting pairs of points is $\Omega(n^2 \log n)$. To bound the number of leaves from below, we find the minimum degree of vertices that lies in each level of the tree and then compute the lower bound by multiplying them.

Lemma II.1. *The difference between the degree of each node of the tree (except the root) and the degree of its parent is at most one.*

Proof: Consider an arbitrary node u of the tree and its corresponding matrix. The degree of the node u is equal to the number of elements in the set of candidate entries of the matrix that can be removed in the next step. In this step, we remove an entry from the candidate set. We have three cases:

Case 1: after removing the current entry, only one new element added to the candidate set (see Fig. 4A). In this case the degree of the node and its children is the same.

Case 2: after removing the current entry, two new elements added to the candidate set (see Fig. 4B). In this case the degree of children is one more than the degree of their parent.

Case 3: after removing the current entry, no new element added to the candidate set (see Fig. 4C). In this case the degree of children is one less than the degree of their parent.

So, we are done. ■

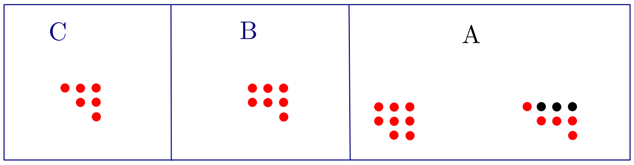


Fig. 4: Three case of the proof of Lemma II.1.

Now, we use this lemma to bound the number of leaves in the decision tree. So, in each level of the tree, we choose the node with minimum degree. On way that comes to mind is to start from the root and the by going down in the tree, we decrease the degree of the node in each step by one. This is true, but it can not go further than $(n-1)/2$th level, because if we go one step further from this level, the degree of the nodes will be zero which is unacceptable.

Considering the proof of Lemma II.1, we should choose the entries such that case 3 in the proof of the lemma happens, i.e., if we choose an entry, the next entry is chosen from a row which is at least two rows above or below the row of current entry's row (if we choose the next entry from the row just above or just below the current row, then the number of selection does not change in the next step). So, the best strategy is to choose elements from the every other row (see Fig. 5).

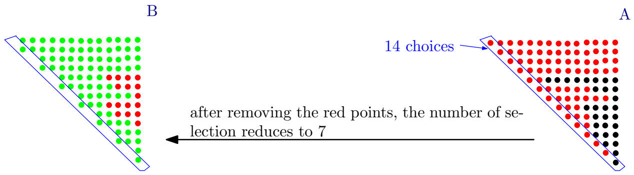

B
A
14 choices
after removing the red points, the number of selection reduces to 7

Fig. 5: The case with minimum selection for the next step.

As one can see in Fig. 5A, by removing the red entries in $(n-1)/2$ steps, the number of selection decreases to $(n-1)/2$ (the green entries in Fig. 5B). Therefore, in step $(n-1)/2+1$, if we choose any of the green entries, by Lemma II.1, the number of candidates in the next step increase by 1, i.e. $(n-1)/2+1$ selection.

As we mentioned before, we are looking for permutations (paths in the tree) such that we have the least selection for the next step. To this end, if one remove the entries showed in Fig. 7, in the order mentioned in the figure, we have $((n-1)/2)+1$, $(n-1)/2$, $(n-1)/2$ and $((n-1)/2)-1$ selection, respectively. After removing these four entries, the number of selection decreases by one and this decrease is irrecoverable. In a similar manner, we can remove the next 4 entries in similar situation which decreases the number of selections to $((n-1)/2)-2$. We continue removing the elements in this manner (see Fig. 6).

So, by summing up all of the degrees, the number of generated permutations are as follows:

$$\begin{aligned}
&\overbrace{\frac{n}{2}\times\cdots\times\frac{n}{2}}^{\frac{n}{2}\ \text{times}}\times\overbrace{\frac{n}{4}\times\cdots\times\frac{n}{4}}^{n\ \text{times}}\times\overbrace{\frac{n}{8}\times\cdots\times\frac{n}{8}}^{2n\ \text{times}}\times\cdots\times\\
&\qquad\cdots\overbrace{\frac{n}{2^i}\times\cdots\times\frac{n}{2^i}}^{n2^{i-2}\ \text{times}}\times\cdots\times\overbrace{\frac{n}{2^{\log n}}\times\cdots\times\frac{n}{2^{\log n}}}^{n2^{\log n-2}\ \text{times}}\\
=\;&(\frac{n}{2})^{\frac{n}{2}}\times\overbrace{(\frac{1}{2})^{n}\times(\frac{n}{2})^{n}}\times\overbrace{(\frac{1}{2})^{2n}\times(\frac{n}{2})^{2n}}\times\cdots\times\\
&\qquad\overbrace{(\frac{1}{2})^{n\times 2^{i-2}\times(i-1)}\times(\frac{n}{2})^{n\times 2^{i-2}}}\times\cdots\\
&\qquad\times\overbrace{(\frac{1}{2})^{n\times 2^{\log n-2}\times(\log n-1)}\times(\frac{n}{2})^{n\times 2^{\log n-2}}}\\
=\;&(\frac{n}{2})^{\sum_{i=1}^{\log n} n\times 2^{i-2}}\times(\frac{1}{2})^{\sum_{i=1}^{\log n} n\times 2^{i-2}\times(i-1)}\\
=\;&(\frac{n}{2})^{\frac{n}{2}\times(n-1)}\times(\frac{1}{2})^{\frac{n}{4}\times(4(1-n)+2n\times\log n)}\\
=\;&(\frac{n}{2})^{\frac{n}{2}\times(n-1)}\times(\frac{1}{2})^{n\times(1-n)}\times(\frac{1}{2})^{\frac{n^2}{2}\times\log n}\\
=\;&n^{\frac{n}{2}\times(n-1)}\times(\frac{1}{2})^{\frac{n}{2}\times(n-1)}\times(\frac{1}{2})^{n\times(1-n)}\times(\frac{1}{2})^{\log n^{\frac{n^2}{2}}}\\
=\;&n^{\frac{n}{2}\times(n-1)}\times(\frac{1}{2})^{\frac{n}{2}\times(1-n)}\times\frac{1}{n^{\frac{n^2}{2}}}\\
=\;&n^{-\frac{n}{2}}\times(\frac{1}{2})^{\frac{n}{2}\times(1-n)}=\frac{2^{\frac{n}{2}\times(n-1)}}{n^{\frac{n}{2}}}=(\frac{2^{n-1}}{n})^{\frac{n}{2}}\\
=\;&(\frac{2^{n-1}}{2^{\log n}})^{\frac{n}{2}}=2^{\frac{n}{2}\times(n-\log n-1)}.
\end{aligned}$$

The bound is much less that the one that gives us the desired lower bound. However, we believe it is possible to find a suitable lower bound using a more sophisticated analysis of the tree.

III. Conclusion and Future Works

In this paper, we studied the time complexity of sorting distances between pairs of n input points in the real line. We could not come-up with a new result, but it seems that more sophisticated analysis of the structure that proposed in this paper will show the complexity of the problem. Our conjecture is that this problem has $\Omega(n^2\log n)$ lower bound on time complexity.

Another way of attacking the problem is to reduce another problem with known time complexity $\Omega(n^2\log n)$ to this problem. There are several algorithm in the class of quadratic time complexity (see [4], [5]). However, we could not find a problem with $\Omega(n^2\log n)$ time complexity. It is also interesting if one study the problem in higher dimensions.

References

[1] G. Narasimhan and M. Smid, *Geometric spanner networks*. Cambridge University Press, 2007.

[2] P. Bose, P. Carmi, M. Farshi, A. Maheshwari, and M. Smid, "Computing the greedy spanner in near-quadratic time," *Algorithmica*, vol. 58, no. 3, pp. 711–729, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00453-009-9293-4

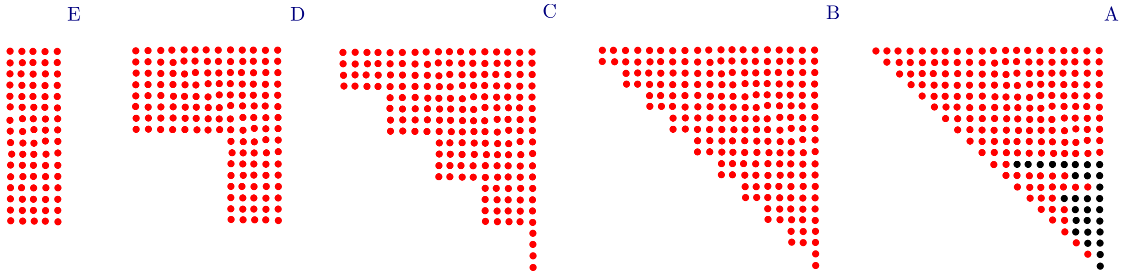


Fig. 6: The case with minimum selection for the next step.

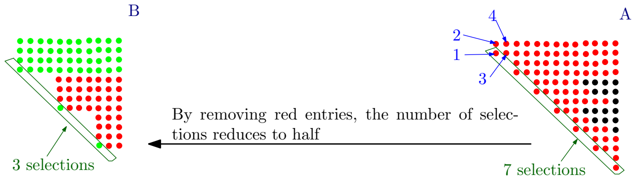

B
A
4
2
1
3
By removing red entries, the number of selections reduces to half
3 selections
7 selections

Fig. 7: The case with minimum selection for the next step.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. The MIT Press, 2009.

[4] A. Gajentaan and M. H. Overmars, "On a class of problems in computational geometry," *Computational Geometry*, vol. 45, no. 4, pp. 140 – 152, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925772111000927

[5] J. King, "A survey of 3sum-hard problems," 2004.

Theoretical and numerical characterization of continuously graded thin layer by the reflection acoustic microscope

Ahmed Markou, Hassan Nounah and Lahcen Mountassir
Laboratory of Metrology and Information Processing
Department of Physics
Faculty of Sciences, Ibno Zohr University, Agadir, Morocco

***Abstract*—This article presents a theoretical and numerical study by the reflection acoustic microscope of the surface acoustic waves propagation at the interface formed by a thin layer and the coupling liquid (water). The thin layer presents a gradient in its acoustical parameters along its depth. A stable transfer matrix method is used to compute the reflectance function of the surface acoustic modes radiated in the coupling liquid. This function is required to calculate the theoretical acoustic material signature which allows to determine the phase velocity of these modes. In order to characterize the influence of the gradient on the acoustic material signature, a few gradient functions are studied. The numerical results obtained show that the acoustic material signature can be used to characterize these profiles.**

Keywords*—*Reflection acoustic microscope; Surface acoustic wave; Thin graded layer; Transfer matrix method; Reflectance function; Acoustic material signature

I. Introduction

The acoustic microscope is an ultrasonic wave sensor, at the presence of an electrical excitation in its input allows to generate an acoustic wave capable to penetrate into the structures with a high resolution. This device is widely used in the various fields of nondestructive testing, such as microelectronics [1] and metallurgy [2] and biology [3], etc. The development of the first acoustic scanning microscope is motivated by the idea of using the acoustic field to study the elastic properties of materials with a resolution equivalent or better than that obtained by the optical microscope [4]. The acoustic material signature or V(z) curve has been used by Atlar et al. [5] to analyze the images obtained by the reflection acoustic microscope. They have obtained the V(z) curve by recording the voltage in the output of the transducer at each defocus z (distance between the acoustic lens and the sample). Weglein et al. [6] found that the periodic oscillations in the V(z) curve are linked to the surface acoustic wave propagation. Parmon and Bertoni proposed a simple formula, by using a ray model, to calculate the velocity of the surface acoustic wave from measuring the period of the oscillations in the V(z) curve [7]. The invention of the line focus acoustic microscope allows the quantitative measurement of the elastic properties and thicknesses of isotropic and anisotropic layered materials [8][9]. Today many techniques concerning the quantitative measurement by the acoustic microscope are developed to characterize the elastic properties of different solids. These techniques involve many type of acoustic microscope, citing for example: the conventional acoustic microscope and time-resolved acoustic microscope [10] and tow-lens acoustic microscope. For the conventional acoustic microscope, the acoustic beam is focused onto a sample by means of an acoustic lens. The transducer is located on the upper part of this lens and its other face is ground into a cavity with a quarter-wavelength matching layer. According to the shape of this cavity, we can distinguish tow conventional acoustic microscope types: line focus acoustic microscope (LFAM) which has a cylindrical cavity form. This lens focuses the acoustic wave along a line and the surface acoustic wave are excited in the direction perpendicular to the focal line. For the acoustic microscope with a spherical cavity shape, the wave is focused onto a spot with the dimensions comparable to the acoustic wavelength in the coupling liquid, which is placed between the lens and the specimen. In this study we will focus on the numerical investigation of the acoustic material signature by the reflection acoustic microscope with spherical cavity to evaluate the influence of the gradient function shape on V(z) curve and on its phase for the continuously thin graded layer on a solid substrate. The numerical example considered in this article is a thin graded Aluminum layer deposited in a solid Silicon substrate, where the Aluminum layer presents a continuous gradient in its elastic parameters and density, from its surface to a given depth (layer thickness). For modeling the inhomogeneous area, a simple gradient functions are used to describe the variation of the elastic properties according to the spatial coordinate. There are several methods for determining the reflectance function $R(f, \theta)$ relative to the inhomogeneous media, such as transfer matrix method and stiffness matrix method [11] and Peano series expansion [12], etc. The model adopted in this study to evaluate the surface acoustic wave propagation is based on the transfer matrix formalism which is detailed in the previous study [13]. The present study offers the possibility to characterize the gradient profiles in the elastic properties of a thin layer by the V(z) curve technique and by the dispersion curve of the surface acoustic wave. The first stage is to establish the reflectance function [13] and the second stage is to determine the expression of the V(z) curve (section II) and deducing the dispersion curve.

II. Background Theory

There are several methods for analyzing the V(z) curve in the output of a focus transducer for the solid structures and this section will expose Auld and Weise et al. and Zinin et al. approaches for determining V(z) curve expression

[14][15][16][17].
The figure 1 shows the defocused acoustic lens schema for a reflection acoustic microscope. The trajectory CO is related to the specular wave and $BFHD$ is the trajectory of the leaky surface acoustic wave (in the ray model). The leaky Rayleigh wave is excited by the ray BF, finally the trajectory $AO'E$ is relative to the edge wave . The acoustic lens is moved mechanically away the sample in the positive values of z. The following assumption is considered: the coupling liquid is parfait and then does not support transverse modes and that the sample is a solid plan reflector.

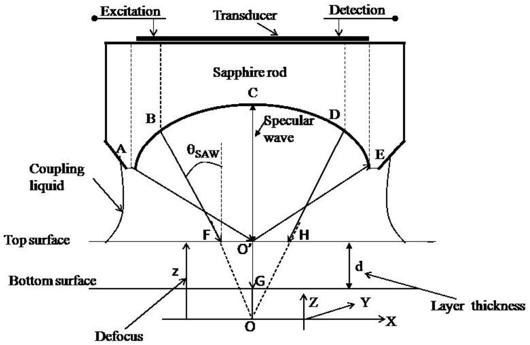


Fig. 1. Defocused acoustic lens in the reflection acoustic microscope

The output signal V of the reflection acoustic microscope at any position z can be expressed as [18]:

$$V = \int_{\infty}^{-\infty} \int_{\infty}^{-\infty} U_{in}(-k_1, -k_2) U_{sc}(k_1, k_2) k_3 dk_1 dk_2 \quad (1)$$

Where U_{in} and U_{sc} are, respectively, the Fourier spectrum of the incident and scattered acoustic fields. k_1 and k_2 and k_3 are the components of the wave-vector **k**, where $k = \omega / v_{liq}$. ω is the angular frequency and v_{liq} is the acoustic wave velocity in the coupling liquid. In the Debye approximation, the spectrum of the incident field at the focal plane can be written as:

$$U_{in} = \frac{P(k_1, k_2)}{k_3} \quad (2)$$

With $P(k_1, k_2)$ represents the pupil function, it is defined as the distribution of the field emitted before crossing the lens. The Fourier spectrum of the acoustic field reflected from a multi-layered system located between tow semi-infinite media is as following [19]:

$$U_{sc} = R(k_1, k_3) U_i(k_1, k_2) \quad (3)$$

At the object plane marked by z coordinate, the expression of the spectrum of the incident field is determined by introducing the propagation factor $\exp(jk_3 z)$. Taking into account the equation (1) and (2) and (3), the expression of the output signal V(z) is then:

$$\begin{aligned} V(z) = \int_{\infty}^{-\infty} \int_{\infty}^{-\infty} & P(-k_1, -k_2) P(k_1, k_2) \\ & \times R(k_1, k_3) \exp(2jk_3 z) dk_1 dk_2 / k_3 \end{aligned} \quad (4)$$

The expression of the V(z) in the equation (4) is relative to an anisotropic sample in the non-par-axial approximation. For an acoustic microscope with spherical lens and taking into account the spherical coordinates and the Debye approach which consider that $\sqrt{k_1^2 + k_2^2} \leq \sin\theta_m$. Where θ_m is the semi-aperture of the pupil. For symmetrical pupil function, the V(z) curve can be expressed as (the constant resulting from the second integral on the azimuthal angle ϕ is omitted) :

$$V(z) = \int_0^{\theta_m} P^2(\theta) R(\theta, \omega) \exp(2jkz\cos(\theta)) \cos\theta \sin\theta d\theta \quad (5)$$

$R(\theta, \omega)$ is the reflectance function of the acoustic wave propagating at the interface constituted by the coupling liquid and the graded layer. This reflectance is of great importance because it contains all informations about the modes which are reflected in the coupling liquid.

According to the ray model, which consider only the contribution of the normal rays to the lens surface of the acoustic microscope, the velocity of the surface acoustic wave v_{SAW} and the periodicity in the acoustic material signature V(z), for each given frequency f, are related by the following formula:

$$v_{SAW} = \frac{v_{liq}}{\sqrt{1 - (1 - \frac{v_{liq}}{2f\Delta z})^2}} \quad (6)$$

Where Δz represents the period of the oscillations in the V(z) curve. This period can be calculated by using the fast Fourier transform of the windowing V(z) curve.

III. Numerical results and discussion

A. Gradient functions

The numerical values of the ultrasonic velocities and densities of the layer and substrate [20] used in numerical simulations are regrouped in the following table:

TABLE I. Input data used in simulation

	$V_L(m/s)$	$V_T(m/s)$	$\rho(kgm^{-1})$	$d(\mu m)$
Silicon	8485	5850	2300	-
Aluminum	6374	3111	2700	10
water	1500	-	1000	-

The gradient functions used in simulations are:

$$\begin{aligned} f_l(x_3) =& f_0 + (f_d - f_0) x_3 \\ f_g(x_3) =& f_0 \exp(-\alpha x_3^2) \\ f_{tanh}(x_3) =& f_0 + (f_d - f_0) \frac{a_1}{a_2} \end{aligned} \quad (7)$$

Where f_l and f_g and f_{Tanh} are, respectively, the linear and Gaussian and tanh profiles, they describe the physical properties along the depth x_3. d denotes the thickness of the graded layer. f_0 and f_d are the acoustical parameters at the top surface of the layer ($x_3 = 0$) and at the substrate ($x_3 = d$). α and a_1 and a_2 are defined as the following:

$$\begin{aligned} \alpha =& \frac{1}{d^2} log(\frac{f_0}{f_d}) \\ a_1 =& tanh(b(x_3 - \frac{d}{2})) - tanh(b\frac{d}{2}) \\ a_2 =& tanh(b(\frac{d}{2}) - tanh(-b(\frac{d}{2})) \end{aligned} \quad (8)$$

With b is a given constant, it allows to provide different Tanh profile shapes. To approach the physical reality of a continuously graded layer, the thin layer is divided into a finite number n of homogeneous elementary layers. For each subdivision n and at a fixed frequency, the phase velocity of the leaky Rayleigh wave which propagates in the layer is determined. When this velocity becomes stable and remains unchanged with n, its convergence is reached. The asymptotic value reached by the phase velocity allows to estimate the relative error related to the convergence.

The figure 2 shows the longitudinal velocity profiles given by the functions in the relationship (7), where these velocity profiles vary between the tow values of the longitudinal velocities in the Aluminum (top surface of the layer) and in the Silicon (in the substrate). The transversal velocity and density are also varied in the direction perpendicular to the surface of layer and follow the same profiles as those presented in the figure 2.

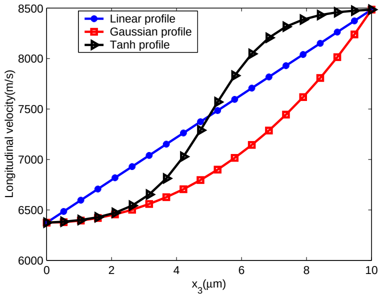


Fig. 2. Gradient functions simulating the gradient of the layer, $d = 10\mu m$

Each point in the figure 2 corresponds an homogeneous elementary layer with thickness equal to $0.5\mu m$. As mention this figure, the graded layer is discretized then into twenty elementary layers.

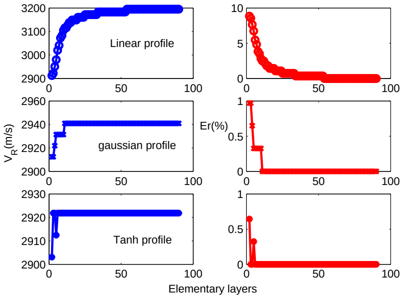


Fig. 3. Convergence of the velocity V_R (blue) and its error (red) versus the number of elementary layers n at f=1GHz

The figure 3 shows the convergence of the phase velocity of the leaky Rayleigh mode (blue curve) and its relative error (red curve) versus the number of elementary layers at the frequency of 1GHz. From analyzing this figure, it is clear that the number of elementary layers is justified by the convergence of the Rayleigh velocity. Indeed, for a linear profile and at the frequency of 1GHz for to have an error less than 1% on the phase velocity of the leaky Rayleigh mode, the graded layer should be slicing into about twenty elements (figure 3). For the other profiles like Gaussian and Tanh, the stabilization of the phase velocity of the leaky Rayleigh mode is reached when the layer is subdivided into about only ten elements. At low frequency (10MHz), the phase velocity of the leaky Rayleigh mode stabilizes at few elementary layers (n=2 for linear profile and n=4 for the Tanh profile). For the Gaussian profile, this velocity remains constant versus the number of elementary layers (figure 4).

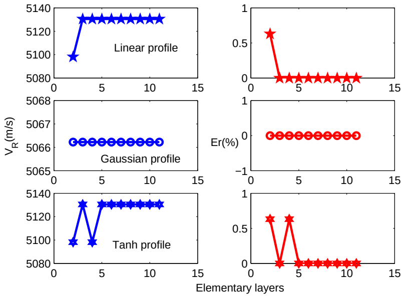


Fig. 4. Convergence of the velocity V_R (blue) and its error (red) versus the number of elementary layers n at f=10MHz

The following points explain the results in the figures 3 and 4:

- At high frequency (f=1GHz) and for great values of slicing n into elementary layers, the thickness $\Delta x_3 = d/n$ of each element becomes small with regards to the wavelength and gives almost continuous variation of the acoustic parameters Δf_i and then V_R becomes steady. For Δx_3 large (n is small), the properties (Δf_i important) of the layer vary discontinuously, thing that lead to a sharp variation in the velocity V_R
- At low frequency (f=10MHz), the wavelength of the leaky Rayleigh wave is very important in comparison with the global thickness of graded layer and then the depth of penetration of the acoustic wave is large and then the disturbance due to the gradient is negligible. This fact leads to the fast convergence of the velocity of the leaky Rayleigh mode V_R.

B. Acoustic material signature and dispersion curve

The images in the figures 5 and 6 (left figures) represent the phase of the reflectance function $R(f, \theta)$ plotted in the plan (f, θ). They show the dispersion of different surface acoustic modes radiated in the coupling liquid. The first mode in these images is called leaky Rayleigh mode, and the other modes are called Sezawa modes. As mention these figures, their phases vary between the tow values $-\pi$ and $+\pi$. In the dispersion curve relative to the linear profile (figure 5 left), the first sezawa mode presents a cut off frequency at the frequency of 280MHz where the phase velocity is about 5850 m/s (θ=14.84 degree), which is also the value of the transversal velocity in the substrate. The mode which corresponds to the phase velocity above the transversal velocity in the substrate is called pseudo-Sezawa mode. For the other profiles, the second mode is continuous and does not show the discontinuity observed in the dispersion curve relative to the linear profile. These images show also that the

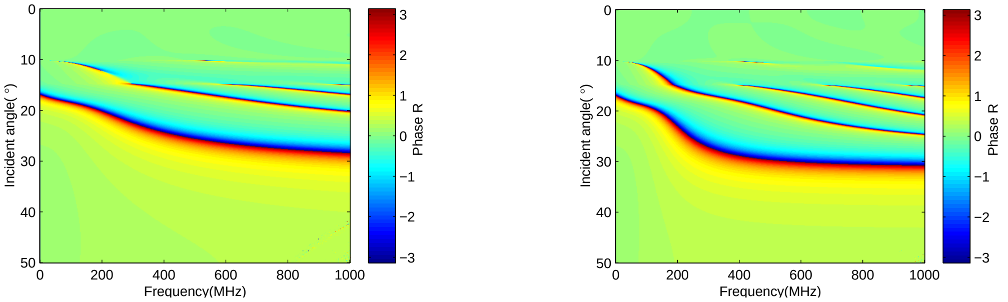


Fig. 5. Phase of the reflectance function $R(f, \theta)$ relative to the linear profile (left) and for the Gaussian profile (Right)

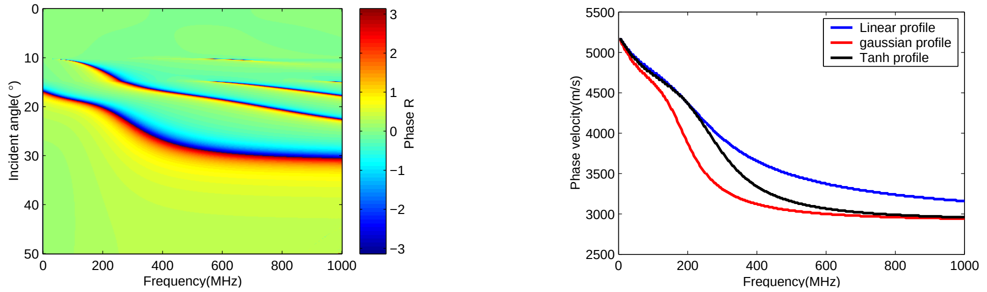


Fig. 6. Phase of the reflectance function $R(f, \theta)$ relative to the Tanh profile (left) and the dispersion curve of Rayleigh mode (right) for the three profiles

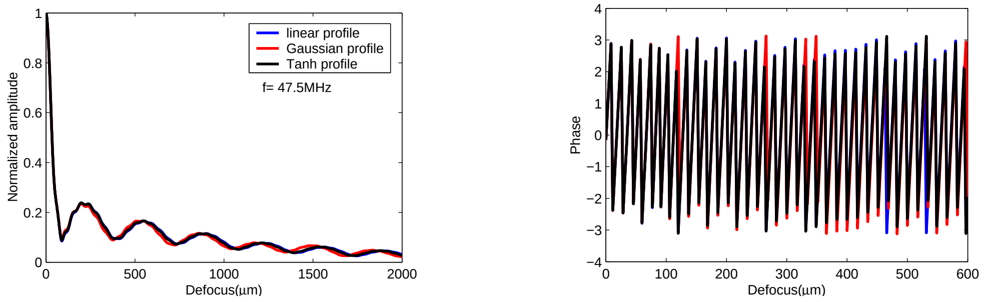


Fig. 7. Variations of V(z) curve (left) and its phase (right) relative to the three profiles simulating the graded layer at the frequency of 47.5MHz

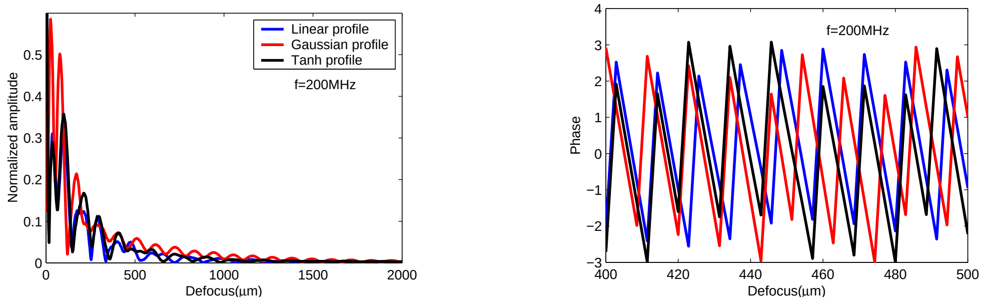


Fig. 8. Variations of V(z) curve (left) and its phase (right) relative to the three profiles simulating the graded layer at the frequency of 200MHz

number of the surface acoustic modes which appear in the dispersion curve differs from a profile to an other. The figure 6 (right) shows the dispersion curve of the leaky Rayleigh modes for the three gradient profiles, where the phase velocity of the modes vary between 5163 m/s at low frequency (2.5 MHz) and 3160 m/s for linear profile and about 2900 m/s for Tanh and Gaussian profiles at high frequency (1GHz). Below 100MHz the influence of gradient function is very less. This influence appears at high frequencies. From the frequency of 800MHz, the dispersion curve relative to the Gaussian and Tanh profiles are confused, this because of the gradient functions which are identical near to surface layer (figure 2). The figures 7 and 8 and 9 show the theoretical V(z) curve (left) and its phase (right) for the three gradient functions used in the simulations. They show the influence of the shape of the gradient functions on the V(z) curve and on its phase. At low frequencies (below 100MHz), this influence is almost null. Above the frequency of 100MHz, the effect of the gradient appears and a lag between the peaks in the phase of V(z) curve appears for the three profiles (figure 8 et 9 left). These peaks are relative to the leaky surface acoustic modes. It is also possible to study these profiles by the V(f) curve which

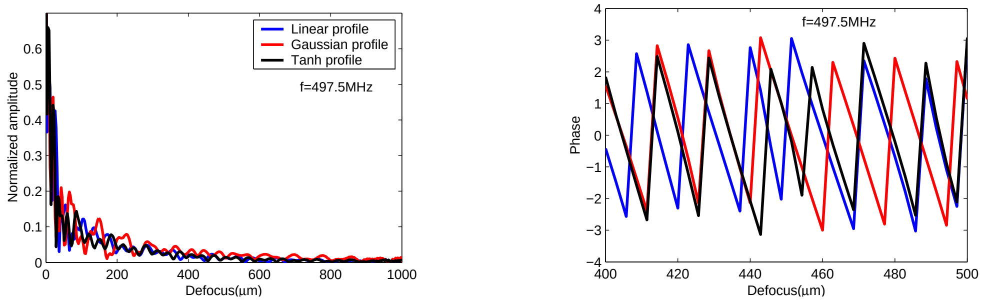


Fig. 9. Variations of V(z) curve (left) and its phase (right) relative to the three profiles simulating the graded layer at the frequency of 497.5MHz

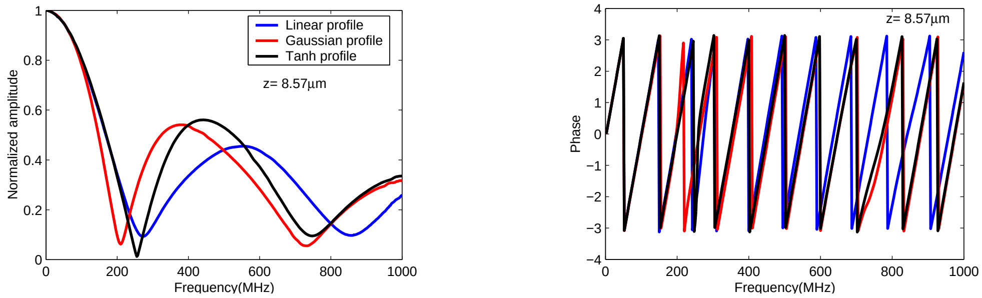


Fig. 10. Variations of V(f) curve (left) and its phase (right) relative to the three profiles simulating the graded layer at the defocus of 8.57μm

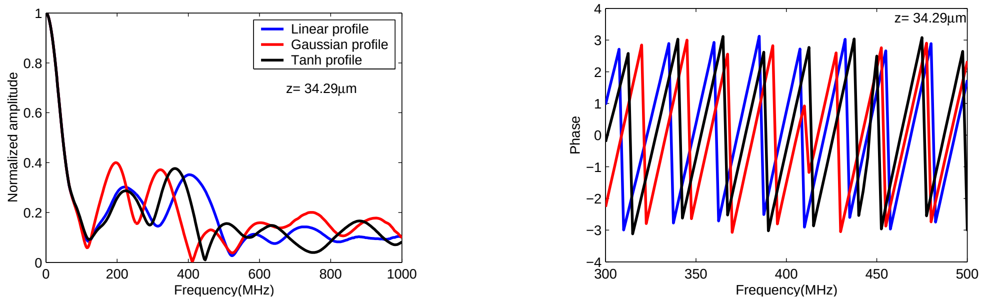


Fig. 11. Variations of V(f) curve (left) and its phase (right) relative to the three profiles simulating the graded layer at the defocus of 34.29μm

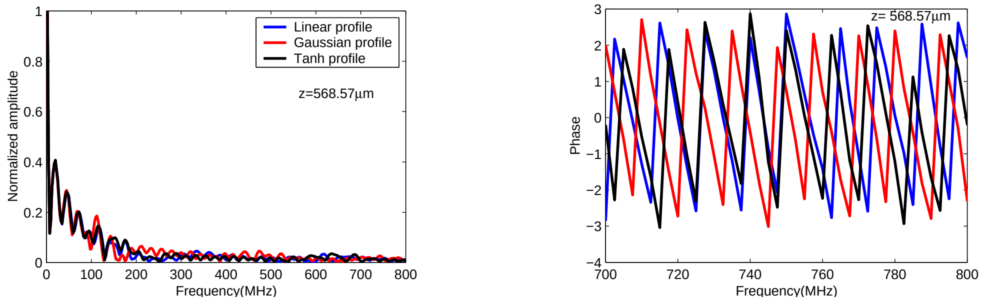


Fig. 12. Variations of V(f) curve (left) and its phase (right) relative to the three profiles simulating the graded layer at the defocus of 568.57μm

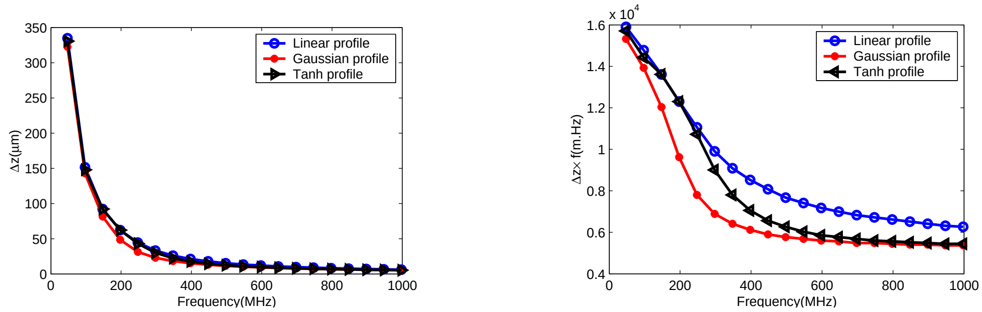


Fig. 13. Variations of the period Δz (left) and $f\Delta z$ (right), relative to the Rayleigh mode, with the frequency for the three gradient profiles

is calculated by setting the defocus z and by varying the frequency f in the integral of the equation 5. The V(f) curve (left) and its phase (right) are presented in the figures 10, 11 and 12. These figures show the variations of the V(f) curve and its phase versus the frequency at the fixed defocus. The difference between these curves, relative to the three profiles, is more pronounced above the frequency of 100MHz. For interpreting all these remarks, we use the dispersion curve in the figure 6 (right). Indeed, at the frequency of 100MHz, the phase velocity of the leaky Rayleigh mode, for the linear and Gaussian and Tanh profiles are, respectively, 4796 m/s and 4660 m/s and 4741 m/s which correspond to the wavelength of λ_{123}=(53.3μm, 51.78μm, 52.68μm). These wavelength are five times higher than the layer thickness (d=10μm). Then the acoustic wave "sees" the graded layer as homogeneous and then the effect of the gradient functions on the acoustic wave is negligible. When the wavelength, relative to the three profiles, approaches or exceeds the layer thickness (at f=300MHz, λ_{123} =(13.1μm, 11μm, 12.5μm)), the acoustic wave is more influenced by the graded area in the thickness of the layer and its propagation becomes dependent to the frequency.

The reflection acoustic microscope offers the possibility to characterize the gradient profiles through the analyze of the dispersion curve which can be determined from the V(z) curve by calculating different periods in this curve. The figure 13 shows the period Δz (left) and the quantity $f\Delta z$ (right) versus the frequency. For the leaky Rayleigh mode, Δz is calculated from the tow first successive minima at each frequency in the V(z) curve. The quantity $f\Delta z$ has the same shapes as the phase velocity V_R, but its values can be determined by the relationship (6) (figure 14).

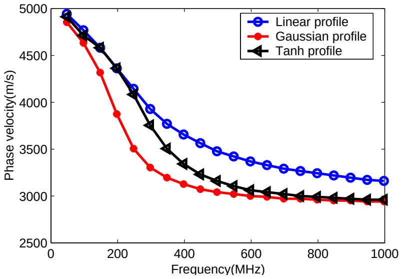


Fig. 14. Dispersion curve of V_R calculated from the equation 6

IV. Conclusion

This work presents a theoretical and numerical model of the characterization, by reflection acoustic microscope, of unidirectional continuous gradient profile in acoustical properties of thin layer. For to be close to the physical reality of a continuously inhomogeneous media, the graded layer is divided into a finite number of homogeneous elementary layers of equal thicknesses. This number is selected in such a way to have a compromise between the accuracy of results and the computing time. A number of twenty elementary layers ensures an error of 1% in the leaky Rayleigh velocity. This result is used to calculate the V(z) and V(f) curves according to the wave theory exposed in the section II. The interest of these tow curves is that they allow the determination of the dispersion curve of the surface acoustic wave by evaluating the periodicity of the V(z) curve at each frequency. The analyze of the dispersion curve of leaky Rayleigh mode and the V(z) curve, for the three gradient profiles, studying in this work, show a clear dependence of the acoustic wave propagation to the frequency when the thickness of the graded layer is the same order of magnitude as the acoustic wavelength.

References

[1] D. A. Hutt, D. P. Webb, K. C. Hung, C. W. Tang, P. P. Conway, D. C. Whalley and Y. C. Chan, *Scanning Acoustic Microscopy Investigation of Engineered Flip-chip Delamination* , 26th IEEE/CPMT International Electronics Manufacturing Technology Symposium, California, pp.191-199, 2000.

[2] C. Pecorari, P. B. Nagy and L. Adler, *Acoustic microscopy to study grain structure*, Review of Progress in Quantitative Nondestructive Evaluation. Vol. 09, Plenum Press : New York, 1990.

[3] C. F. Quate, A. Atalar and H. K. Wickramasinghe, *Acoustic Microscopy With Mechanical Scanning-A Review*, Proceedings of IEEE. Vol. 67 No. 8, 1979.

[4] R. A. Lemons and C. F. Quate, *Acoustic Microscopy*, Physical Acoustics Vol. XIV, Mason, London Academic, 1979.

[5] A. Atalar, C. F. Quate and H. K. Wickramasinghe, *Phase imaging in reflection with the acoustic microscope*, Appl. Phys. Lett. 31(12) p. 791-793, 1977.

[6] R. D. Weglein and R. G. Wilson, *Characteristic material signatures by acoustic microscopy*, Electron.Lett .14(12) pp.352-354, 1978. .

[7] W. Parmon and H. L. Bertoni, *Ray interpretation of the material signature in the acoustic microscope*, Electron.Lett.15(12) pp. 684-686, 1979.

[8] J. Kushibiki, A. Ohkubo and N. Chubachi *Linearly focused acoustic beams for acoustic microscope*, Electron.Lett.17(15) pp. 520-522, 1981.

[9] J. Kushibiki, A. Ohkubo and N. Chubachi *Anisotropy detection in sapphire by acoustic microscope using line-focus beam*, Electron.Lett.17(15) pp. 534-536, 1981.

[10] E. C. Weiss, P. Anastasiadis, G. Pilarczyk, R. M. Lemor and P. V. Zinin, *Mechanical Properties of Single Cells by High-Frequency Time-Resolved Acoustic Microscopy*, IEEE Transactions On Ultrasonics, Ferroelectrics, and Frequency Control, Vol.54, No.11, pp.2257-2271, 2007.

[11] X. Deng, T. Monnier, P. Guy and J. Courbon, *Acoustic microscopy of functionally graded thermal sprayed coatings using stiffness matrix method and Stroh formalism*, J.Appl.Phys.113(22), pp.224508:1-10, 2013.

[12] C. Baron, *Series expansion of the matricant for studying the elastic waves propagation in the mediums with continuously variable properties*, PhD dissertation, Bordeaux, France, 2005.

[13] A. Markou and H. Nounah, *Numerical Evaluation of the Effect of Gradient On reflection Coefficient of Continuously Graded Layer*, IJACSA, Vol.6(5), pp.97-102, 2015.

[14] A. Atalar, *An angular spectrum approach to contrast in reflection acoustic microscopy*, J. Appl. Phys. 49(10), pp.5130-5139, 1978.

[15] B. A. Auld, *general electromechanical reciprocity relations applied to the calculation of elastic wave scattering coefficients*, Wave Motion 1:3-10, 1979.

[16] W. Weise, P. Zinin and S. Boseck, *Angular spectrum approach for imaging spherical-particles in reflection and transmission scanning acoustic microscope*, Acoustical imaging, pp.707-712 vol.22, Plenum: New York, 1996.

[17] P. Zinin, W. Weise, O. Lobkis and S. Boseck , *The theory of three-dimensional imaging of strong scatterers in scanning acoustic microscope* , Wave Motion 25(3):213-236, 1997.

[18] A. Atalar, *A backscattering formula for acoustic transducers*, J. Appl. Phys.51(6)pp. 3093-3098, 1980.

[19] W. Weise, P. Zinin and S. Boseck *Modeling of inclined and curved surfaces in the reflection scanning acoustic microscope*, J. microsc.176(3) pp. 245-253, 1994.

[20] H. Nounah, *Modeling and characterization of materials with gradient in the mechanical properties by the micro acoustic methods*, PhD dissertation, Montpellier II, France, 1995.

Toward Information Diffusion Model for Viral Marketing in Business

Lulwah AlSuwaidan
Information Systems Department
Collage of Computer and Information Sciences
King Saud University
Collage of Computer and Information Sciences
Al-Imam Muhammad Ibn Saud Islamic University
Riyadh, Saudi Arabia

Mourad Ykhlef
Information Systems Department
Collage of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

***Abstract*—Current obstacles in the study of social media marketing include dealing with massive data and real-time updates have motivated to contribute solutions that can be adopted for viral marketing. Since information diffusion and social networks are the core of viral marketing, this article aims to investigate the constellation of diffusion methods for viral marketing. Studies on diffusion methods for viral marketing have applied different computational methods, but a systematic investigation of these methods has limited. Most of the literature have focused on achieving objectives such as influence maximization or community detection. Therefore, this article aims to conduct an in-depth review of works related to diffusion for viral marketing. Viral marketing has applied to business-to-consumer transactions but has seen limited adoption in business-to-business transactions. The literature review reveals a lack of new diffusion methods, especially in dynamic and large-scale networks. It also offers insights into applying various mining methods for viral marketing. It discusses some of the challenges, limitations, and future research directions of information diffusion for viral marketing. The article also introduces a viral marketing information diffusion model. The proposed model attempts to solve the dynamicity and large-scale data of social networks by adopting incremental clustering and a stochastic differential equation for business-to-business transactions.**

Keywords*—*information diffusion; viral marketing; social media marketing; social networks

I. Introduction

Online social networks have become part of our daily lives and hundreds of millions of accounts have spread over different social media channels. The spread and development of the Internet and mobile technology has affected the growth of social networks. For instance, Twitter has half a billion users worldwide who produce around 175 million tweets per day or 10,000 tweets per second [1], [2]. The massive volume of information traded over social networks should be analyzed and used for marketing purposes. The study reviews diffusion methods including data mining, evolutionary methods, and statistics. It will also introduce the viral marketing information diffusion (VMID) model that aims to utilize data mining and a stochastic differential equation to achieve the best possible diffusion outcomes.

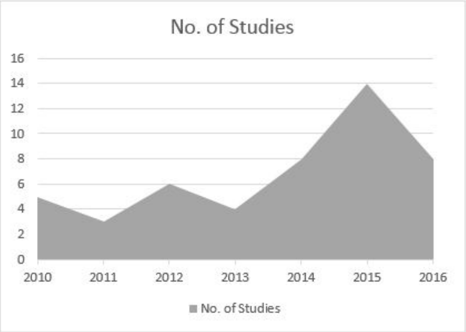


Fig. 1: Research on information diffusion for viral marketing, 2010–2016.

According to Bennett [3], around 70% of customers trust purchasing recommendations from their friends or close relatives, and only 34% under the age of 30 view product ads on television. Also, 87% of marketers say they would like to know how to measure the return on investment (ROI) through online social networks [3]. Brown and Fiorella [4] raised the question of why businesses should be on social media. Marketers should ask the same question before starting a marketing campaign via a social network. Other relevant questions include “Why should we choose this certain social media channel instead of others?” and “Which users have the power to influence other users?”

To spot new research advancements in information diffusion for viral marketing, we performed a search for various topics related to information diffusion for viral marketing in business. The search encompassed the last few years. Fig.1 shows the recent enormous increase in research interest on the topic. This motivates a depth discussed around it in the following sections.

This article asks the following research question: “Are there methods or techniques that consider information systems to enhance and optimize the marketing message diffusion?” The goals of this article are to find gaps and limitations in the existing research and to provide new insights into the topic. All the studies reviewed in this article consider a computer and information systems perspective. The answers to the research

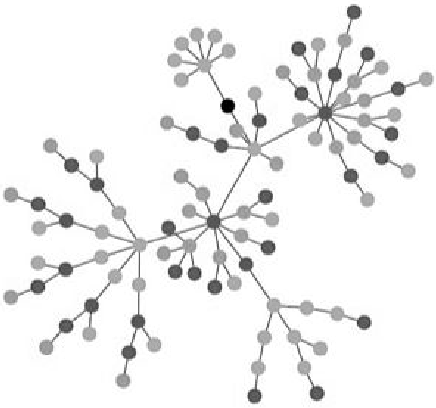


Fig. 2: Graph representation of social networks.

question led to the creation of the VMID model to fill the gap in existing research.

The remainder of this article is organized as follows: Section II provides a general background of online social networks and viral marketing. Section III discusses the existing information diffusion models in viral marketing and divides them into four according to their role: synchronous, asynchronous, influence maximization, and social network mining. Section IV discusses social media marketing in business. Section V presents the methodology and discusses the current challenges to diffusion in online social networks for viral marketing. It also introduces the proposed VMID model. Section VI concludes the article and provides recommendations for further research.

II. Background

This section provides a background of the two core concepts of this article: online social networks and viral marketing. Online social networks are the underlying environment and platform for the diffusion of viral marketing messages.

A. Online Social Networks

Social Network can be represented as graph. Easley and Kleinberg [6] defined a graph as "a way of specifying relationships among a collection of items. A graph consists of a set of objects, called nodes, with certain pairs of these objects connected by links, called edges." Social networks can be represented as a set of nodes and edges forming a network or graph, where the nodes are the participants and the edges are the types of connections. More specifically, in social networks, people or groups of people are the nodes, and the types of social interaction they engage in are the edges. Attention should be directed toward the connectedness in these networks. Easley and Kleinberg [6] emphasized this through two fundamental observations: interconnecting links and interdependence in users' behavior. Kempe, Kleinberg, and Tardos [7] also described a social network as "the graph of relationship[s] and interactions within a group of individuals." Fig. 2 illustrates the graph structure, which demonstrates the general case (unweighted and undirected).

There is much interweaving between static and dynamic social networks; they all have the same structure of representation of nodes and relationships. However, The nodes and their relationships continuously change over time in dynamic networks. The nodes and relationships in dynamic social networks are generated once there is direct communication or information flow between two or more nodes. Kempe et al. [7] showed that successful marketing is achieved by dealing with the information dynamics in the network rather than studying the structural properties of the graph. This has directed interest toward dynamic social networks, unlike most of the existing research that focuses on static networks [8]–[10]. This leads to a need to examine both the dynamics of information and the network structure to achieve the best marketing strategy.

B. Viral Marketing

Online social networks have become part of people's daily activities; therefore, companies and organizations tend to harness the rapidity provided by social networks. Viral marketing is a marketing method that uses social networks. It is based on people who have influence over their relatives and friends in social networks. Researchers can measure the influence of these people by nodes and ratios of propagation in online social networks. Companies target these influencers because marketing through them will have a great impact on their products. The most crucial part of viral marketing is choosing the right influential nodes, called seeds, and then choosing the best diffusion method.

Viral marketing is a way of advertising products to a specific group of interested people using online social network relationships. Long and Wong [9] defined it as "targeting a limited number of users (seeds) in the social network by providing incentives, and these targeted users would then initiate the process of awareness spread by propagating the information to their friends via their social relationships." The viral marketing phenomenon is described as influence-spreading over social networks [11]. The process of viral marketing, influence marketing, or WOM is divided into three stages: 1) initiating the advertising message , 2) locating the best seeding nodes, and 3) diffusing the marketing message to others [12].

The complexity of social network structures requires that marketers build and follow a procedure to spread information accurately. Social networks concentrate on human behavior such as opinions, recommendations, and reviews; understanding this behavior and its consequences is a key factor to success in viral marketing [8], [9]. Generally, the viral marketing process consists of two main stages: seeding and diffusion [9]. Kitsak et al. [13] found that the core spreaders are more effective than those that have more connections. Viral marketing targets a limited number of users to begin marketing; these users have a sufficient impact on another group of interested users. Choosing the optimal seeding is an NP-hard problem, as proven by [7]. Minimizing the number of seeds to reduce the overall cost is the objective of viral marketing. To satisfy this objective, Long and Wong [9] introduced and studied the J-MIN-Seed problem. Since this review concentrates on information diffusion in viral marketing, the following section will provide a detailed discussion of information diffusion methods.

III. Information Diffusion

The Oxford Dictionary defines diffusion as "the spread of something." In the analysis of social networks, diffusion is the process of information diffusion via the network. The majority of current research focuses on information diffusion in online social networks [14]–[16]. Viral marketing and diffusion in online social networks are largely interwoven because they share the same underlying environment, diffuse the same content, and have the same objectives. The basic idea of information diffusion in online social networks originated from the concept of virus spread. Most epidemiological diffusion models are considered non-graphical approaches because they assume no stability in the network structure [15]. Newman [17] categorized epidemiological models into two: the susceptible-infected-removed model (SIR) [18] and the susceptible-infected-susceptible model (SIS) [19]. Another diffusion approach is based on rumors spread. Boccalettia, Latorab, Morenod, Chavez, and Hwang [20] classified rumor receivers into ignorant (does not care about it), spreader (willing to spread it), and stifler (ceases to spread it).

This section divides the models according to the technique used: synchronous, asynchronous, influence maximization, and social network mining.

A. *Synchronous Diffusion Models*

This section discusses the two dominant diffusion models: linear threshold (LT) and independent cascade (IC). These models assume a discrete time event and a set of predefined and activated nodes. The diffusion process is mainly based on the probability defined over each edge and how one node influences others to diffuse information. Kempe et al. [7] proposed a generalized version of each model and showed that they are mostly equivalent. For this reason, Lu, Wen, and Cao [21] developed a community-based algorithm and a distributed set-cover algorithm based on the probabilistic diffusion model because of the limited ability of LT and IC in large-scale networks.

1) Linear Threshold Model: This model relies on randomly choosing a threshold θ between 0 and 1 for each node [22], [23]. Each node i has a weighted edge between this node and all of its neighbors, or $W_{i,j}$. If the total weight of all the adjacent nodes satisfies the following condition, then the node becomes active:

$$\sum_{j=1}^{n} W_{i,j} \geq \theta_i \tag{1}$$

where n is the total number of edge weights of active adjacent nodes. In particular, at time t, all the active nodes in time $t-1$ remain active. All the nodes that satisfy the above condition also become active [7], [11], [15].

There are number of studies worked under LT model to solve problems related to diffusion [24]–[27]. The work done by Galuba, Aberer, Chakraborty, Despotovic, and Kellerer [24] aimed to predict an information cascade graph. Chen, Yuan, and Zhang [25] conducted a study of influence maximization in an LT model. They introduced the first scalable heuristic algorithm aimed at maximizing influence within the LT model, called the local directed acyclic graphs (LDAG) algorithm. Khalil, Dilkina, and Song [26] proved that the LT model has an enormous effect on the problem of network modification. However, Guisheng, Jijie, and Hongbin [27] claimed that the LT model was inefficient in real-life social networks. Therefore, they introduced the cellular automaton-based network diffusion (CAND) model.

TABLE I: Differences and similarities between LT and IC models

Dimensions		Linear Threshold	Independent Cascade
Similarities	Network structure	Static	
	Network links	Directed	
	Model processing	Iterative and synchronous (early adoption)	
Differences	Edge measures	Influence degree	Diffusion probability
	Node measures	Influence threshold	No specific node measure required
	Communication pattern	Receiver-centric	Sender-centric

2) Independent Cascade Model: The IC model is a mathematical model for the diffusion process in directed graph $G= < V, E>$, where V represents nodes and E represents connected links [11], [28]. The model uses the probability defined over each edge that connects the nodes [29]. The activated nodes try to activate inactive nodes using the probability value of each edge [15]. The model initially defines a random set of active nodes to activate the inactive nodes connected to them. Similar to the LT model, the activated nodes in time t-1 remain active in time t [7], [23]. The process starts using the active node i, which activates the connected inactive node j.

Studying the node strength is a challenge because there is no unique measurement for it. Arnaboldi et al. [16] claimed that knowing the node strength will assist in inferring the information diffusion. Zhu, Wang, Wu, and Zhu [30] proposed SpreadRank, which measures the spread ability of each node. It is a generalized version of the diffusion model CTMC-ICM, which introduces the theory of continuous-time Markov chain (CTMC) into the IC model. A number of studies have enhanced the IC model, such as that of Nazemian and Taghiyareh [31], who proposed the IC with positive and negative WOM (ICPM).

TableI summarizes the main differences and similarities between the two diffusion models. Both models are applied on static and directed graphs. The main difference between the two is the diffusion method. The LT model uses influence degree, while the IC model uses probabilities. IC has greater coverage than LT because most of the graph nodes are covered. On the other hand, LT is more accurate than IC because only the nodes that satisfy the condition/rule are involved in the diffusion process. Although both models work effectively, many enhancements have been made to increase their performance, especially regarding time.

B. *Asynchronous Diffusion Models*

The methods presented in this section have been used in viral marketing diffusion and information diffusion in online social networks. This section will discuss asynchronous methods that could have a significant impact on dynamic viral marketing diffusion. Saito, Ohara, Yamagishi, Kimura, and Motoda [32] proposed an estimated parameter based on information diffusion resulting at a specific time; this ensures an asynchronous pattern in the same diffusion models (AsIC and AsLT) [33], [34]. Opinion propagation was one of the problems solved by Kimura et al. [34] via an extension of AsIC and AsLT. They created the extension using a value-weighted voter model with multiple opinions. Similarly, using the same learning performance model, Kimura, Saito, Ohara, and Motoda [35] used the value-weighted voter model to detect anti-majority opinions in social networks.

The T-BaSIC model presented by Guille and Hacid [36] predicts the temporal dynamics of diffusion in social networks. The approach is based on machine learning techniques and the inference of time-dependent diffusion probabilities from a multidimensional analysis of individual behaviors. In addition, Wonyeol, et al. [37] generalized the IC model by including two important aspects, continuous trials and time restriction, in a new model called the continuously activated and time-restricted IC (CT-IC) model. This is an influence diffusion model dedicated to viral marketing. They also proposed the continuously activated and time-restricted independent path algorithm (CT-IPA) to compute the exact influence spread path. Zhu et al. [30] presented one of the applications of the IC model. They introduced a new method that enhances the influence maximization problem by concentrating on a small group of influences.

C. Influence Maximization

The majority of current research considered static network topology. A static network topology assumes that the structure remains the same over time. The literature has begun to examine the effect of choosing the best initial influential nodes. Kempe et al. [7] identified the challenge of optimizing the selection of the most influential nodes in the network and introduced an approximation algorithm for influence maximization. The problem with their approach is the time required for simulation processing. The limitation of the approximation algorithm [7] was its ineffective performance in large-scale mobile social networks, especially in dealing with the diffusion minimization problem. The challenge is choosing the best set of users who can maximize spread in the network. Chaudhury et al. [14] found that their method of spreading information was faster than the greedy k-center method. They proposed the degree-based scaling method applied in the graph. The method aimed to increase the active set of nodes that have the topmost degree with the least possible amount of time. Both algorithms ensured optimal seeding regardless of the amount of time consumed. Abadi and Khayyambashi [38] discussed the problem of influence maximization with viral marketing and introduced a new algorithm. Their proposed algorithm aimed to select the experts and leaders in social networks based on spatiality and knowledge. Although they claimed that the algorithm was quick, it required long processing steps to produce reasonable results. Zhou, Zhang, and Cheng [39] addressed the influence maximization problem by integrating greedy algorithms and mining the top influences. They proposed GAUP to mine the most influential nodes in the network. In addition, Kim, Beznosov, and Yoneki [40] introduced a decentralized influence maximization problem by influencing k-neighbors rather than randomly selected users in the network.

All the aforementioned studies focused on maximizing influence by concentrating on the social network structure. Research on social network dynamics has been increasing. For instance, Zhuang, Sun, Tang, Zhang, and Sun [41] concentrated on maximizing influence in dynamic social networks. In response to dynamic challenges, they proposed an influence maximization algorithm called Maximum Gap Probing (MaxG).

D. Social Network Mining For Viral Marketing

Researchers have only recently begun to examine systematically the effect of applying mining methods such as clustering or classification to viral marketing. This section explores mining methods for viral marketing.

1) Clustering: In general, clustering groups objects that have similar attributes. Han, Kamber, and Pei [42] defined clustering as "the process of partitioning a set of data objects into subsets." The subset represents a cluster; each cluster contains objects that are similar to each other and different from others in another cluster. Among the current trends in mining social networks for viral marketing, clustering analysis methods have gained the most interest because of their ability to deal with a social network's structure, correlations, and groups. The idea of clustering as a machine learning model is to identify homogenous groups of people [43].

Few studies have considered clustering for diffusion in viral marketing. Banerjee, Al-Qaheri, and Hassanien [44] combined fuzzy logic, game theory, and clustering within social network mining for viral marketing. Since marketing in social networks depends on the dynamics of social influence interaction, Banerjee et al. [44] applied game theory. Clustering was used to determine the representative group of customers. Sharma and Shrivastava [45] used clustering to identify the highly influential nodes for viral marketing. Two clusters were used as a strong tie cluster and weak tie cluster. They also introduced two algorithms: cluster mining and influence mining. AlSuwaidan, Ykhlef, and Alnuem [46] introduced a novel spreading framework for viral marketing that ensures optimization of cost and time and predicts the required cost and time to reach sufficient coverage. The framework is based on incremental clustering and activity networks and is directed toward the most active user in the social networks. Among the works presented in the literature that have discussed the information diffusion in online social networks, Niu, Long, and Li [47] used a k-means algorithm to cluster users into different groups. They also proposed a continuous time diffusion model by incorporating users' heterogeneous temporal diffusion patterns. This model is also applicable to viral marketing diffusion because it is applied on the same constrained environment and has the same objective

Some other methods, such as influencer selection, use clustering for identification problems. Lou and Tang [48] developed a model for mining top-k structural hole spanners in large-scale social networks. The general idea behind this was to measure how a node bridges different communities. This was the first attempt to prove the NP-hardness of maximizing the decrease of minimal cut in an unweighted graph. Shrihari, Hudli, and Hudli [49] proposed an approach for identifying influencers that uses a k-means clustering algorithm. The identification process of opinion leaders (influencers) was based on many factors such as the amount of time they spend in social networks, their positive comments, and their responses to others.

2) Classification: Classification is a supervised modeling technique. The majority of produced classification methods attempt to predict or estimate future events based on historical data. What is important to note in modeling using classification is that the target groups or classes are known from the beginning; therefore, classification techniques are not useful in mining social networks [50]. However, most classification studies on mining social networks for diffusion consider classification as a method for making classes or organizing user behavior or properties before the diffusion process occurs. Surma and Furmanek [51] used classification in mining social networks for viral marketing. They used a classification and regression tree (C&RT) model to identify users of online social networks who will respond to marketing campaigns.

Classification has also been used for predicting user behaviors in online social networks. This simplifies the diffusion process because it determines the best possible node to start from based on past behavior patterns. Ortigosa, Carro, and Quiroga [52] used classification as a mining method for predicting user personality based on interaction behavior patterns. The prediction process was based on parameters such as number of friends and wall posts. Ortigosa et al. [52] developed an application on Facebook called TP2010 to gather information from approximately 20,000 users to determine their interaction personality patterns. They used the classifier as a machine learning technique to observe interaction patterns among users

3) Community Detection: Community detection is one of the most widely used methods in the literature on analyzing social influence. It relies on the idea of targeting the communities that are most closely related to a certain domain through social networks. Viral marketing has the same objective, since locating the right community is the first step toward effective viral marketing. However, Most of existing methods perform community detection at random basis [53]. In this section, a review of existing community detection methods and models that have been applied to online social networks will be presented to emphasize the importance and current lack of applying community detection to viral marketing.

The focus of community detection is on the community structure because of its usefulness in knowing how the network is structured [54]. Many researchers have raised questions about the definition of community and its structure. Shen [55] highlighted the need for a clear definition of community; he claimed that this definition is dependent on its context and applications. Tang and Liu [56] and Shen [55] defined community as a group of nodes that are densely connected with each other more frequently than with those outside the group.

Bhat and Abulaish [12] introduced a community detection approach to viral marketing. Their work concentrated on identifying overlapping communities in online social networks by examining the weighted online social network of email communications from Enron [57]. Two measures, node betweenness and probability, were used to measure the overlapping influence for nodes. If the probability lay between 1% and 5%, then it became a top influencer node. This supported the hypothesis proposed by Bhat and Abulaish [12]: the more a community overlaps, the greater the individual's influence in the whole network. In addition, the outlier nodes were considered noisy nodes, which proved to be non-influencing nodes when the probability reached 60%. Bhat and Abulaish [12] attributed this to the lack of applying the concept of community detection and its attributes to viral marketing. Meng, Zhang, Zhu, Xing, Wang, and Shi [58] have proposed an incremental density-based link clustering algorithm for community detection in dynamic networks called iDBLINK. This algorithm directed to solve the problem in dynamic social network. It ansured its accuracy and efficiency.

4) Evolutionary Methods: Evolutionary or intelligence methods have been widely integrated into several domains because of their effective and optimized consequences. Applying intelligence methods to online social network analysis has had a large impact in terms of social influence. Swarm intelligence methods have been used to analyze social networks; a genetic algorithm (GA), differential evolution (DE), and particle swarm optimization (PSO) were applied to maximize the spread [11]. The study of Gui-sheng et al. [11] selected optimal seeds. GA and DE were used to increase the number of active nodes, while PSO focused on recording the best node position and speed within the population. In the chain of evolutionary computation and swarm algorithms, the ant colony optimization (ACO) algorithm as a swarm intelligence method has been applied to the competitive maximization problem [59]. ACO aims to select a set of nodes and links that maximizes influence by increasing the spread within the network. The algorithm starts by selecting random nodes with an initial pheromone value and then creates solutions corresponding to each node. The process continues iteratively until all the ants probabilistically create solutions. The limitation of this algorithm is that it depends on a random generation of node probability and pheromones. An ant algorithm, a type of intelligence method, was used to analyze social networks to stimulate the spread of opinions [60]. Enhancing the marketing strategy was another purpose of this approach. The approach concentrated on three steps: understanding the network structure, mining entities' opinions, and applying the ant-mining algorithm to formulate opinion rules.

IV. Social Media Marketing In Business

A high proportion of studies in the field of social marketing and especially viral marketing are concerned with business-to-consumer (B2C) business. Constantinides [61] claimed that the development of information technology and communication has kept the customers in control of marketing communication because they are involved in the product development lifecycle. As recent marketing research has focused on the effects and potential of marketing through social networks, Constantinides [61] introduced a classification model for social media as a marketing tool. The model consists of passive and active approaches, where the former depends on the customers' public voice and the latter depends on direct communication with the customers.

Social media platforms have become an effective way to reach customers from different levels, but they need enhancements in targeting suitable individual customers and businesses. Jussila, Kärkkäinen, and Aramo-Immonen [62] mentioned the misconception that social media marketing is directed to the B2C market. They found limited social media use among customers and partners in B2B companies; therefore, their article focused on B2B marketing. They referred to two surveys directed to chief executive officer (CIO) and strategic marketers to measure the adoption of social media in B2B, and they found that the respondents had contrasting perspectives regarding social media in B2B. This is because of the unclear perception of the status of social media. The main difference between social media marketing in B2C and B2B is the customer type. In B2C, customers are ordinary people seeking to know other customer reviews and to have special treatment. In B2B, on the other hand, organizations and companies are looking for direct contact and special offers.

Recent research has shown a lack of models and frameworks for B2B social media marketing [63]–[65]. In particular, [64] listed the most influential factors affecting social media marketing for B2B businesses. Although marketing executives are apprehensive about adopting social media marketing for B2B businesses [63], computerized or automated systems will have a significant impact on the overall process [65].

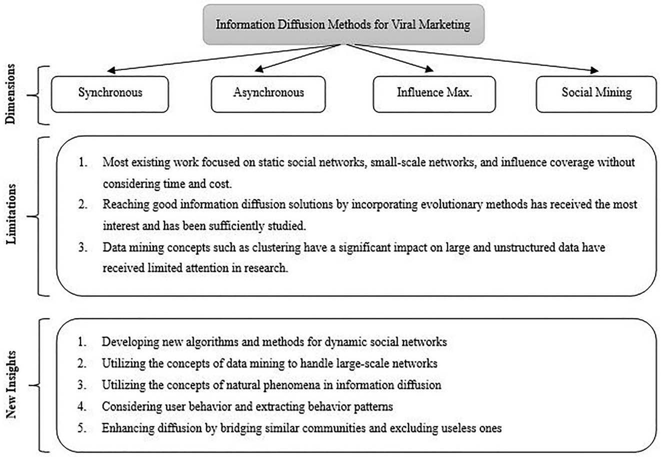
V. Viral Marketing Information Diffusion Model

The previous sections presented a general background of two fundamental concepts related to information diffusion: social networks and viral marketing. This article addresses the state of social media marketing in business and sheds light on recent work on diffusion in online social networks for viral marketing. These works are classified according to the method used: synchronous, asynchronous, influence maximization, and social network mining. There is a huge gap in developing viral marketing methods directed to business. This article also addresses the limitations of some information diffusion models for viral marketing. This section will focus on the challenges and problems related to diffusion in online social networks for viral marketing; these will serve as input to build the VMID model. The taxonomy shown in Fig. 3 presents the four main research dimensions of information diffusion in viral marketing, the related limitations, and areas for improvement.

A. Challenges In Information Diffusion For Viral Marketing

There are some general challenges related to online social networks that affect the process of diffusion for viral marketing. Noisy, and incomplete data are the most common challenges in the online social network structure [66], [67]. Gatti et al. [68] listed major challenges to diffusion in online social networks, such as collecting real-world samples, recognizing user behavior patterns, and large-scale simulation. Most of the discussed methods and algorithms were applied to small-scale networks (see Table II); therefore, large-scale networks still pose a challenge and can lead to new research trends in the coming years. Kleinberg [69] outlined a set of challenges related to large-scale social networks, including the inference of social processes from data and the problem of maintaining individual privacy in studies of social networks.

Table II shows that the majority of methods and algorithms are applied to static social networks. Static social networks are easier to use in experiments and testing than dynamic social networks. Dynamic, continuous, and time-variance testing are needed for more iterations of simulation. Guille et al. [15] discussed technical and crawling API limitations. There are also some challenges related to social network mining. Aggarwal [70] stated that the challenge was mining the linkage behavior of the social network. Jones and Liu [71] presented a mixture of challenges in mining social networks, such as sentiment analysis, trust prediction, privacy and vulnerability, and user migration.



1. Most existing work focused on static social networks, small-scale networks, and influence coverage without considering time and cost.
2. Reaching good information diffusion solutions by incorporating evolutionary methods has received the most interest and has been sufficiently studied.
3. Data mining concepts such as clustering have a significant impact on large and unstructured data have received limited attention in research.

1. Developing new algorithms and methods for dynamic social networks
2. Utilizing the concepts of data mining to handle large-scale networks
3. Utilizing the concepts of natural phenomena in information diffusion
4. Considering user behavior and extracting behavior patterns
5. Enhancing diffusion by bridging similar communities and excluding useless ones

Fig. 3: Taxonomy for information diffusion in viral marketing, related limitations, and new insights.

The problems have also been extracted from the observations gathered in Table II. The most notable point is the dominance of the LT and IC models in the majority of the reviewed work on information diffusion. An attempt has been made to enhance the LT and IC models to maintain their fit into the time-variant application, but only to a limited extent. It is a preliminary attempt that needs further improvement. Influence maximization has been studied in a variety of directions. A few suggestions for its enhancement could be made, such as looking for optimality or fast influence identification selection. Influence maximization was applied in static, dynamic, large-scale, and small-scale networks and achieved optimal or at least good results in experiments and simulations. Diffusion has not been addressed in most of the existing studies that use mining methods. Much of the research on social network mining focused on static social networks. Community detection methods have attracted widespread interest in covering overlapping and hierarchical communities. Clustering was typically used in seeding and identifying the greatest number of influencer nodes; however, a current major focus is on basic clustering methods such as k-means and cliques. Classification, on the other hand, has only used in mining social networks for diffusion as a method for organizing user behavior or properties before the diffusion process occurs. Although data mining is considered a powerful method, there remains a need to test and apply the rest of the mining methods. For decades, data mining methods have effectively analyzed noisy, incomplete, and unstructured data, especially in large data collection, which is the main property of social networks. Time series, association rule mining, and prediction are among the data mining methods that can produce a good analysis.

TABLE II: Summary of Existing Diffusion Models for Viral Marketing

Direction	Domain	Reference	Method/Algorithm Name	Network Type		Network Structure		Outcomes		
				Large-Scale	Small-Scale	Static	Dynamic	Optimal	Good	Weak
Synchronous	LT	[24]	Information Propagation Model		Yes	Yes			Yes	
		[25]	LDAG	Yes		Yes		Yes		
		[26]	Edge Deletion, and Edge Adding	Yes		Yes		Yes		
		[27]	CAND		Yes		Yes		Yes	
	IC	[30]	SpreadRank		Yes		Yes		Yes	
		[31]	ICPN		Yes	Yes			Yes	
Asynchronous		[32], [34], [35]	AsLT, AsIC		Yes		Yes		Yes	
		[36]	T-BaSIC		Yes		Yes	Yes		
		[37]	CT-IC	Yes			Yes		Yes	
		[30]	CTMC-ICM		Yes		Yes		Yes	
Influence Maximization		[7]	Approximation Algorithm		Yes	Yes		Yes		
		[14]	The Degree-Based Scaling Method		Yes	Yes		Yes		
		[38]	Expert and Influential Leader Discovery Approach		Yes	Yes			Yes	
		[41]	MaxG	Yes			Yes	Yes		
		[40]	Hybrid Method		Yes		Yes		Yes	
Social Network Mining	Clustering	[44]	Fuzzy Logic, Game Theory, and Clustering		Yes		Yes	Yes		
		[47]	K-Means Clustering		Yes		Yes		Yes	
		[45]	Cluster Mining, Influence Mining		Yes	Yes			Yes	
		[46]	Spreading Framework	Yes			Yes		Yes	
	Classification	[51]	Classification and Regression Tree (C&RT)		Yes	Yes				Yes
		[52]	Classification TP2010		Yes	Yes			Yes	
	Community Detection	[12]	Overlapping Social Network Communities and Viral Marketing		Yes	Yes			Yes	
		[58]	iDBLINK	Yes			Yes	Yes		
	Evolutionary Algorithm	[60]	Ant Algorithm		Yes	Yes		Yes		
		[11]	GA, DE, and PSO	Yes		Yes		Yes		
		[59]	ACO		Yes	Yes			Yes	

Each has its own aims and methods that will fit into the different steps of diffusion for viral marketing and will ensure the optimality of the overall process.

Evolutionary methods have been more than adequately studied in relation to the diffusion optimization problem. Swarm intelligence, genetic, and heuristic-based algorithms have ensured valuable consequences for the diffusion process. Integrating them with other methods leads to good results in terms of time. Most of the evolutionary research has dealt with huge data collection and has found that evolutionary methods are the best choice.

B. The VMID Model Structure

Addressing all the aforementioned issues and challenges requires an information systems model to maintain the effectiveness of viral marketing diffusion. Findings from the literature indicate the benefit of social media marketing in businesses. Viral marketing as a method based on social media marketing will have better outcomes compared with social media marketing because of its unique procedure. The purpose of the VMID model (Fig. 4) is to address the problem of information diffusion for viral marketing directed to industrial B2B businesses. This model attempts to fill the gap between businesses and social media marketing by adopting viral marketing. It consists of three main parts: business producer, marketing message diffusion, and consumer. The majority of core processing within this model resides in marketing message diffusion because it consists of computational processing that aims to increase the effectiveness of marketing. Normally, the message starts from the business producer, who initiates the marketing campaign. Marketing executives are responsible for formulating the message content and ensuring its structure [70], [71]. The model requires a suitable keyword describing the campaign to match it with the appropriate cluster in the next step. The following subsections will describe each part.

1) Social Analysis: A pre-processing step is required to maintain the targeted social media. This stage consists of two correlated steps. The first step is collecting the dataset from the selected social media. It is important to note the problem related to real-time processing [15], [72]; social media owners are still placing restrictions on collecting real-time datasets. The second step is analyzing the collected dataset. This step requires the classification process to match the suitable cluster for further processing in the next step. The matching process consists of knowing the marketing message and the objective of diffusion. After these steps have been taken and the determined cluster is ready, it will be sent to incremental clustering, which is included in structure modeling in the next module.

2) Structure Modeling: Utilizing incremental clustering for structure modeling deals with the dynamicity of the social network. As discussed in Section V, most problems related to information diffusion are relevant to dynamic and large-scale networks. Introducing incremental clustering solves these two issues because data mining concepts are effective in large-scale data and adopting incremental clustering will handle the dynamicity problem.

The processing in this step receives the collected and analyzed data from social analysis. Then, incremental clustering is performed over it in real time. Incremental clustering has proven its value in terms of cost, space, and time [73]–[75]. The basic idea behind incremental clustering is processing one sample at a time. Takaffoli, et al. [76] introduced a framework for incremental local communities in dynamic networks. However, their work only considered updates from historical sampling, which is ineffective and time consuming in processing. The added value of the VMID model is combining the activity network with incremental clustering; therefore, every newly updated network is passed to activity network processing. In particular, the activity network can be considered as analyzing

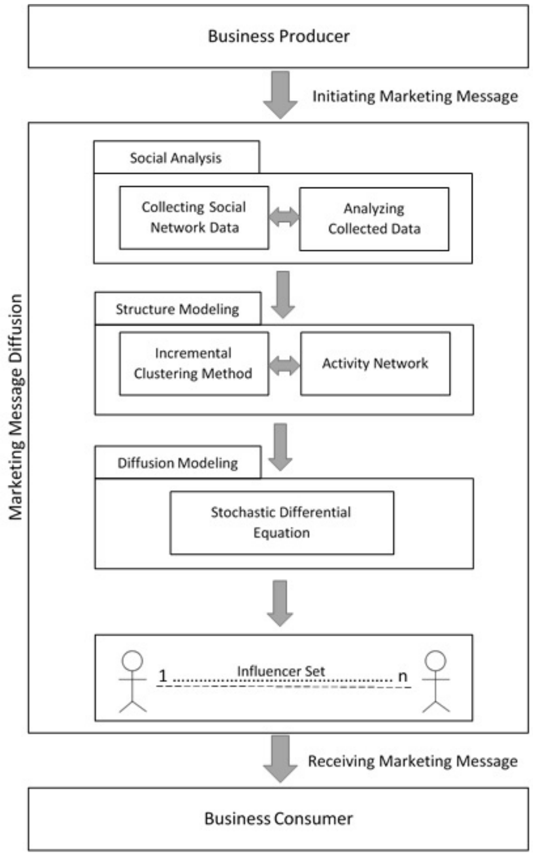

Business Producer
Initiating Marketing Message
Marketing Message Diffusion
Social Analysis
Collecting Social Network Data
Analyzing Collected Data
Structure Modeling
Incremental Clustering Method
Activity Network
Diffusion Modeling
Stochastic Differential Equation
1 Influencer Set n
Receiving Marketing Message
Business Consumer

Fig. 4: VMID model.

user interaction in social networks and microblogging websites. Chun, et al. [77] examined the interaction between users in a guestbook by tracking users' comments. They proposed an activity network where users are the nodes and comments represent the links. The link is constructed only if two users exchange comments. Similarly, Wilson, et al. [78] studied user interaction by proposing an interaction graph that models the users' interaction and communication instead of focusing only on social link relationships. The proposed interaction graph consists of all the nodes existing in the examined social graph. However, the key point of the interaction graph is the link formation, which only contains the links between nodes that interact through communication or an application. The interaction graph was examined using data derived from the Facebook network. The aim of producing the activity network is to optimize the spreading process by excluding the least important users who are inactive in the social network. This reduces the time and cost required to spread a certain message. The resulting network will be saved in the system storage for a matching process between message keywords and cluster keywords.

3) Diffusion Modeling: The concept of diffusion was inspired by natural phenomena such as diseases, waves, fluid, and water. For information diffusion models, this article proposes the adoption of the statistical concept "stochastic differential equation" (SDE), which has been systematically examined for its effectiveness in different domains including engineering, applied mathematics, and computers [79]. SDE was chosen because the platform for marketing diffusion evolves over time, and SDE handles problems with similar objectives. The general mathematical definition in (2) is as follows:

$$dX_t = b\,(t;X_t)\,dt + \ \sigma\,(t;X_t)\,dB_t \qquad (2)$$

where B_t is the standard Brownian motion and b and σ are given functions of time t and the current state x.

A stochastic real-valued process $(X_t)_{t0}$ is said to be a diffusion process if it satisfies the following conditions:

1) $(X_t)_{t\geq 0}$ is a Markov process.
2) The following limits exist:
   a)

$$b\,(t;X_t) = \underset{\Delta\to 0}{lim}\,\frac{1}{\Delta} E(X(\ t+\Delta) - X(t))|\ X(t) = x) \qquad (3)$$

   b)

$$\sigma^2\,(t;X_t) = \underset{\Delta\to 0}{lim}\,\frac{1}{\Delta}E \{(X\,(t+\Delta) - X\,(t))^2|\ X(t) = x)) \qquad (4)$$

3) $X(t)_{t\geq 0}$ is a continuous process

$$(P\,(|X_t - X_s| \geq \ \epsilon \mid X_s = x) = o\ \ (t-s)) \qquad (5)$$

$b\,(t;X_t)$ is called the drift (coefficient, parameter) and $\sigma^2\,(t;X_t)$ is the diffusion (coefficient, parameter). The definition of the diffusion process suggests a relationship between drift and diffusion, as shown in (2).

VI. Conclusion

Research in the field of information diffusion for viral marketing has increased enormously in recent years. This article, to the best of our knowledge, is the first to examine studies related to information diffusion in viral marketing. This article discusses some of the challenges and issues that can be used to direct future research. It also addresses the gap in social media marketing by introducing the VMID model. The proposed model has based on incremental clustering and SDE. It attempts to adopt information system modeling for marketing in B2B business. In industrial markets, it is a challenge to align with a systematic method that facilitates the connection between partners. The VMID model requires more experiments and testing to ensure its validity in the real world.

References

[1] S. Bennett, "Just How Big Is Twitter In 2012?," ed, 2012.

[2] S. Harden, "Statistic Brain Percentage, Numbers, Financial, Ranking," in *Twitter Statistics*, ed, 2014.

[3] S. Bennett, "All Twitter," in *Revolutionized Marketing: The Rise Of Social Media*, ed, 2014.

[4] D. Brown and S. Fiorella, Influence Marketing How to Create, Manage, and Measure Brand Influencers in *Social Media Marketing*. Indiana: Que, 2013.

[5] M. Cohn, "Social Media vs Social Networking," in *compukol connection*, ed, 2011.

[6] D. Easley and J. Kleinberg, *Networks,Crowds, and Markets: Reasoning about a Highly Connected World.*: Cambridge University Press 2010.

[7] D. Kempe, J. Kleinberg, and E. Tardos, ”Maximizing the Spread of Influence through a Social Network,” in *KDD ’03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, 2003, pp. 137-146.

[8] D. Centola, ”The spread of behavior in an online social network,” *Science*, vol. 329, pp. 1194–1197, 2010.

[9] C. Long and R. C.-W. Wong, ”Minimizing Seed Set for Viral Marketing,” in *11th IEEE International Conference on Data Mining*, Vancouver,BC, 2011, pp. 427 - 436.

[10] C. Jiang, Y. Chen, and K. J. Ray Liu. (2013, December 2). Evolutionary Dynamics of Information Diffusion over Social Networks. *Cornell University Library*. Available: http://arxiv.org/abs/1312.0317

[11] Y. Gui-sheng, W. Ji-jie, D. Hong-bin, and L. Jia, ”Intelligent Viral Marketing algorithm over online social network,” in *Second International Conference on Networking and Distributed Computing*, Beijing, 2011, pp. 319-323.

[12] S. Y. Bhat and M. Abulaish, ”Overlapping Social Network Communities and Viral Marketing,” in *In proceeding of: International Symposium on Computational and Business Intelligence*, New Delhi, 2013.

[13] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, ”Identification of influential spreaders in complex networks,” *Nature Physics*, vol. 6, pp. 888–893, 2010.

[14] A. Chaudhury, P. Basuchowdhuri, and S. Majumdar, ”Spread of Information in a Social Network Using Influential Nodes,” in *Advances in Knowledge Discovery and Data Mining*, ed Berlin: Springer-Verlag, 2012, pp. 121-132.

[15] A. Guille, H. Hacid, C. Favre, and D. Zighed, ”Information Diffusion in Online Social Networks: A Survey,” *SIGMOD Rec.*, pp. 17-28, 2013.

[16] V. Arnaboldi, M. Conti, M. L. Gala, A. Passarella, and F. Pezzoni, ”Information diffusion in OSNs: the impact of nodes’ sociality,” in *SAC’14, Gyeongju*, Republic of Korea, 2014, pp. 616-621.

[17] M. Newman, ”The Structure and Function of Complex Networks,” *SIAM Review*, vol. 45, pp. 167-256, 2003.

[18] M. Opuszko and J. Ruhland, ”Impact of the Network Structure on the SIR Model Spreading Phenomena in Online Networks,” in *ICCGI 2013*, Nice, France, 2013, pp. 22-28.

[19] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, ”Cascading behavior in large blog graphs,” in *SDM ’07*, Patterns of Cascading Behavior in Large Blog Graphs, 2007, pp. 551-556.

[20] S. Boccalettia, V. Latorab, Y. Morenod, M. Chavez, and D.-U. Hwang, ”Complex networks: Structure and dynamics,” *Physics Reports*, vol. 424, pp. 175-308, 2006.

[21] Z. Lu, Y. Wen, and G. Cao, ”Information Diffusion in Mobile Social Networks: The Speed Perspective,” in *IEEE INFOCOM ’14*, Toronto, ON, 2014, pp. 1932-1940.

[22] M. Granovetter, ”Threshold Models of Collective Behavior,” *Amarican Journal of Sociology*, pp. 1420-1443, 1978.

[23] J. Goldenberg, B. Libai, and E. Muller, ”Using Complex Systems Analysis to Advanced Marketing Theory Development,” *Academy of Marketing Science Review*, 2001.

[24] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, ”Outtweeting the Twitterers - Predicting Information Cascades in Microblogs,” in *WOSN’10 Proceedings of the 3rd conference on Online social networks*, Boston, MA, 2010, pp. 3-11.

[25] W. Chen, Y. Yuan, and L. Zhang, ”Scalable Influence Maximization in Social Networks under the Linear Threshold Model,” in *IEEE 10th International Conference on Data Mining (ICDM)*, Sydney, NSW, 2010, pp. 88-97.

[26] E. B. Khalil, B. Dilkina, and L. Song, ”Scalable diffusion-aware optimization of network topology,” in *KDD ’14*, New York, New York, USA, 2014, pp. 1226-1235.

[27] Y. Guisheng, W. Jijie, and D. Hongbin, ”A Cellular Automaton based Network Diffusion model: Preparation for more scalable Viral Marketing,” in *2012 International Conference on Collaboration Technologies and Systems (CTS)*, Denver, CO, USA, 2012, pp. 308-315.

[28] M. Kimura, K. Saito, and H. Motoda, ”Blocking Links to Minimize Contamination Spread in a Social Network,” *ACM Trans. Knowl. Discov. Data*, pp. 1-22, 2009.

[29] T. C. Schelling, *Micromotives and Macrobehavior*. New York: W. W. Norton and Company 2006.

[30] T. Zhu, B. Wang, B. Wu, and C. Zhu, ”Maximizing the spread of influence ranking in social networks,” *Information Sciences*, vol. 278, pp. 535-544, 2014.

[31] A. Nazemian and F. Taghiyareh, ”Influence maximization in Independent Cascade model with positive and negative word of mouth,” in *2012 Sixth International Symposium on Telecommunications (IST)*, 2012, pp. 854-860.

[32] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda, ”Learning Diffusion Probability based on Node Attributes in Social Networks,” in *Proceedings of the 19th international conference on Foundations of intelligent systems*, Warsaw, Poland, 2011, pp. 153-162.

[33] K. Saito, M. Kimura, K. Ohara, and H. Motoda, ”Learning Continuous-Time Information Diffusion Model for Social Behavioral Data Analysis,” in *Advances in Machine Learning*, ed Berlin Heidelberg: Springer, 2009, pp. 322-337.

[34] M. Kimura, K. Saito, K. Ohara, and H. Motoda, ”Learning to Predict Opinion Share in Social Networks,” in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, Atlanta, Georgia., 2010, pp. 1364–1370.

[35] M. Kimura, K. Saito, K. Ohara, and H. Motoda, ”Learning to predict opinion share and detect anti-majority opinionists in social networks,” *Journal of Intelligent Information Systems*, p. Springer Science+Business Media, 2013.

[36] A. Guille and H. Hacid ”A predictive model for the temporal dynamics of information diffusion in online social networks,” in *WWW ’12 Companion*, Lyon, 2012, pp. 1145-1152.

[37] L. Wonyeol, K. Jinha, and Y. Hwanjo, ”CT-IC: Continuously Activated and Time-Restricted Independent Cascade Model for Viral Marketing,” in *2012 IEEE 12th International Conference on Data Mining (ICDM)*, 2012, pp. 960-965.

[38] N. S. N. Abadi and M. R. Khayyambashi, ”Influence maximization in viral marketing with expert and influential leader discovery approach,” in *2014 8th International Conference on e-Commerce in Developing Countries: With Focus on e-Trust (ECDC)*, 2014, pp. 1-8.

[39] J. Zhou, Y. Zhang, and J. Cheng, ”Preference-based mining of top- influential nodes in social networks,,” *Future Generation Computer Systems*, vol. 31, pp. 40-47, 2014.

[40] H. Kim, K. Beznosov , and E. Yoneki, ”Finding influential neighbors to maximize information diffusion in twitter,” in *WWW Companion ’14*, Seoul, Korea, 2014, pp. 701-706.

[41] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, ”Influence Maximization in Dynamic Social Networks,” *in IEEE International Conference on Data Mining (ICDM)*, Dallas, 2013.

[42] J. Han, M. Kamber, and J. Pei,*Data Mining: Concepts and Techniques*. Walthman, USA: Morgan Kaufmann Publishers, 2006.

[43] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. UK: John Wiley & Sons Ltd., 2009.

[44] S. Banerjee, H. Al-Qaheri, and A. E. Hassanien, ”Mining Social networks for viral marketing using fuzzy logic,” in *Fourth Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation (AMS)*, Kota Kinabalu, Malaysia, 2010, pp. 24-28.

[45] S. Sharma and V. Shrivastava, ”Viral Marketing in Social Network Using Data Mining,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 1, pp. 380-384, 2013.

[46] L. AlSuwaidan, M. Ykhlef, and M. A. Alnuem, ”A novel spreading framework using incremental clustering for viral marketing,” in *11th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA’ 2014)*, Doha, 2014.

[47] G. Niu, Y. Long, and V. O. K. Li, ”Temporal Behavior of Social Network Users in Information Diffusion,” in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Warsaw, 2014, pp. 150 - 157.

[48] T. Lou and J. Tang, ”Mining structural hole spanners through information diffusion in social network,” in *IW3C2*, Rio de Janeiro, Brazil, 2013, pp. 837-847.

[49] H. A. Shrihari, A. A. Hudli, and A. V. Hudli, ”Identifying Online Opinion Leaders Using K-means Clustering,” in *12th International Conference on Intelligent Systems Design and Applications (ISDA)*, Kochi, 2012, pp. 416-419.

[50] K. Tsiptsis and A. Chorianopoulos, *Data mining techniques in CRM*. UK: John Wiley and Son Ltd., 2009.

[51] J. Surma and A. Furmanek, ”Data Mining in On-line Socail Networks for Marketing Response Analysis,” in *IEEE International Conference on Privacy, Security, Risk, and IEEE International Conference on Social Computing*, Boston, MA, 2011, pp. 537-540.

[52] A. Ortigosa, R. M. Carro, and J. I. Quiroga, ”Predicting user personality by mining social interactions in Facebook,” *Journal of Computer and System Sciences*, Volume 80, Issue 1, February 2014, vol. 80, pp. 57-71, 2014.

[53] P. Zhang, C. Moore, and M. E. J. Newman, ”Community detection in networks with unequal groups,” *Physical review E*, vol. 93, p. 012303, 2016.

[54] D. Kempe, *Structure and Dynamics of Information in Networks*. Los Angeles: University of South California, 2011.

[55] H.-W. Shen, *Community Structure of Complex Networks*. Springer Berlin Heidelberg, 2013.

[56] L. Tang and H. Liu, *Community Detection and Mining in Social Media*. Morgan & Claypool Publishers, 2010.

[57] B. Klimt and Y. Yang, ”The Enron corpus: A new dataset for email classification research,” in *Proceedings of 15th European Conference on Machine Learning*, Pisa, Italy, 2004, pp. 217-226.

[58] F. Meng, F. Zhang, M. Zhu, Y. Xing, Z. Wang, and J. Shi, ”Incremental Density-Based Link Clustering Algorithm for Community Detection in Dynamic Networks,” *Mathematical Problems in Engineering*, vol. 2016, p. 11, 2016.

[59] W.-S. Yang and S.-X. Weng, ”Application of the ant colony optimization algorithm to competitive viral marketing,” in *Proceedings of the 7th Hellenic conference on Artificial Intelligence: theories and applications*, Lamia, Greece, 2012.

[60] C. Kaiser, J. Kröckel, and F. Bodendor, ”Ant-Based Simulation of Opinion Spreading in Online Social Networks,” in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, ON, 2010, pp. 537 - 540.

[61] E. Constantinides, ”Foundations of Social Media Marketing,” *Procedia - Social and Behavioral Sciences*, vol. 148, pp. 40-57, 2014.

[62] J. J. Jussila, H. Kärkkäinen, and H. Aramo-Immonen, ”Social media utilization in business-to-business relationships of technology industry firms,” *Computers in Human Behavior*, vol. 30, pp. 606-613, 2014.

[63] S. Jari, L. Tuula, S. Henri, and M. Matti, ”Social Media Marketing in the Scandinavian Industrial Markets,” in *Marketing and Consumer Behavior: Concepts, Methodologies, Tools, and Applications*, ed Hershey, PA, USA: IGI Global, 2015, pp. 1136-1152.

[64] M. I. Dahnil, K. M. Marzuki, J. Langgat, and N. F. Fabeil, ”Factors Influencing SMEs Adoption of Social Media Marketing,” *Procedia - Social and Behavioral Sciences*, vol. 148, pp. 119-126, 2014.

[65] H. Kim, ”The role of WOM and dynamic capability in B2B transactions,” *Journal of Research in Interactive Marketing*, vol. 8, pp. 84-101, 2014.

[66] Y. Liang, J. Caverlee, Z. Cheng, and K. Y. Kamath, ”How Big is the Crowd? Event and Location Based Population Modeling in Social Media,” in *24th ACM Conference on Hypertext and Social Media*, Paris, France, 2013.

[67] G. S. Bindra, K. K. Kandwal, P. K. Singh, and S. Khanna, ”Tracing Information Flow and Analyzing the Effects of Incomplete Data in Social Media,” in *2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN)*, Phuket, 2012.

[68] M. Gatti, A. P. Appel, C. Pinhanez, C. d. Santos, D. Gribel, P. Cavalin, and S. B. Neto, ”Large-Scale Multi-agent-Based Modeling and Simulation of Microblogging-Based Online Social Network,” in *Multi-Agent-Based Simulation XIV*, ed: Springer Berlin Heidelberg, 2014, pp. 17-33.

[69] J. Kleinberg, ”Challenges in Mining Social Network Data: Processes,Privacy, and Paradoxes,” in *KDD ’07*, San Jose, CA, USA, 2007, pp. 4-5.

[70] C. C. Aggarwal, ”An introduction to social network data analytics,” in *Social Network Data Analytics*, ed: Springer US, 2011, pp. 1-15.

[71] I. Jones and H. Liu, ”Mining Social Media: Challenges and Opportunities,” in *2013 International Conference on Social Intelligence and Technology*, State College, PA, 2013, pp. 90-99.

[72] I. Taxidou and P. Fischer, ”Realtime analysis of information diffusion in social media,” *Proc. VLDB Endow.*, vol. 6, pp. 1416-1421, 2013.

[73] F. Can, ”Incremental clustering for dynamic information processing,” *ACM Trans. Inf. Syst.*, vol. 11, pp. 143-164, 1993.

[74] M. Hai-Dong, S. Yu-Chen, S. Fei-Yan, and W. Shu-Ling, ”Clustering for Complex and Massive Data,” in *Information Engineering and Computer Science*, 2009. ICIECS 2009. International Conference on, 2009, pp. 1-4.

[75] S. Young, I. Arel, T. P. Karnowski, and D. Rose, ”A Fast and Stable Incremental Clustering Algorithm,” in *2010 Seventh International Conference on Information Technology: New Generations (ITNG)*, 2010, pp. 204-209.

[76] M. Takaffoli, R. Rabbany, and O. R. Zaiane, ”Incremental local community identification in dynamic social networks,” in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagara, Ontario, Canada, 2013.

[77] H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong, ”Comparison of online social relations in volume vs interaction: a case study of cyworld,” in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, Vouliagmeni, Greece, 2008.

[78] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, ”User interactions in social networks and their implications,” in *Proceedings of the 4th ACM European conference on Computer systems*, Nuremberg, Germany, 2009.

[79] G. Gradoni, R. Pastor, D. Micheli, F. Moglie, V. M. Primiani, and M. Marchetti, ”Stochastic differential equation for wave diffusion in random media,” in *2013 International Conference on Electromagnetics in Advanced Applications (ICEAA)*, 2013, pp. 1176-1177.

A New Threshold Based Penalty Function Embedded MOEA/D

Muhammad Asif Jan
Department of Mathematics
Kohat University of Science & Technology
Khyber Pakhtunkhwa, Pakistan

Nasser Mansoor Tairan
College of Computer Science
King Khalid University Abha,
Saudi Arabia

Rashida Adeeb Khanum
Jinnah College for Women
University of Peshawar Khyber
Pakhtunkhwa, Pakistan

Wali Khan Mashwani
Department of Mathematics
Kohat University of Science & Technology
Khyber Pakhtunkhwa, Pakistan

***Abstract*—Recently, we proposed a new threshold based penalty function. The threshold dynamically controls the penalty to infeasible solutions. This paper implants the two different forms of the proposed penalty function in the multiobjective evolutionary algorithm based on decomposition (MOEA/D) framework to solve constrained multiobjective optimization problems. This led to a new algorithm, denoted by CMOEA/D-DE-ATP. The performance of CMOEA/D-DE-ATP is tested on hard CF-series test instances in terms of the values of IGD-metric and SC-metric. The experimental results are compared with the three best performers of CEC 2009 MOEA competition. Experimental results show that the proposed penalty function is very promising, and it works well in the MOEA/D framework.**

***Keywords*—Constrained multiobjective optimization; decomposition; MOEA/D; penalty function; threshold.**

I. Introduction

In this paper, we consider the following constrained multiobjective optimization problem (CMOP) [1]:

$$\begin{array}{ll} \text{Minimize} & F(x) = (f_1(x), f_2(x), \ldots, f_m(x))^T; \\ \text{Subject to} & g_j(x) \geq 0,\ j = 1, \ldots, p; \\ & l_k \leq x_k \leq u_k,\ k = 1, \ldots, n, \end{array} \qquad (1)$$

where $x = (x_1, \ldots, x_n)^T \in \mathcal{R}^n$ is an n dimensional vector of decision variables, F is the objective vector function that consists of m real-valued objective functions, and $g_i(x) \geq 0$ are inequality constraints. The objective and constraint functions, f_i's and g_j's, could be linear or non linear real-valued functions. l_k and u_k are the lower and upper bounds (called bound constraints) of x_k, $k = 1, \ldots, n$, respectively, which define the search region $\mathcal{S} = \{x = (x_1, \ldots, x_n)^T |\ l_k \leq x_k \leq u_k, k = 1, \ldots, n\}$.

A solution $x \in S$ is called a feasible solution, if it satisfies all the inequality constraints in (1). The set of all feasible solutions is called the feasible region. Mathematically, we can write:

$$\mathcal{F} = \{x \in \mathcal{S} \subset \mathcal{R}^n | g_j(x) \geq 0, j = 1, \cdots, p\}.$$

However, If a solution is not feasible, we call it infeasible. The set of all infeasible solutions is called the infeasible region. The feasible attainable objective set (AOS) can be defined as $\{F(x) | x \in \mathcal{F}\}$.

Since the objectives in (1) more often contradict each other, so it is hard to find a single solution in $\mathcal{F}$ that could minimize all the objectives at the same time. Instead, one looks for a set of optimal compromising/tradeoff feasible solutions. The best tradeoffs among the objectives can be defined in terms of Pareto-optimality [2], [3].

A solution x is said to Pareto-dominate or simply dominate another solution y, mathematically denoted as $x \preceq y$, if $f_i(x) \leq f_i(y)$, $\forall i = 1, \ldots, m$ and $f_j(x) < f_j(y)$ for at least one $j \in \{1, \ldots, m\}$[1]. This definition of domination is sometimes referred to as a weak dominance relation.
A solution $x^* \in \mathcal{F}$ is Pareto-optimal to (1) if there is no solution $x \in \mathcal{F}$ such that $F(x) \preceq F(x^*)$. $F(x^*)$ is then called a Pareto-optimal (objective) vector. The set of all Pareto-optimal solutions is called the Pareto Set (PS) in the decision space and Pareto Front (PF) in the objective space [2].

In the majority of constrained optimization problems, the optimal solutions lie on the constraints' boundaries. Thus, to arrive at these solutions, some algorithms evolve some good infeasible solutions with less constraint violation along with their feasible counterparts during the evolutionary process (e.g., see [4]–[6]). The primary purpose of evolving infeasible solutions in the search procedure is to utilize the information they transport. As EAs are stochastic search and optimization methods, rejecting infeasible individuals might lead the EA being stuck in local optima, particularly in problems with disconnected search space [7], [8]. Moreover, in some highly constrained optimization problems, it could be a demanding problem to find a single feasible solution [9], [10]. Therefore, constraint handling techniques used in multiobjective optimization (MOO) can be mainly distinguished by knowing how infeasible solutions are mixed up and evolved in the evolutionary process.

[1] This definition of domination is for minimization. All the inequalities should be reversed if the goal is to maximize the objectives in (1). "dominate" means "be better than".

In [1], we introduced a new threshold based penalty function in the replacement and update scheme of MOEA/D-DE [11], an improved version of MOEA/D [12], to penalize infeasible solutions. The threshold is adaptively adjusted by using the minimum and maximum constraint violation in the neighborhood of a solution. The infeasible solutions with constraint violation less than the threshold are less penalized than the ones with constraint violation greater than the threshold. As a result, we expect that some good infeasible solutions with less constraint violation will have a chance to evolve in the evolutionary process. The some preliminary experimental results, presented in [1], have proven the capability of the proposed algorithm for solving CF-series [13] test instances. In this paper, we present detailed experimental results and comment on the pitfalls of the proposed algorithm.

The rest of this paper is organized as follows. Section II presents some basic concepts and the two versions of the proposed penalty function. Section III briefly introduces MOEA/D and the modified algorithmic framework of MOEA/D-DE. Section IV discusses the experimental settings. Section V presents and discusses experimental results on CF-series [13] test instances. Section VI compares our experimental results with the three best performers [14]–[16] of CEC 2009 MOEA competition. Finally, Section VII concludes this paper with an outline of the work carried out.

II. Basic Concepts and The Proposed Penalty Function

A. Degree of Constraint Violation

The degree of constraint violation of a solution $x \in S$ can be defined as [1], [3]:

$$V(x) = |\sum_{j=1}^{p} \min(g_j(x), 0)|. \tag{2}$$

Obviously, if $V(x) = 0$, x is feasible; otherwise, it is infeasible.

B. Tchebycheff Aggregation Function

MOEA/D [12] decomposes an MOP into a number of single objective subproblems. This paper uses the Tchebycheff aggregation function for this purpose, which is given as under [17]:

$$\begin{array}{ll} \text{Minimize} & g^{te}(x|\lambda, z^*) = \max_{1\leq i\leq m}\{\lambda_i|f_i(x) - z_i^*|\}; \\ \text{Subject to} & x \in \mathcal{F} \subset \mathcal{R}^n; \end{array} \tag{3}$$

where $z^* = (z_1^*, \ldots, z_m^*)^T$ is the reference point, i.e., $z_i^* = \min\{f_i(x)|x \in \ \mathcal{F}\}\ \forall i = 1, \ldots, m$ and $\lambda = (\lambda_1, \ldots, \lambda_m)^T$ is a weight vector such that $\lambda_i \geq 0\ \forall i = 1, \ldots, m$ and $\sum_{i=1}^m \lambda_i = 1$. Some theorems related to the Pareto-optimality of Tchebycheff aggregation function can be found in [2].

C. The Proposed Penalty Function

The proposed penalty function uses a threshold value, τ for dynamically controlling the amount of penalty.

Suppose MOEA/D [12] decomposes the MOP into N subproblems. At each generation, MOEA/D retains N solutions $x^1, \ldots, x^N$, where x^i is the current solution to subproblem i. Let P be the mating and update range set in MOEA/D. Then define [1]:

$$V_{min} = \min\{V(x^i), i \in P\}, \tag{4}$$

$$V_{max} = \max\{V(x^i), i \in P\}, \tag{5}$$

where $V(x^i)$ is the degree of constraint violation of solution x^i.

The threshold value, τ is defined as [1]:

$$\tau = V_{min} + s(V_{max} - V_{min}), \tag{6}$$

where the parameter s controls the threshold value. In [1], we used $s = 0.3$.

Our suggested penalty function encourages the algorithm to search the feasible region and the infeasible region near the feasible region. It is defined in the following two different ways: For $i = 1, \ldots, m$

$$f_p^i(x) = \begin{cases} f_i(x) + s_1 V^2(x), & \text{if } V(x) < \tau; \\ f_i(x) + s_1\tau^2 + & \\ s_2(V(x) - \tau), & \text{otherwise,} \end{cases} \tag{7}$$

$$g_p^{te}(x|\lambda, z^*) = \begin{cases} g^{te}(x|\lambda, z^*) + s_1 V^2(x), & \text{if } V(x) < \tau; \\ g^{te}(x|\lambda, z^*) + s_1\tau^2 + & \\ s_2(V(x) - \tau), & \text{otherwise,} \end{cases} \tag{8}$$

where s_1 and s_2 are two scaling parameters with $s_1 << s_2$. In the penalty functions, the penalty increases sharply when $V(x)$ exceeds the threshold. This is realized by scaling the degree of constraint violation, $V(x)$ of an infeasible solution by relatively high value of parameter s_2 than parameter s_1 in our penalty function formulations. In Eq. 7, the penalty is added to individual objective function values of an infeasible solution, while in Eq. 8, it is added directly to Tchebycheff aggregation function value of an infeasible solution. Furthermore, in [1], we tested Eq. 8 only.

III. Multiobjective Evolutionary Algorithm Based on Decomposition

Zhang and Li [12] suggested a simple yet efficient MOEA, multiobjective evolutionary algorithm based on decomposition (MOEA/D). MOEA/D approximates the PF by explicitly decomposing an MOP into several single objective optimization subproblems. These subproblems are then optimized concurrently and collaboratively by evolving population of solutions. An EA is employed for this purpose. The Euclidean distances between the aggregation coefficient vectors of these subproblems are calculated to identify the neighborhood of each subproblem. The information gathered from the neighboring subproblems is then used to optimize a subproblem.

In this work, we employed the penalty functions defined by Eqs. 7, 8 in the update scheme of MOEA/D-DE [11], one of the efficient versions of MOEA/D to solve CF-series [13] test instances. This resulted in a new algorithm, denoted by CMOEA/D-DE-ATP (For details of CMOEA/D-DE, please see [18]). The pseudo-code of the modified update scheme is given in Algorithm 1.

Algorithm 1 Pseudo-code of the update scheme of CMOEA/D-DE-ATP.

```
1:  Each new child solution $y$ updates $n_r$ solutions from the set $P$ of its neighboring solutions as follows:
2:  Set $c = 0$ and then do the following:
3:  if $c = n_r$ or $P$ is empty then
4:      return;
5:  else
6:      Randomly pick an index $j$ from $P$;
7:      Compute the Tchebycheff aggregation function values of $y$ and $x^j$ with the new objective values of Eq. 7 (or the new aggregation function values of $y$ and $x^j$ with Eq. 8);
8:      if $g^{te}(y|\lambda^j, z) \leq g^{te}(x^j|\lambda^j, z)$ (or $g_p^{te}(y|\lambda^j, z) \leq g_p^{te}(x^j|\lambda^j, z)$ ) then
9:          $x^j = y$, $F(x^j) = F(y)$, $V(x^j) = V(y)$, and $c = c + 1$;
10:     end if
11:     Remove $j$ from $P$ and go to step 3;
12: end if
```

IV. Experimental settings

In our experiments, we use the same parameters' settings and weight vectors' selection criteria as is used in [13]. Further, we use statistics of the inverted generational distance metric (IGD-metric) [12], [19] for comparing results on CF-series test instances, CF1-CF10. Also, the set coverage metric (SC-metric) [12] is used to compare the nondominated solutions obtained by different algorithms. Unless otherwise stated, we will use Eq. 6 with $s = 0.7$ and Eqs. 7, 8 with $s_1 = 0.01$ and $s_2 = 20$ in all experiments.

V. Experimental Results

TABLE I: THE IGD-METRIC STATISTICS OF CMOEA/D-DE-ATP USING EQS. 7, 8. THE RESULTS IN **BOLDFACE** INDICATE THE BETTER RESULTS; IF NOT, THEY ARE IDENTICAL.

	best (lowest)		mean		st. dev.	
Test Instance	Eq. 7	Eq. 8	Eq. 7	Eq. 8	Eq. 7	Eq. 8
CF1	0.0003	0.0003	0.0006	**0.0005**	0.0003	**0.0002**
CF2	0.0028	**0.0027**	**0.0037**	0.0041	**0.0013**	0.0019
CF3	0.0632	0.0632	0.1382	0.1382	0.0441	0.0441
CF4	0.0060	**0.0051**	0.0097	**0.0095**	**0.0042**	0.0043
CF5	0.0406	**0.0297**	**0.1606**	0.1663	**0.1084**	0.1107
CF6	**0.0049**	0.0053	0.0197	**0.0192**	**0.0141**	0.0144
CF7	0.0344	**0.0304**	**0.1188**	0.1310	0.0729	**0.0722**
CF8	**0.0332**	0.0356	**0.0370**	0.0371	0.0020	**0.0010**
CF9	**0.0428**	0.0434	**0.0468**	0.0479	**0.0022**	0.0030
CF10	**0.1068**	0.1108	**0.1509**	0.1630	**0.0396**	0.0409

Table I presents the best (i.e., lowest), mean, and standard deviation of the IGD-metric values for CF-series test instances found by CMOEA/D-DE-ATP with Eqs. 7, 8. These statistics are based on 30 independent runs. As it can be seen from this table that CMOEA/D-DE-ATP can find better best values with Eq. 7 for one 2-objective, CF6 and three 3-objective, CF8-CF10 and with Eq. 8 for four 2-objective, CF2, CF4, CF5, and CF7 test instances. The best values for test instance CF1 are identical. This table also shows that CMOEA/D-DE-ATP with both Eqs. 7, 8 performs similarly on test instance CF3. However, improved mean and st. dev. values can be found when CMOEA/D-DE-ATP employs Eq. 7 for most of the test instances. In particular, the improved performance can be seen for the three 3-objective test instances, CF8-CF10. This suggests that adding the penalty to individual objective function values as is done in Eq. 7 before calculating the aggregation function values is a good choice for better performance on CF-series test instances.

It can also be seen from Table I that CMOEA/D-DE-ATP with both Eqs. 7, 8 finds small values for the mean of IGD-metric on CF1, CF2, CF4, CF6, CF8, CF9. Empirically, these results illustrate that the final nondominated solutions found by CMOEA/D-DE-ATP for these test instances approximate the PF very well in a sense.

TABLE II: THE AVERAGE SET COVERAGE BETWEEN CMOEA/D-DE-ATP WITH EQ. 7 AND WITH EQ. 8 ON CF-SERIES TEST INSTANCES. THE RESULTS IN **BOLDFACE** INDICATE THE BETTER RESULTS; IF NOT, THEY ARE IDENTICAL.

Test Instance	C(Eq. 7, Eq. 8)	C(Eq. 8, Eq. 7)
CF1	0.46	**0.48**
CF2	0.13	0.13
CF3	0.65	0.65
CF4	0.26	**0.28**
CF5	0.19	**0.22**
CF6	0.17	**0.21**
CF7	0.24	0.24
CF8	**0.05**	0.04
CF9	0.03	0.03
CF10	**0.39**	0.31

Table II presents the average set coverage between the nondominated solutions of CMOEA/D-DE-ATP with Eq. 7 and Eq. 8. The results of this table reveal that, in terms of the SC-metric, the nondominated solutions found by CMOEA/D-DE-ATP with Eq. 8 are better than those obtained with Eq. 7 for test instances CF1, CF4-CF6, but are worse for test instances CF8 and CF10 vice versa. The table also shows that the nondominated solutions acquired from CMOEA/D-DE-ATP with both Eqs. 7, 8 are same for test instances CF2, CF3, CF7, and CF9. However, looking at the results of this table, it can be inferred that the performance of CMOEA/D-DE-ATP is comparable with both Eqs. 7, 8, as there is no big difference in the SC-metric values.

Figures 1 and 2 show, in the objective space, the distributions of the 100 and 150 nondominated population members for the seven 2-objective, CF1-CF7, and the three 3-objective, CF8-CF10, CF-series test instances. These solutions are selected based on the criteria as mentioned in [13] from the final population of the run with the best (i.e., lowest) IGD-metric value among the 30 independent runs. These figures also show

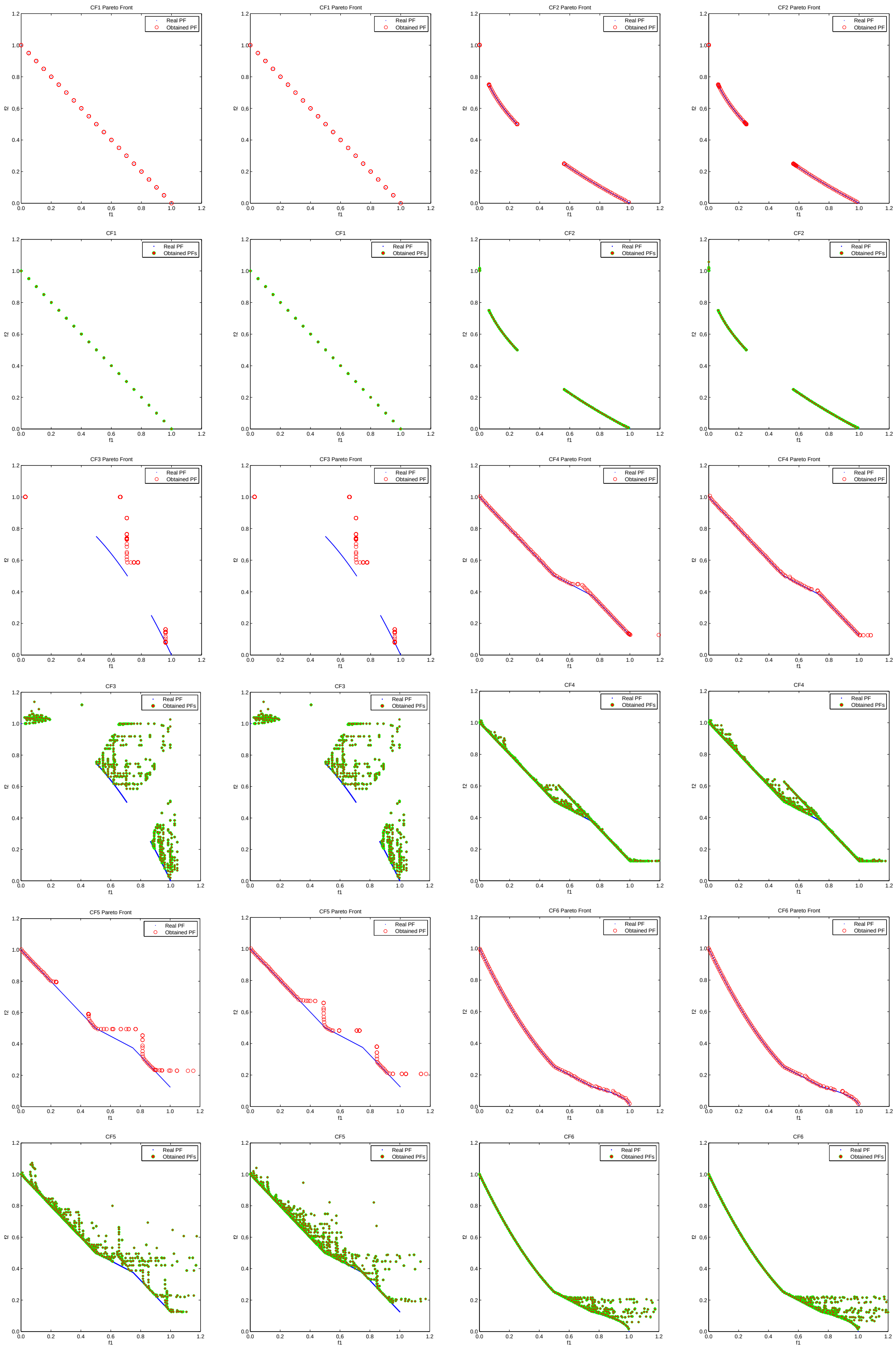


Fig. 1: Plots of the nondominated front with the best IGD value and all the 30 final nondominated fronts found by CMOEA/D-DE-ATP when using Eq. 7 (columns 1 and 3) and Eq. 8 (columns 2 and 4) for CF1-CF6.

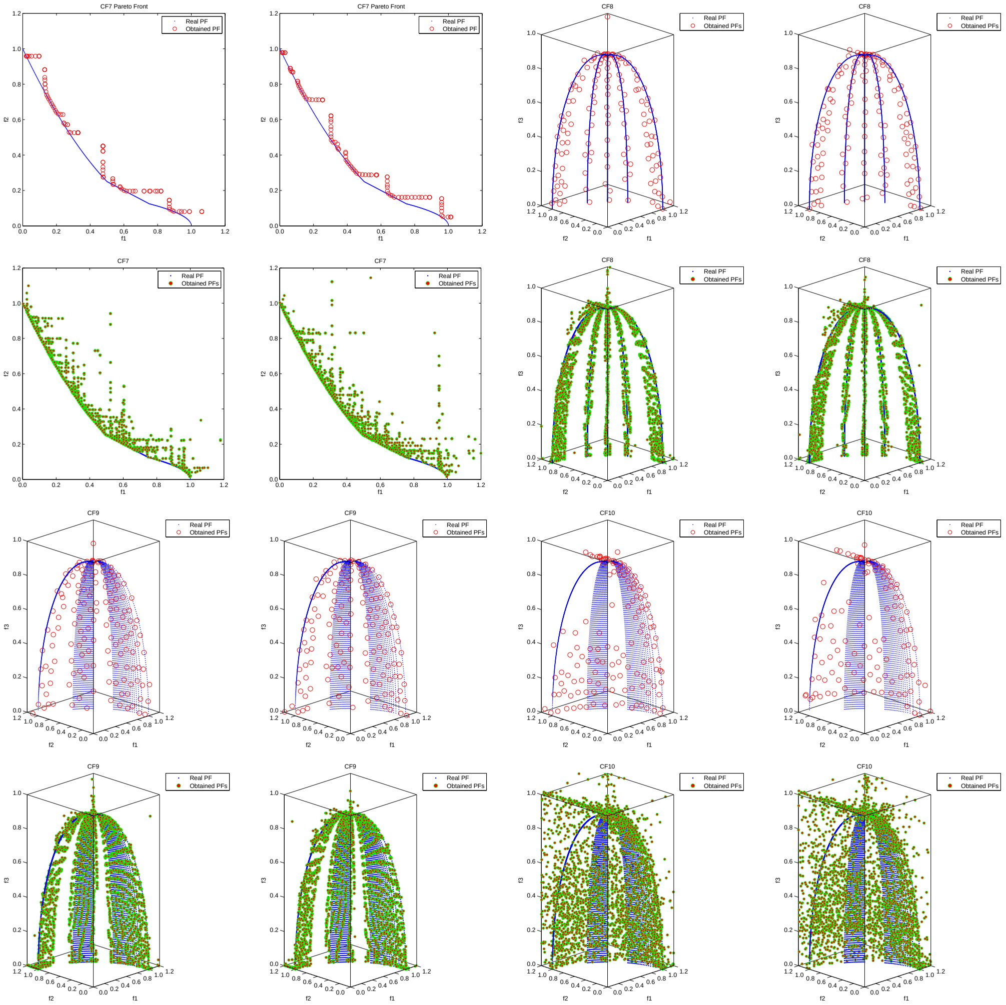


Fig. 2: Plots of the nondominated front with the best IGD value and all the 30 final nondominated fronts found by CMOEA/D-DE-ATP when using Eq. 7 (columns 1 and 3) and Eq. 8 (columns 2 and 4) for CF7-CF10.

all the 30 final nondominated fronts of these selected 100 and 150 nondominated solutions.

It is very clear from these figures that CMOEA/D-DE-ATP with both Eqs. 7, 8 found good approximations for the four 2-objective, CF1, CF2, CF4, and CF6, and two 3-objective, CF8, CF9 test instances. However, it performed poorly on test instances CF3, CF5, CF7, and CF10. It is also apparent from the plots of 30 nondominated fronts of test instances CF4 and CF6 that CMOEA/D-DE-ATP fails to find the whole PF in some runs.

The PF of CF3 is concave and discontinuous. Therefore, it could be hard for the algorithm than all other 2-objective test instances. Although the PFs of CF4 and CF5, CF6 and CF7, and CF9 and CF10 are identical, the poor performance of CMOEA/D-DE-ATP on test instances CF5, CF7 and CF10 could be due to the presence of harder objective and constraint functions in these test instances than test instances CF4, CF6, and CF9.

Figure 3 shows the evolution of the average IGD-metric values versus function evaluations of the nondominated solutions in the current population. This figure shows that CMOEA/D-DE-ATP with both Eqs. 7, 8 converges at the same rate in terms of IGD-metric values for six CF-series test instances CF1, CF3, CF4, CF6, CF8 and CF9. However, it converges slightly faster in terms of IGD-metric values for the other four CF-series test instances CF2, CF5, CF7, and CF10 with Eq. 7 than with Eq. 8.

Figure 4 depicts the average generation feasibility versus generations' graphs. This figure demonstrates that CMOEA/D-DE-ATP with both Eqs. 7, 8 approaches to the feasible regions at the same rate for test instances CF1-CF6. It converges slower to the feasible regions with Eq. 7 than with Eq. 8 for the test instances CF7-CF10. This permits further exploration of the infeasible regions near the PF and could be one of the reasons for the better performance of CMOEA/D-DE-ATP with Eq. 7 on the three 3-objective test instances, CF8-CF10.

As it can be seen from Figure 4 that 50 % or more of the initial populations for test instances CF1, CF2, CF4 and CF5 are feasible. These feasible solutions are propagated in the subsequent generations by the replacement and update scheme

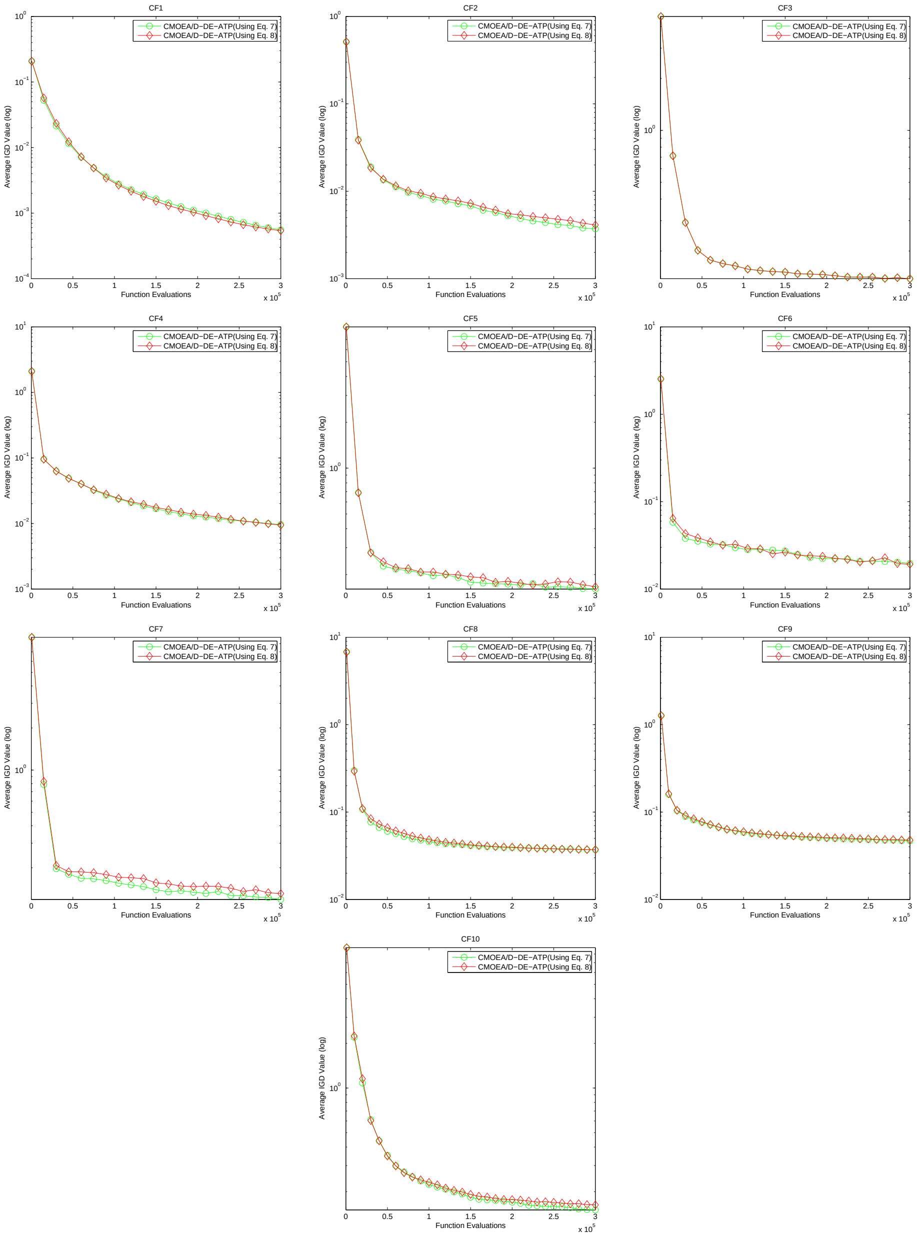


Fig. 3: Evolution of the IGD-metric values versus function evaluations when CMOEA/D-DE-ATP uses Eqs. 7, 8 for CF1-CF10.

of the algorithm and thus produce better feasible solutions due to the DE operator. Furthermore, the feasibility ratio becomes 1 after the initial 30 to 40 generations for these test instances. The reason for the quick convergence to the feasible region is the higher adopted update number of neighboring parent solutions (as in our settings $n_r = 6$ when $T = 60$ and $n_r = 10$ when $T = 100$) that are replaced by a better child solution in the update scheme of the algorithm. This speedy convergence to the feasible region is good for test instances like CF1, CF2, and CF4, but it causes problems for harder test instance like CF5. The PF of CF5 is a piecewise continuous curve with three pieces like CF4, but its objective and constraint functions are quite different and harder than CF4.

On the other hand, about 25 $\%$ or below of the initial populations for test instances CF6-CF10 is feasible. Here, the proposed constraint handling technique has more chances to evolve better infeasible solutions during the evolutionary process. Particularly, in the two 3-objective test instances CF8 and CF9, the average feasibility ratio at the last generations of the algorithmic runs is 0.6 and about 0.9, respectively (see Figure 4). This way the infeasible regions near the feasibility boundaries in these two instances are well explored and could be a reason for the better performance of the algorithm on these two instances.

Moreover, in test instance CF6, the feasibility ratio of CMOEA/D-DE-ATP with both Eqs. 7, 8 becomes 1 after the initial 40 generations, while in test instance CF7, it takes 200 generations of CMOEA/D-DE-ATP with Eq. 7 and 50 gener-

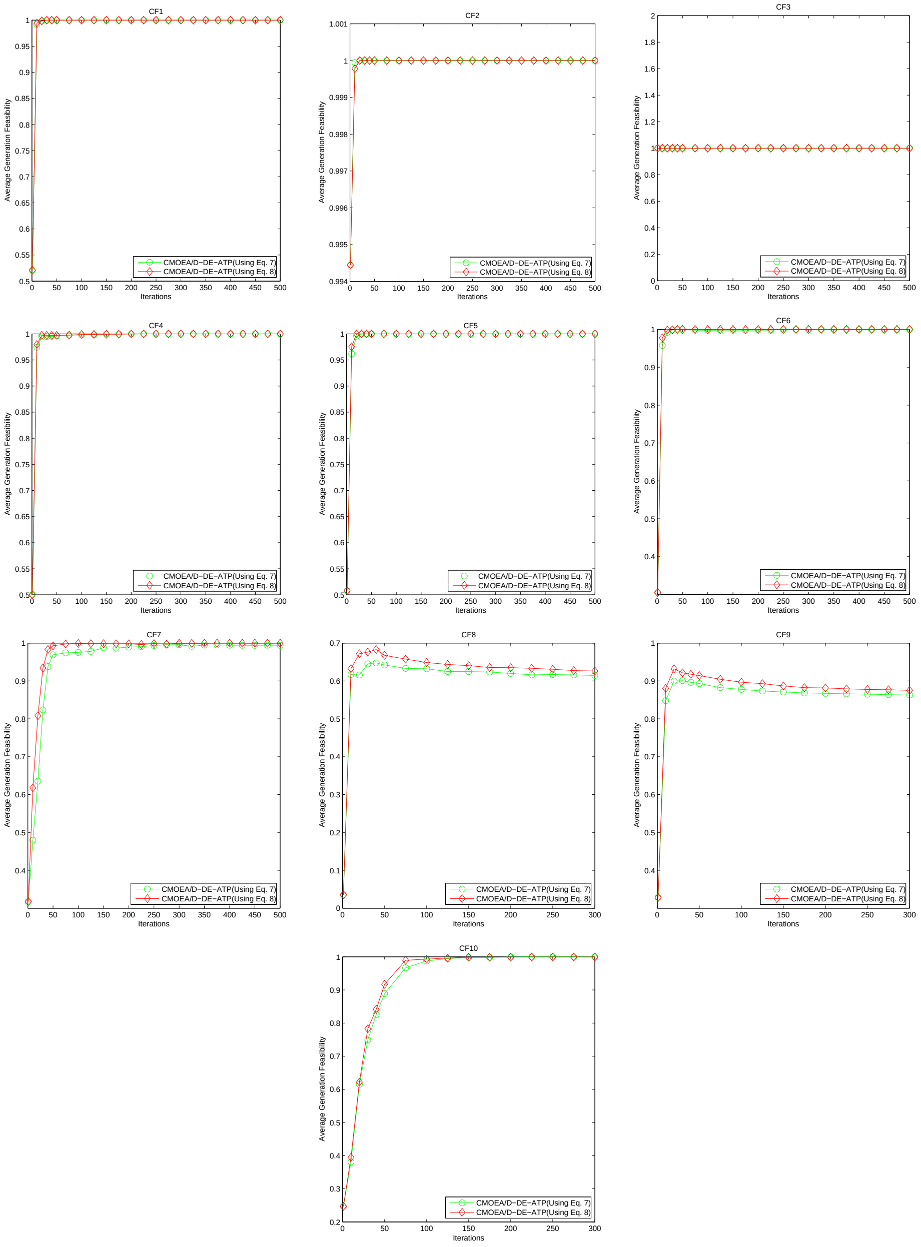


Fig. 4: Evolution of the generation feasibility versus generations when CMOEA/D-DE-ATP uses Eqs. 7, 8 for CF1-CF10.

ations with Eq. 8 to become 1. Thus, the quick convergence to the feasible region is good in case of CF6, but could be a reason for the poor performance of CMOEA/D-DE-ATP with both Eqs. 7, 8 in case of CF7.

In test instance CF10, the feasibility ratio takes about 140 generations of CMOEA/D-DE-ATP with Eq. 7 and about 80 generations of CMOEA/D-DE-ATP with Eq. 8 to become 1. Again, the less exploration of the infeasible regions could be the reason for the poor performance of CMOEA/D-DE-ATP on test instance CF10.

VI. Comparison with the Three Best Performers of CEC 2009 MOEA Competition

In this section, we compare the results of CMOEA/D-DE-ATP with Eqs. 7, 8 with the three best performers [14]–[16] in CEC 2009 MOEA competition on the CF-series test instances.

Table III compares the best (i.e., lowest), mean, and standard deviation values of the IGD-metric obtained from our algorithm, CMOEA/D-DE-ATP with Eqs. 7, 8, and the three best performers [14]–[16] in CEC 2009 MOEA competition for the CF-series test instances. The table clearly shows that CMOEA/D-DE-ATP has found the best (i.e., lowest) IGD-metric values for four test instances CF1, CF6, CF8 and CF9 and the second best value for one test instance CF3 with Eq.

TABLE III: COMPARISON BETWEEN CMOEA/D-DE-ATP WITH EQ. 7 (INDICATED BY JZ1) AND WITH EQ. 8 (INDICATED BY JZ2), TSENG AND CHEN'S [14] (INDICATED BY TC), LIU AND LI'S [15] (INDICATED BY LL), AND LIU ET. AL'S [16] (INDICATED BY LI) ALGORITHMS IN TERMS OF THE IGD VALUES BASED ON 30 INDEPENDENT RUNS. THE RESULTS IN **BOLDFACE** AND IN *ITALIC* INDICATE THE BETTER AND THE SECOND BETTER RESULTS.

	best (lowest)					mean					st. dev.				
Test Instance	JZ1	JZ2	TC	LL	LI	JZ1	JZ2	TC	LL	LI	JZ1	JZ2	TC	LL	LI
CF1	**0.0003**	**0.0003**	0.0139	*0.0007*	0.0071	*0.0006*	**0.0005**	0.0192	0.0009	0.0113	0.0003	*0.0002*	0.0026	**0.0001**	0.0028
CF2	0.0028	*0.0027*	0.0041	*0.0027*	**0.0016**	*0.0037*	0.0041	0.0268	0.0042	**0.0021**	*0.0013*	0.0019	0.0147	0.0026	**0.0005**
CF3	*0.0632*	*0.0632*	0.0753	0.0908	**0.0381**	0.1382	0.1382	*0.1045*	0.1829	**0.0563**	0.0441	0.0441	*0.0156*	0.0421	**0.0076**
CF4	0.0060	**0.0051**	0.0089	0.0090	*0.0055*	0.0097	*0.0095*	*0.0111*	0.0142	**0.0070**	0.0042	0.0043	**0.0014**	0.0033	*0.0015*
CF5	0.0406	0.0297	*0.0176*	0.0588	**0.0079**	0.1606	0.1663	*0.0208*	0.1097	**0.0158**	0.1084	0.1107	**0.0024**	0.0307	*0.0067*
CF6	**0.0049**	0.0053	0.0096	0.0090	*0.0062*	0.0197	0.0192	0.0162	**0.0139**	*0.0150*	0.0141	0.0144	*0.0060*	**0.0026**	0.0065
CF7	0.0344	0.0304	*0.0187*	0.0535	**0.0104**	0.1188	0.1310	*0.0247*	0.1045	**0.0191**	0.0729	0.0722	**0.0047**	0.0351	*0.0061*
CF8	**0.0332**	*0.0356*	0.6220	0.0473	0.0388	**0.0370**	*0.0371*	1.0854	0.0607	0.0475	*0.0020*	**0.0010**	0.2191	0.0130	0.0064
CF9	**0.0428**	*0.0434*	0.0721	0.0460	0.1191	**0.0468**	*0.0479*	0.0851	0.0505	0.1434	**0.0022**	*0.0030*	0.0082	0.0034	0.0214
CF10	0.1068	0.1108	0.1173	*0.1055*	**0.0984**	*0.1509*	0.1630	**0.1376**	0.1974	0.1621	0.0396	0.0409	**0.0092**	0.0760	*0.0316*

7. It has also found the best IGD-metric values for two test instances CF1 and CF4 and the second best values for four test instances CF2, CF3, CF8, and CF9 with Eq. 8. Particularly, for test instances CF1, CF8 and CF9 better statistics are found by our algorithm except the standard deviation value on CF1 (although both our standard deviation values are very close to the best standard deviation value).

VII. Conclusions

A penalty function that penalizes infeasible solutions based on an adaptive threshold value has been introduced into the update and replacement scheme of MOEA/D-DE. This resulted in a new algorithm, CMOEA/D-DE-ATP for CMOO. The proposed penalty function is presented in two forms given by Eqs. 7, 8. The performance of CMOEA/D-DE-ATP is tested on CF-series test instances in terms of the values of IGD-metric and SC-metric.

From the experimental results in this paper, we can make the following conclusions.

- Overall, CMOEA/D-DE-ATP produced better results with the proposed penalty function defined by Eq. 7 than when it is defined by Eq. 8. That is, it is better to add the penalty to individual objective function values before calculating the aggregation function values than directly adding the penalty to aggregation function values of an infeasible solution for better performance achievement on CF-series test instances.
- The comparison of CMOEA/D-DE-ATP with the three best performers in CEC 2009 special session and competition indicated that CMOEA/D-DE-ATP has found the best (i.e., lowest) IGD-metric values for five test instances CF1, CF4, CF6, CF8 and CF9 and the second best values for two test instances CF2 and CF3. In particular, our algorithm overall found better statistics for tests instances CF1, CF8, and CF9.

References

[1] M. A. Jan and Q. Zhang, "MOEA/D for constrained multiobjective optimization: Some preliminary experimental results," in *UK Workshop on Computational Intelligence (UKCI)*. IEEE, 2010, pp. 1–6.

[2] K. Miettinen, *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.

[3] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons LTD, 2001.

[4] M. Gen and R. Cheng, "Interval Programming using Genetic Algorithms," in *Proceedings of the Sixth International Symposium on Robotics and Manufacturing*, Montpellier, France, 1996.

[5] E. M. Montes and C. A. Coello Coello, "A simple multimembered evolution strategy to solve constrained optimization problems: Smes," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 1, pp. 1–17, February 2005.

[6] T. Ray, H. K. Singh, A. Isaacs, and W. Smith, "Infeasibility driven evolutionary algorithm for constrained optimization," in *Constraint-Handling in Evolutionary Computation*, E. Mezura-Montes, Ed. Berlin: Springer. Studies in Computational Intelligence, Volume 198, 2009, ch. 7, pp. 145–165, ISBN 978-3-642-00618-0.

[7] Y. G. Woldesenbet, G. G. Yen, and B. G. Tessema, "Constraint handling in multiobjective evolutionary optimization," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 3, June 2009.

[8] G. Yen, "An adaptive penalty function for handling constraint in multi-objective evolutionary optimization," *Constraint-Handling in Evolutionary Optimization*, pp. 121–143, 2009.

[9] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W H Freeman and Co San Francisco, 1979.

[10] S. S. Rao, *Engineering Optimization*, 3rd ed. John Wiley & Sons, 1996.

[11] H. Li and Q. Zhang, "Multiobjective Optimization Problems with Complicated Pareto Sets, MOEA/D and NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 284–302, April 2009.

[12] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[13] Q. Zhang, W. Liu, and H. Li, "The Performance of a New Version of MOEA/D on CEC09 Unconstrained Mop Test Instances," in *Special Session on Performance Assessment of Multiobjective Optimization Algorithms/CEC 09 MOEA Competition*, Norway, 18-21 May 2009.

[14] L. Tseng and C. Chen, “Multiple trajectory search for unconstrained/constrained multi-objective optimization,” in *IEEE Congress on Evolutionary Computation, CEC2009*. IEEE, 2009, pp. 1951–1958.

[15] H. Liu and X. Li, “The multiobjective evolutionary algorithm based on determined weight and sub-regional search,” in *IEEE Congress on Evolutionary Computation, CEC2009*. IEEE, pp. 1928–1934.

[16] M. Liu, X. Zou, Y. Chen, and Z. Wu, “Performance assessment of DMOEA-DD with CEC 2009 MOEA competition test instances,” in *IEEE Congress on Evolutionary Computation, CEC2009*. IEEE, 2009, pp. 2913–2918.

[17] K. M. Miettinen, *Nonlinear Multiobjective Optimization.* Kluwer Academic Publishers, 1999.

[18] M. A. Jan and R. A. Khanum, “A study of two penalty-parameterless constraint handling techniques in the framework of MOEA/D,” *Appl. Soft Comput.*, vol. 13, no. 1, pp. 128–148, 2013.

[19] Q. Zhang, A. Zhou, S. Zhao, P. N. Suganthan, W. Liu, and S. Tiwari, “Multiobjective optimization test instances for the CEC 2009 special session and competition,” University of Essex and Nanyang Technological University, Tech. Rep. CES-487, 2008.

Threshold Based Penalty Functions for Constrained Multiobjective Optimization

Muhammad Asif Jan
Department of Mathematics
Kohat University of Science & Technology
Khyber Pakhtunkhwa, Pakistan

Nasser Mansoor Tairan
College of Computer Science
King Khalid University Abha,
Saudi Arabia

Rashida Adeeb Khanum
Jinnah College for Women
University of Peshawar Khyber
Pakhtunkhwa, Pakistan

Wali Khan Mashwani
Department of Mathematics
Kohat University of Science & Technology
Khyber Pakhtunkhwa, Pakistan

***Abstract*—This paper compares the performance of our recently proposed threshold based penalty function against its dynamic and adaptive variants. These penalty functions are incorporated in the update and replacement scheme of the multiobjective evolutionary algorithm based on decomposition (MOEA/D) framework to solve constrained multiobjective optimization problems (CMOPs). As a result, the capability of MOEA/D is extended to handle constraints, and a new algorithm, denoted by CMOEA/D-DE-TDA is proposed. The performance of CMOEA/D-DE-TDA is tested, in terms of the values of IGD-metric and SC-metric, on the well known CF-series test instances. The experimental results are also compared with the three best performers of CEC 2009 MOEA competition. Empirical results show the pitfalls of the proposed penalty functions.**

***Keywords*—Constrained multiobjective optimization; decomposition; MOEA/D; dynamic and adaptive penalty functions; threshold.**

I. Introduction

In this paper, we consider the following constrained multiobjective optimization problem (CMOP) [1]:

$$\begin{array}{ll} \text{Minimize} & F(x) = (f_1(x), f_2(x), \dots, f_m(x))^T; \\ \text{Subject to} & g_j(x) \geq 0,\ j = 1, \dots, p; \\ & l_k \leq x_k \leq u_k,\ k = 1, \dots, n; \end{array} \tag{1}$$

where $x = (x_1, \dots, x_n)^T \in \mathcal{R}^n$ is an n dimensional vector of decision variables, F is the objective vector function that consists of m real-valued objective functions, and $g_i(x) \geq 0$ are inequality constraints. The objective and constraint functions, f_i's and g_j's, could be linear or non linear real-valued functions. l_k and u_k are the lower and upper bounds (also called bound constraints) of x_k, $k = 1, \dots, n$, respectively, which define the search region, $\mathcal{S} = \{x = (x_1, \dots, x_n)^T |\ l_k \leq x_k \leq u_k, k = 1, \dots, n\}$.

A solution $x \in S$ satisfying all the inequality constraints in (1) is called a feasible solution; otherwise, we call it an infeasible solution. The set of all feasible solutions is called the feasible region, denoted by $\mathcal{F}$, and the set of all infeasible solutions is called the infeasible region. Also, we define feasible attainable objective set (AOS) by $\{F(x)|x \in \mathcal{F}\}$.

Because the objectives in (1) often contradict one another, a single solution in the feasible search region could not be found that minimizes all the objectives simultaneously. Therefore, a set of optimal compromising/tradeoff solutions that satisfy all constraints (i.e., feasible solutions) is desired. The best tradeoffs among the objectives can be defined in terms of Pareto-optimality [2], [3].

A solution x is said to Pareto-dominate or simply dominate another solution y, mathematically denoted as $x \preceq y$, if $f_i(x) \leq\ f_i(y)$, $\forall i = 1, \dots, m$ and $f_j(x) <\ f_j(y)$ for at least one $j \in \{1, \dots, m\}$[1]. This definition of domination is sometimes referred to as a weak dominance relation.
A solution $x^* \in \mathcal{F}$ is Pareto-optimal to (1) if there is no solution $x \in \mathcal{F}$ such that $F(x) \preceq F(x^*)$. $F(x^*)$ is then called a Pareto-optimal (objective) vector. The set of all Pareto-optimal solutions is called the Pareto Set (PS) in the decision space and Pareto Front (PF) in the objective space [2].

A common way to deal with constraints in constrained optimization is to use penalty functions. In a penalty function approach, the penalty coefficients balance the objective and penalty functions. However, finding appropriate penalty coefficients to strike the right balance is a big challenge in itself [4]. They depend on the problems in hand. Thus, some researchers suggested dynamic penalty functions [5]–[7] to avoid the difficulty of setting penalty coefficients in static penalty functions and to explore infeasible regions.

In dynamic penalty functions, the current generation number (or the number of solutions searched) is considered in the calculation of the penalty coefficients. The penalty to infeasible solutions is small at the beginning of the search due to the initial small generation numbers used in the formulation, and it then increases by the increase in the generation number. As a result, these methods converge to feasible solution (s) at the end of evolution. Generally, dynamic penalty functions are effective, but they are not adaptive to the ongoing success (or failure thereof) of the search and cannot guide the search

[1] One has to reverse all the inequalities if the goal is to maximize the objectives in (1). By the term "dominate" we mean "better than"

to particularly promising regions or away from unpromising regions based on what has already been observed [6]. Furthermore, like static penalty functions, they also need problem specific tuning to perform well [6].

Adaptive penalty functions not only incorporate the current generation number (or search length), but they also consider the feedback from the search in their formulations. Thus, they are benefited from the search history. The penalty coefficients in such methods are adjusted based on what has already been accomplished.

In [1], [8], we proposed a novel threshold-based penalty function for handling constraints in constrained multiobjective optimization (CMOO). The threshold using the minimum and maximum constraint violation in the neighborhood of a solution and controlled by a scaling parameter dynamically adjusts the penalty to infeasible solutions. Moreover, Infeasible solutions with constraint violation less than the threshold are less penalized than the ones with constraint violation greater than the threshold. For this purpose, two additional parameters are used.

In this paper, first we propose a parameterless threshold value contrary to the one used in [8]. Secondly, we define the dynamic and adaptive versions of the threshold based penalty function [8]. These penalty functions are then implemented in one of the improved frameworks of MOEA/D [9], MOEA/D-DE [10] to solve hard CF-series [11] test instances.

The remainder of this paper is organized as follows. Section II presents the Tchebycheff aggregation function and the threshold-based penalty function and its dynamic and adaptive variants. Section III briefly introduces MOEA/D and modifies the algorithmic framework of MOEA/D-DE [10] for CMOPs. Section IV discusses the experimental settings. Section V presents and discusses experimental results on CF-series [11] test instances. Section VI compares our experimental results with the three best performers [12]–[14] of CEC 2009 MOEA competition. Finally, Section VII outlines a summary of the paper.

II. Tchebycheff Aggregation Function and Threshold Based Penalty Functions

A. Tchebycheff Aggregation Function

MOEA/D [9] needs to decompose a multiobjective optimization problem (MOP) into a number of scalar objective subproblems. For this purpose, this work uses the Tchebycheff aggregation function. The reasons for choosing this function include: first, it is less sensitive to the shape of PF. Second, it can be used to find the Pareto-optimal solutions in both convex and nonconvex PFs. It is defined as follows [15]:

$$\begin{array}{ll} \text{Minimize} & g^{te}(x|\lambda, z^*) = \max_{1\le i\le m}\{\lambda_i|f_i(x) - z_i^*|\}; \\ \text{Subject to} & x \in \mathcal{F} \subset \mathcal{R}^n; \end{array}$$

where $z^* = (z_1^*, \ldots, z_m^*)^T$ is the reference point, i.e., $z_i^* = \min\{f_i(x)|x \in \mathcal{F}\}\ \forall i = 1, \ldots, m$ and $\lambda = (\lambda_1, \ldots, \lambda_m)^T$ is a weight vector such that $\lambda_i \geq 0\ \forall i = 1, \ldots, m$ and $\sum_{i=1}^m \lambda_i = 1$.
Theorems describing the Pareto-optimality conditions of Tchebycheff aggregation function are available in [2].

B. Threshold Based Penalty Functions

The proposed penalty function in [1], [8] uses a threshold value, τ for dynamically controlling the amount of penalty to infeasible solutions. It is computed as follows.

Assume that MOEA/D [9] decomposes the given MOP into N subproblems. In each iteration, MOEA/D keeps N individual solutions $x^1, \ldots, x^N$, where x^i is the current solution to subproblem i. If P is supposed to be the mating and update range in MOEA/D (for details, please see Algorithm 1, Section III). Then we define [1], [8]:

$$V_{min} = \min\{V(x^i), i \in P\}. \tag{2}$$

$$V_{max} = \max\{V(x^i), i \in P\}. \tag{3}$$

Where $V(x^i) = |\sum_{j=1}^p \min(g_j(x^i), 0)|$ is the degree of constraint violation of solution x^i. Note that the constraints are normalized before calculating this value.

The threshold value, τ is defined as [1], [8]:

$$\tau = V_{min} + s(V_{max} - V_{min}), \tag{4}$$

where parameter s controls the threshold value, τ.

The proposed threshold based penalty function encourages the algorithm to search the feasible region and the infeasible region near the feasible region. It is defined as follows [1], [8]:
For $i = 1, \ldots, m$

$$f_p^i(x) = \begin{cases} f_i(x) + s_1 V^2(x), & \text{if } V(x) < \tau\ ; \\ f_i(x) + s_1\tau^2 + s_2(V(x) - \tau), & \text{otherwise.} \end{cases} \tag{5}$$

In [8], we have tested the algorithm with a fixed value of parameter s (i.e., we used $s = 0.3$ in Eq. 4). However, due to the changing values of V_{min} and V_{max} during the evolutionary process, the resulting threshold value, τ still remains adaptive.

It might also possible that if the values of V_{min} and V_{max} are quite away from one another, then the resulting threshold value, τ will be large. Consequently, more infeasible solutions will have a chance to propagate in the evolution process, and the search might quickly converge to the feasible region. As a result, the low quality optimal solutions might be obtained. Thus, in order to avoid such situation, this paper sets the threshold value, τ equal to the mean value of the degree of constraint violations of all infeasible solutions in the neighborhood of a solution. That is [1]:

$$\tau = \frac{1}{n_{inf}} \sum_{i\in P} V(x^i), \tag{6}$$

where n_{inf} is the number of infeasible solutions in the neighborhood of a solution. This reduces the effort to choose various values for parameter s.

Eq. 5 employs two additional fixed penalty parameters, s_1 and s_2, where $s_1 << s_2$, to penalize infeasible solutions. Infeasible solutions whose degree of constraint violation is smaller than τ are less penalized than the ones with degree of constraint violation greater than τ. This is realized by scaling the respective violations by parameters s_1 and s_2. However, this paper adjusts s_1 and s_2 dynamically and adaptively. As

a result, the dynamic and adaptive variants of Eq. 5 are established. These variants are defined as follows [1]
For $i = 1, \ldots, m$

$$f_p^i(x) = \begin{cases} f_i(x) + (g/G)^2 V^2(x), \text{ if } V(x) < \tau; \\ f_i(x) + (g/G)^2 \tau^2 + \\ (g/G)(V(x) - \tau), \text{ otherwise,} \end{cases} \tag{7}$$

where $f_p^i(x)$ is the i-th penalized objective function value, g is the current generation number and G is the total number of generations.
When $r_{inf} \neq 0$,

$$f_p^i(x) = \begin{cases} f_i(x) + (r_{inf})^2 V^2(x), \text{ if } V(x) < \tau; \\ f_i(x) + (r_{inf})^2 \tau^2 + \\ (r_{inf})(V(x) - \tau), \text{ otherwise,} \end{cases} \tag{8}$$

where $f_p^i(x)$ is the i-th penalized objective function value and r_{inf} is the infeasibility ratio (i.e., the ratio of the number of infeasible solutions to the total number of neighboring solutions) in the neighborhood of a solution.

Initially in Eq. 7, infeasible solutions are less penalized due to small g values. Thus, infeasible solutions with degree of constraint violation less than τ get particularly more chances to evolve. As a result, infeasible regions are well explored in the beginning of the search. However, later on with the increase in generation number, g, the penalty to infeasible solutions increases. As a result, the search converges to the feasible region at the later stage of the algorithmic run.

On the other hand in Eq. 8, the penalty to infeasible solutions depends on the number of infeasible solutions in the neighborhood of a solution. Thus, the less is the number of infeasible solutions in the neighborhood of a solution, the smaller the penalty is; otherwise, it will increase with the increase in the number of infeasible solutions. Here again, because of the employed penalty parameters, infeasible solutions with degree of constraint violation less than τ get more opportunities to evolve.

Algorithm 1 Pseudo-code of Update Scheme of CMOEA/D-DE-TDA. n_r is the number of solutions updated by a better child solution.

```
1:  Each new child solution $y$ updates $n_r$ solutions from the
    set $P$ of its neighboring solutions as follows:
2:  Set $c = 0$ and then do the following:
3:  if $c = n_r$ or $P$ is empty then
4:      return;
5:  else
6:      Randomly pick an index $j$ from $P$;
7:      Compute the Tchebycheff aggregation function values of $y$ and $x^j$
        with the new objective values of Eqs. 5, 7, and 8;
8:      if $g^{te}(y|\lambda^j, z) \leq g^{te}(x^j|\lambda^j, z)$ then
9:          $x^j = y$, $F(x^j) = F(y)$, $V(x^j) = V(y)$, and $c = c + 1$;
10:     end if
11:     Remove $j$ from $P$ and go to step 3;
12: end if
```

III. MOEA/D AND CMOEA/D-DE-TDA

Zhang and Li [9] proposed a simple but efficient MOEA called MOEA/D. It approximates the PF by explicitly decomposing an MOP into a number of scalar objective optimization subproblems (e.g., Tchebycheff aggregation function is employed for this purpose). These subproblems are then optimized concurrently and collaboratively by evolving a population of solutions. An EA is employed for this purpose. The neighborhood relations among these subproblems are defined based on the Euclidean distances between their aggregation coefficient vectors. Optimization of a subproblem uses the information, mainly from its neighboring subproblems.

MOEA/D-DE [10] is an updated and efficient version of MOEA/D. The framework of this algorithm is altered for CMOPs. The modified framework is denoted by CMOEA/D-DE (pleas see [16] for more details).

We employ the penalty functions defined by Eqs. 5, 7, and 8 with τ given by Eq. 6 in the replacement and update scheme of CMOEA/D-DE to solve CF-series [11] test instances. This resulted in a new algorithm, denoted by CMOEA/D-DE-TDA. The pseudo-code of the update scheme of CMOEA/D-DE-TDA is given in Algorithm 1 [1].

IV. EXPERIMENTAL SETTINGS

In our experiments, the same parameters' settings are used as in [16]. The weight vectors used in Eq. (2) are set as per criteria mentioned in [11]. We employ the inverted generational distance metric (IGD-metric) [9], [17] statistics for comparing the obtained results. For calculating the IGD-metric values, we select 100 feasible nondominated solutions in the case of 2-objective and 150 in the case of 3-objective test instances from each final population. The final solution set P is selected as per criteria given in [11]. We also employ the set coverage metric (SC-metric) [9] to compare the obtained nondominated solutions. One of the reasons for choosing these metrics is that they could measure both convergence and diversity of the approximated solutions in a sense. Secondly, the algorithms in comparison have also used these performance metrics.

V. EXPERIMENTAL RESULTS

In this section, we present the experimental results obtained from CMOEA/D-DE-TDA on CF-series test instances.

Table I presents the best (i.e., lowest), mean, and standard deviation of the IGD-metric values based on 30 independent runs found by CMOEA/D-DE-TDA with Eqs. 5, 7, and 8 for CF-series test instances. As it can be seen from this table that CMOEA/D-DE-TDA with Eq. 5 found better statistics for three test instances CF1, CF6, and CF7 except the standard deviation value on CF7. The table also shows that the algorithm with Eq. 7 found better results than those obtained with Eqs. 5, 8 for test instances CF2, CF4 and CF5 except the best values on CF2 and CF5. Similarly, it showed superior performance with Eq. 8 for the 3-objective test instances except the standard deviation value and best value for test instances CF9 and CF10, respectively. The performance of the algorithm may be considered as comparable with all three formulations for test instances CF2, CF8, and CF9, as there is a marginal difference in the IGD-metric statistics for these test instances.

TABLE I: COMPARISON OF THE IGD-METRIC STATISTICS OF THE ALGORITHM FOR CF1-CF10. THE RESULTS IN **BOLDFACE** AND IN *ITALIC* INDICATE THE BETTER AND THE SECOND BETTER RESULTS.

	best (lowest)			mean			st. dev.		
Test Instance	Eq. 5	Eq. 7	Eq. 8	Eq. 5	Eq. 7	Eq. 8	Eq. 5	Eq. 7	Eq. 8
CF1	**0.0003**	*0.0034*	0.0035	**0.0005**	*0.0051*	0.0076	**0.0002**	*0.0017*	0.0027
CF2	**0.0026**	*0.0028*	**0.0026**	*0.0038*	**0.0037**	0.0045	*0.0014*	**0.0011**	0.0020
CF3	0.0632	0.0632	0.0632	0.1382	0.1382	0.1382	0.0441	0.0441	0.0441
CF4	*0.0056*	**0.0054**	0.0062	*0.0083*	**0.0076**	0.0106	*0.0018*	**0.0016**	0.0041
CF5	**0.0326**	0.0526	*0.0328*	0.1560	**0.1175**	*0.1356*	0.0932	**0.0348**	*0.0921*
CF6	**0.0069**	0.0423	*0.0387*	**0.0220**	0.1667	*0.1241*	**0.0160**	0.1162	*0.0303*
CF7	**0.0374**	*0.0551*	0.0556	**0.1227**	*0.1474*	0.1567	0.0787	**0.0684**	*0.06590*
CF8	0.0344	*0.0336*	**0.0335**	0.0380	*0.0359*	**0.0357**	0.0017	*0.0016*	**0.0013**
CF9	*0.0431*	0.0433	**0.0429**	0.0475	*0.0473*	**0.0471**	*0.0023*	**0.0022**	0.0025
CF10	**0.1020**	*0.1054*	0.1062	0.1602	*0.1487*	**0.1430**	0.0464	*0.0380*	**0.0308**

Overall, the algorithm is to be run with Eq. 5 for better best IGD-metric values on most of the CF-series test instances. However, it is to be run with Eq. 7 or with Eq. 8 for the better mean and standard deviation values.

Moreover, the small values for the mean of IGD-metric for test instances CF1, CF2, CF4, CF8, CF9 show that the final nondominated solutions obtained by the algorithm with all three Eqs. 5, 7, and 8 approximate the PF very well for these test instances, in a sense.

TABLE II: THE AVERAGE SET COVERAGE AMONGST EQs. 5 (S), 7 (D), 8 (A) WHEN USED BY THE ALGORITHM. THE RESULTS IN **BOLDFACE** INDICATE THE BETTER RESULTS; IF NOT, THEY ARE IDENTICAL.

Test Instance	C(S, D)	C(D, S)	C(S, A)	C(A, S)	C(D, A)	C(A, D)
CF1	0.01	**0.94**	0.11	**0.84**	**0.82**	0.16
CF2	0.19	**0.22**	0.18	**0.20**	0.19	**0.22**
CF3	0.65	0.65	0.65	0.65	0.65	0.65
CF4	0.26	**0.32**	**0.30**	0.27	**0.34**	0.26
CF5	**0.22**	0.18	0.21	0.21	0.21	0.21
CF6	0.03	**0.08**	0.03	**0.06**	**0.09**	0.04
CF7	**0.26**	0.23	**0.22**	0.21	**0.25**	0.22
CF8	0.03	**0.06**	0.04	**0.06**	0.05	0.05
CF9	0.03	**0.04**	0.02	**0.03**	0.03	0.03
CF10	0.28	**0.39**	0.31	**0.37**	**0.36**	0.31

Table II presents the average set coverage among the nondominated solutions of CMOEA/D-DE-TDA obtained with Eqs. 5, 7, and 8. The results of this table disclose that, in terms of the SC-metric, the nondominated solutions found by CMOEA/D-DE-TDA with Eq. 7 are better than those obtained with Eq. 5 except for test instances CF5 and CF7. However, due to a small difference in the SC-metric values, the nondominated solutions found by the algorithm with both Eqs. 5, 7 may be considered as comparable for all CF-series test instances except CF1 and CF10. The difference in the SC-metric values is much bigger in case of test instance CF1, while it is reasonably big in case of test instance CF10. Similarly, by the same reason, the nondominated solutions obtained with Eqs. 5, 8 may be considered as comparable except CF1 and CF10, and the ones obtained with Eqs. 7, 8 may be thought as similar except CF1.

Figures 1–5 show, in the objective space, the distributions of the 100 and 150 nondominated population members for the seven 2-objective, CF1-CF7, and the three 3-objective, CF8-CF10, CF-series test instances, respectively. These solutions are selected based on the criteria given in [11] from the final population of the run with the best (i.e., lowest) IGD-metric value among the 30 independent runs. These figures also show all the 30 final nondominated fronts of these selected 100 and 150 nondominated solutions.

From these figures, it is very clear that CMOEA/D-DE-TDA with all three Eqs. 5, 7, and 8 found good approximations of the PFs for the three 2-objective, CF1, CF2, and CF4 and two 3-objective, CF8, CF9, test instances. However, it performed poorly on three 2-objective, CF3, CF5, CF7, and one 3-objective, CF10 test instances. The algorithm found good approximation of the PF for test instance CF6 with Eq. 5 than those obtained with the other two equations Eqs. 7, 8. In the latter case, solutions in the lower part of the PF are not found. Furthermore, it is also evident from the plots of 30 nondominated fronts of test instance CF4 that CMOEA/D-DE-TDA fails to find the whole PF in some runs.

Since the PF of CF3 is discontinuous and concave, so, it might be harder as compared to all other 2-objective test instances for the algorithm. Furthermore, although the PFs of CF4 and CF5, CF6 and CF7, and CF9 and CF10 are identical, the poor performance of CMOEA/D-DE-TDA with all three Eqs. 5, 7, and 8 on test instances CF5, CF7 and CF10 could be due to the presence of relatively harder objective and constraint functions in these test instances than test instances CF4, CF6, and CF9 (see [11]).

Figure 6 presents the evolution of the average IGD-metric values versus function evaluations of the nondominated solutions in the current population. This figure shows that CMOEA-DE-TDA with Eq. 5 converges faster, in terms of IGD-metric values, than with Eqs. 7, 8 for test instances CF1, CF6 and CF7.

For test instances CF2 and CF4, the IGD-metric values obtained by the algorithm with Eq. 8 are higher than those obtained with Eqs. 5, 7. Although initially different, the algorithm with latter two equations performs similarly. Particularly, the IGD-metric values obtained by the algorithm are almost the same in the later generations.

For test instance CF5, the algorithm with Eq. 8 can converge faster in terms of IGD-metric values than with the other two Eqs. 5, 7.

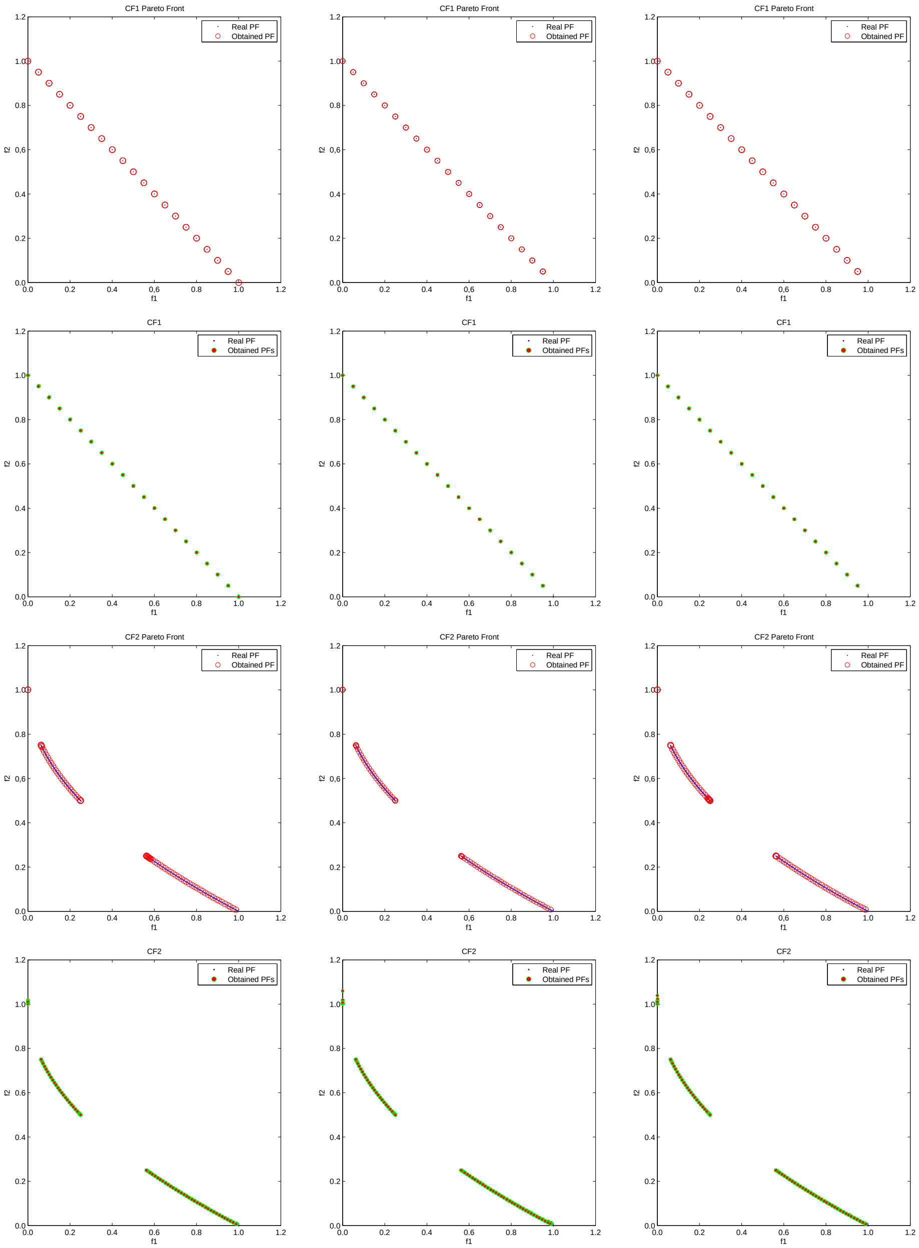


Fig. 1: Plots of the nondominated front with the best IGD value and all the 30 final nondominated fronts found by the algorithm with Eq. 5 (left column), Eq. 7 (middle column), and Eq. 8 (right column) for CF1 and CF2.

Figure 6 also shows that the algorithm with all three Eqs. 5, 7, and 8 converges at the same rate in terms of IGD-metric values for the three 3-objective test instances, CF8-CF10.

Figure 7 shows the average generation feasibility versus generations's graphs. This figure shows that CMOEA-DE-TDA with Eq. 5 converges to the feasible region faster than with Eqs. 7, 8 for all test instances except CF10, where it converges to the feasible region at the same rate with all three Eqs. 5, 7, and 8. Specifically, the feasibility ratio equates to 1 by generations 10 to 50 for test instances CF1-CF7.

The algorithm approaches to the feasible region faster with Eq. 7 than with Eq. 8 for test instances CF1, CF5, CF7, while the situation is vice versa for test instances CF4 and CF6. Moreover, it converges to the feasible region at the same rate for the test instances CF2, CF8, and CF9 with both Eqs. 7, 8.

Figure 7 shows that 50 % or more of the initial populations for test instances CF1-CF5 is feasible. These feasible solutions propagate quickly in the subsequent generations of the algorithm with Eq. 5, and the feasibility ratio becomes 1 by generations 10-50. However, the algorithm retains infeasible solutions until the end of the algorithmic runs with Eqs. 7, 8, and the final feasibility ratios remain in the range (0.6 1) for these test instances. Additionally, with Eq. 7, the feasibility ratio initially remains small even in some cases drops to very

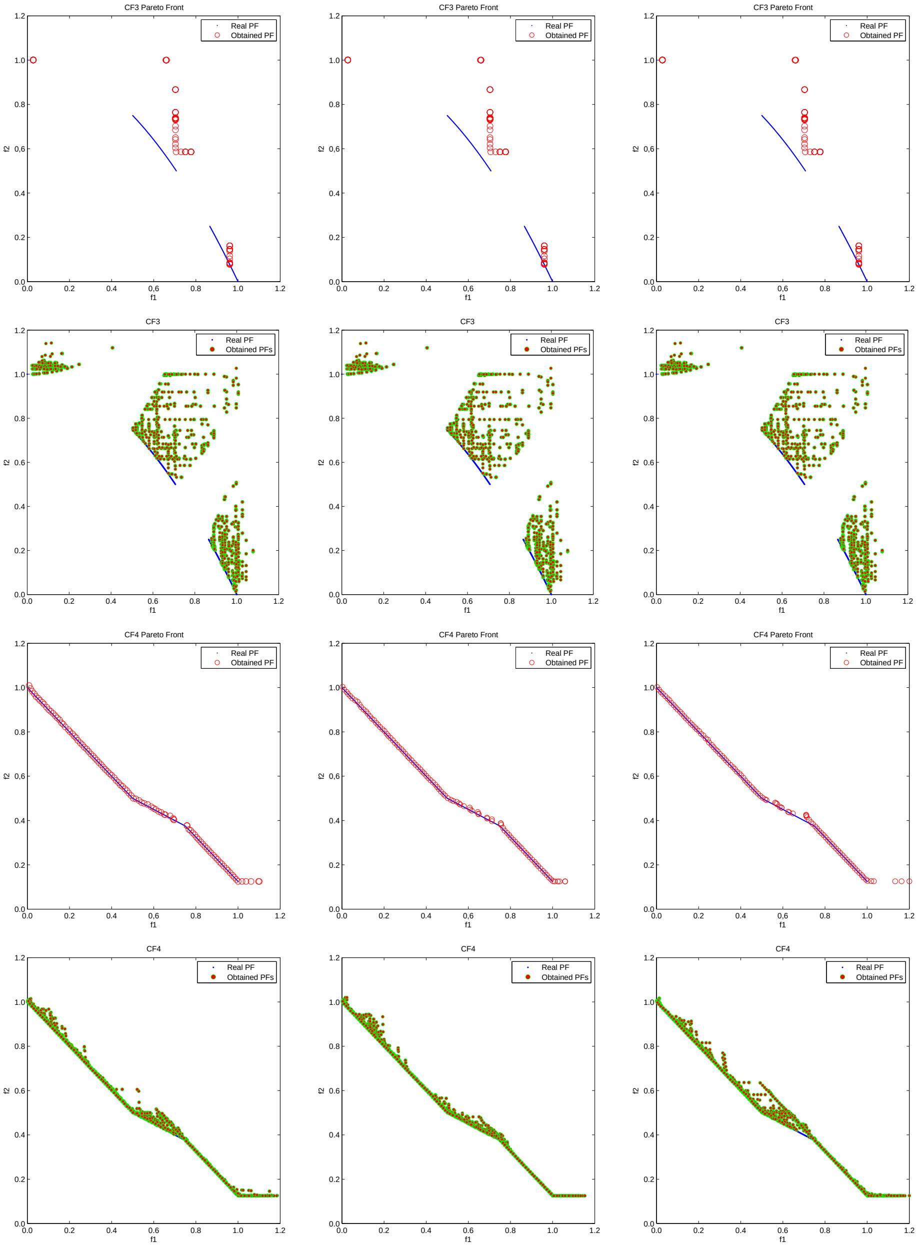


Fig. 2: Plots of the nondominated front with the best IGD value and all the 30 final nondominated fronts found by the algorithm with Eq. 5 (left column), Eq. 7 (middle column), and Eq. 8 (right column) for CF3 and CF4.

low value and then goes on increasing and approaching 1 from there (see Figure 7 for feasibility ratio graphs of test instances CF1 and CF2). The quick convergence to the feasible region is advantageous for test instances like CF1, CF2, and CF4, but can cause problems for harder test instance like CF5, where the PF is a piecewise continuous curve with three pieces like CF4, but its objective and constraint functions are quite different and harder than CF4 (see [11]).

On the other hand, about 25 % or below of the initial populations for test instances CF6-CF10 is feasible. Because of the threshold defined in this paper and individually defined scaling factors in the three penalty functions, the algorithm has more chances to evolve better infeasible solutions during the evolutionary process in these test instances. Particularly, in the two 3-objective test instances CF8 and CF9, the average feasibility ratio at the last generations of the algorithmic runs is about 0.6 and about 0.85 with Eqs. 7, 8 and about 0.7 and about 0.9 with Eq. 5, respectively (see Figure 7). This way the infeasible regions near the feasibility boundaries in these two instances are well explored and could be a reason for the better performance of the algorithm on these two instances with all three equations.

Moreover, in test instance CF6, the feasibility ratio of CMOEA/D-DE-TDA with Eq. 5 becomes 1 after the initial 20

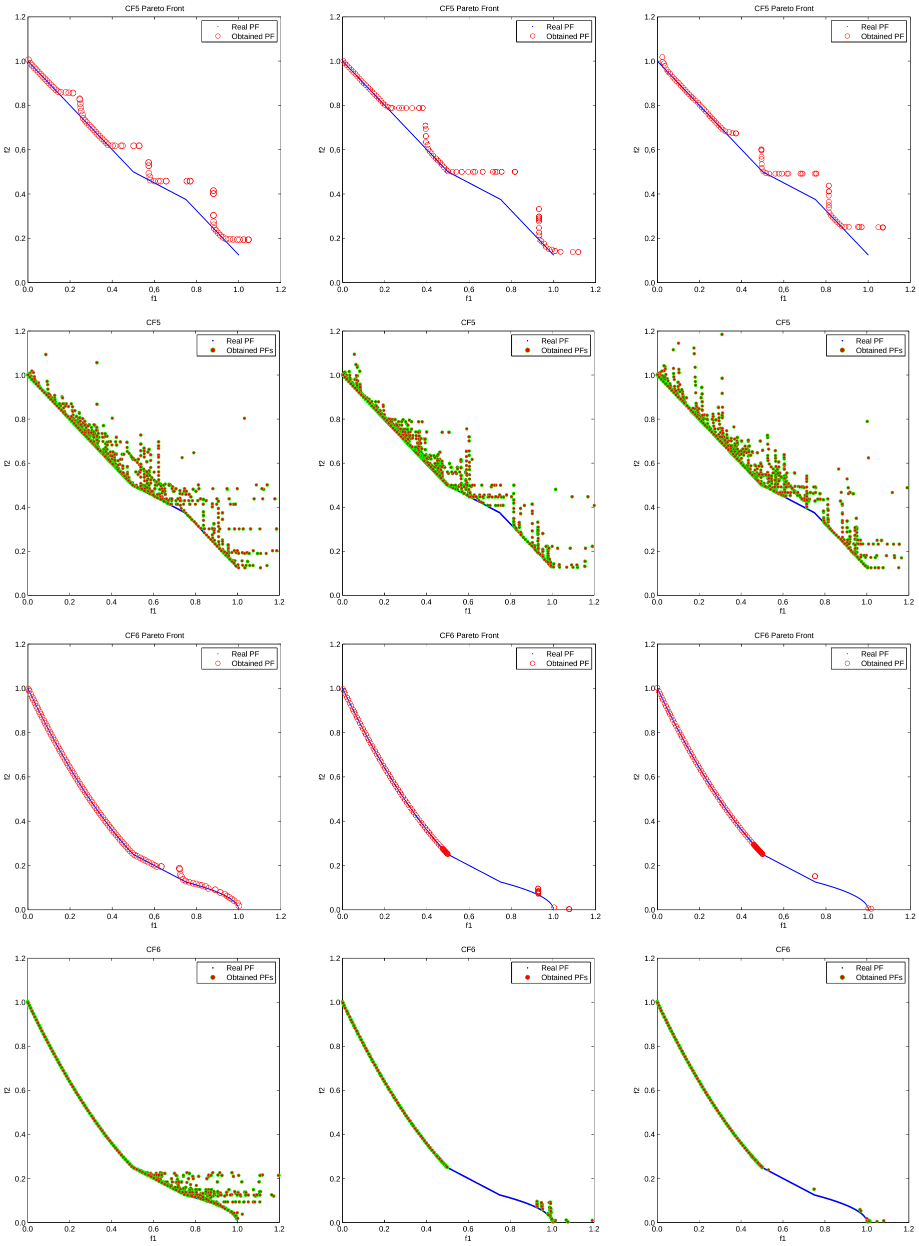


Fig. 3: Plots of the nondominated front with the best IGD value and all the 30 final nondominated fronts found by the algorithm with Eq. 5 (left column), Eq. 7 (middle column), and Eq. 8 (right column) for CF5 and CF6.

generations, and it remains mostly in the range (0.60 0.80) with Eqs. 7, 8, while in test instance CF7, it takes 50 generations of CMOEA/D-DE-TDA with Eq. 5 to become 1 and finally approaches to about 0.72 and about 0.8 with Eq. 8 and Eq. 7, respectively. Thus, the quick convergence to the feasible region and as a result more exploration of it is beneficial in case of Eq. 5 for test instances CF6 and CF7. While, retaining more infeasible solutions could be a reason for a comparatively poor performance of CMOEA/D-DE-TDA with Eqs. 7, 8 on these two test instances.

In test instance CF10, the feasibility ratio takes about 200 generations of CMOEA/D-DE-TDA with Eq. 5 and about 250 generations of CMOEA/D-DE-TDA with Eqs. 7, 8 to become 1. Here, the less exploration of the feasible region could be the reason for the poor performance of CMOEA/D-DE-TDA on this test instance.

VI. Comparison with the Three Best Performers of CEC 2009 MOEA Competition

In this section, the results of CMOEA/D-DE-TDA are compared with the three best performers [12]–[14] in CEC 2009 MOEA competition on the CF-series test instances.

Table III compares the best, mean, and standard deviation values of the IGD-metric obtained from our algorithm, CMOEA/D-DE-TDA with Eqs. 5, 7, and 8 and the three best

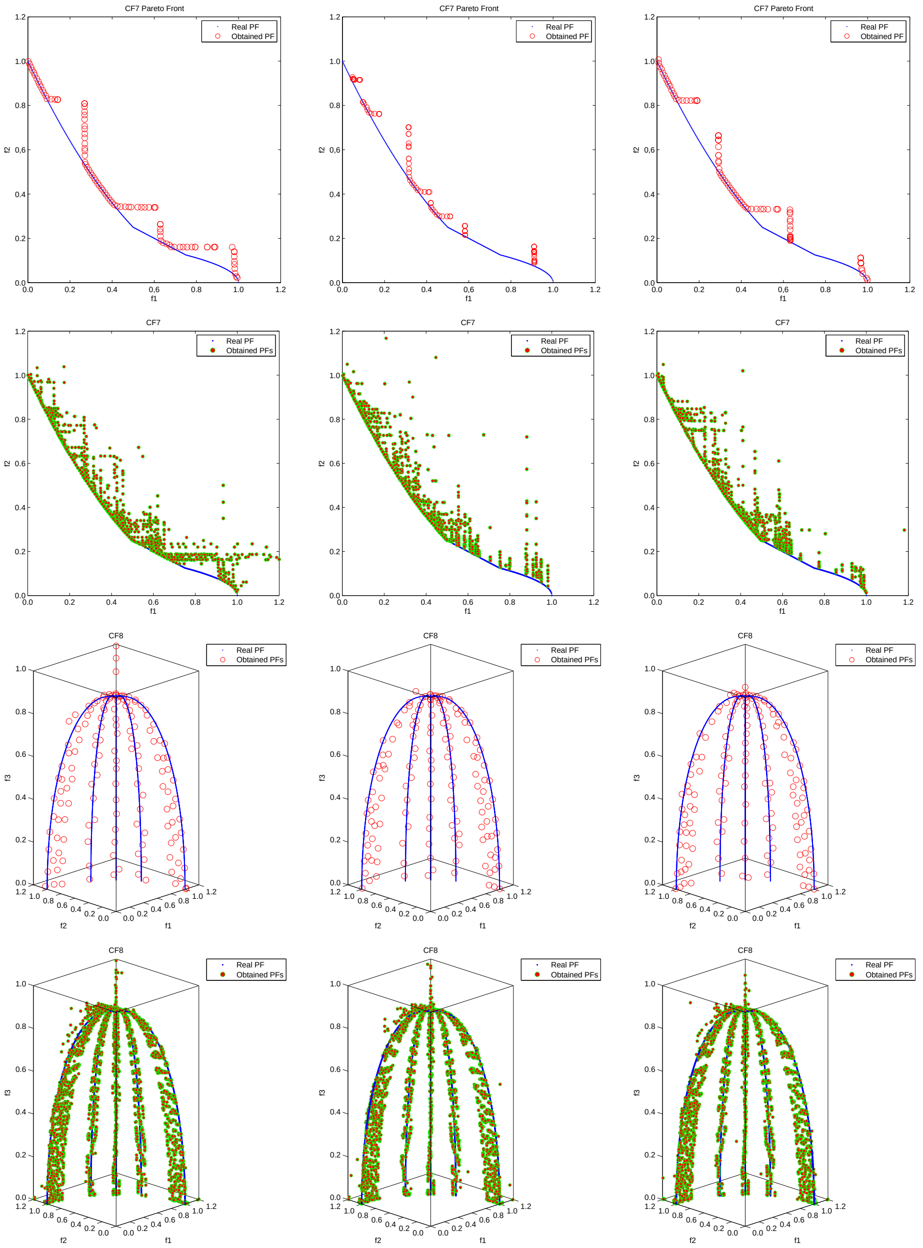


Fig. 4: Plots of the nondominated front with the best IGD value and all the 30 final nondominated fronts found by the algorithm with Eq. 5 (left column), Eq. 7 (middle column), and Eq. 8 (right column) for CF7 and CF8.

performers [12]–[14] in CEC 2009 MOEA competition for the CF-series test instances. It is clear that CMOEA/D-DE-TDA has achieved the best (i.e., lowest) IGD-metric value with Eq. 5 for test instance CF1, with Eq. 7 for test instance CF4, and with Eq. 8 for test instances CF8 and CF9. The table also shows that the algorithm has found the second best values for three test instances CF2, CF6, and CF10. Particulary, better statistics are found by our algorithm for test instances CF1, CF8 and CF9 except the standard deviation value on CF1 (It may be noted that our standard deviation value obtained with Eq. 5 is very close to the best standard deviation value). Further, our algorithm is the second best performer in terms of the mean and standard deviation IGD-metric values with Eq. 8 for test instance CF10.

VII. Conclusions

In this paper, the performance of our proposed algorithm, CMOEA/D-DE-TDA is evaluated with three penalty functions given by Eqs. 5, 7, and 8 on the CF-series test instances. Here, Eqs. 7, 8 are the dynamic and adaptive versions of Eq 5. The performance metrics used for comparison of the obtained results were IGD-metric and SC-metric. Moreover, the experimental results are compared with the three best performers of CEC 2009 MOEA competition.

We can conclude the following points from our experiments

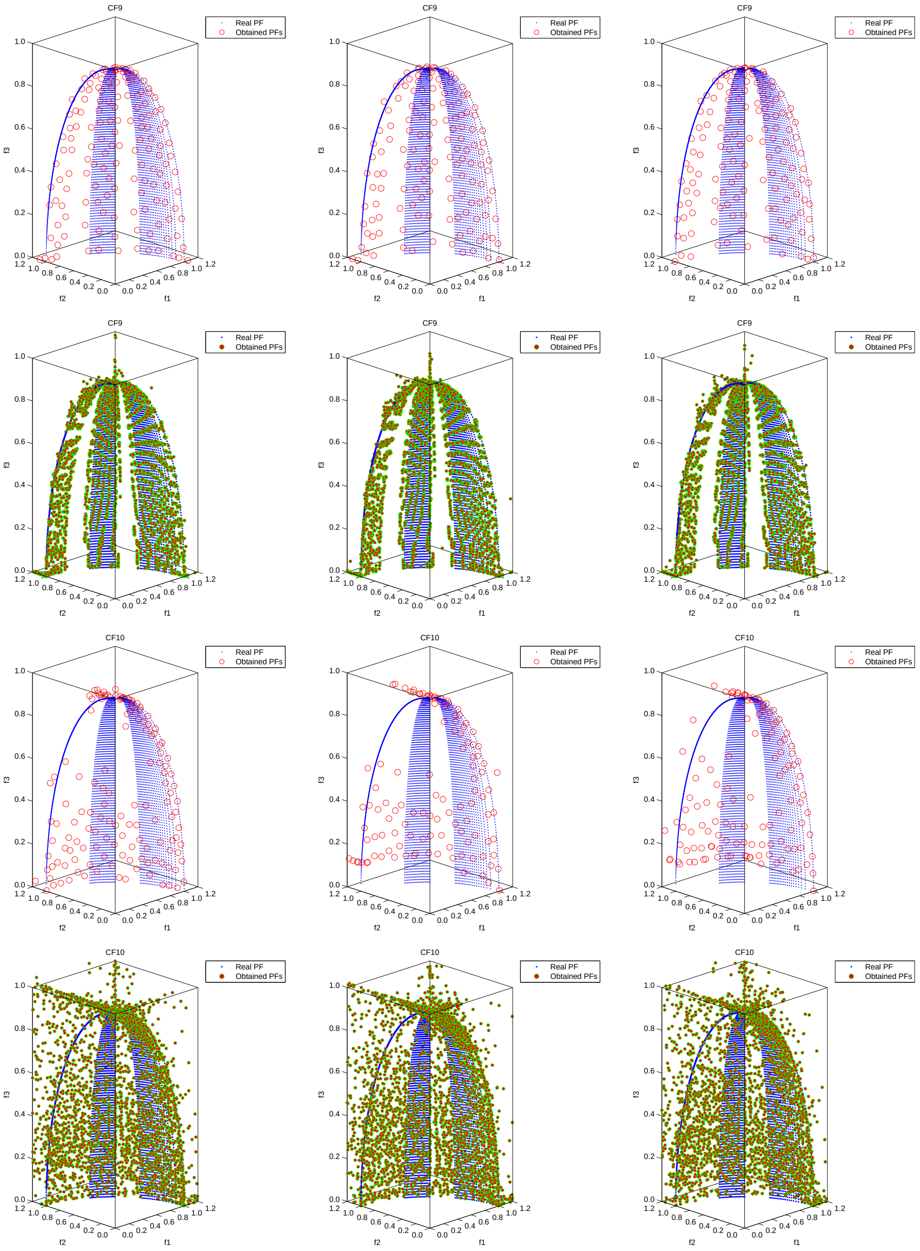


Fig. 5: Plots of the nondominated front with the best (lowest) IGD value and all the 30 final nondominated fronts found by the algorithm with Eq. 5 (left column), Eq. 7 (middle column), and Eq. 8 (right column) for CF9 and CF10.

conducted in this paper.

- CMOEA/D-DE-TDA with all three penalty functions works well even if there are few feasible solutions in the initial population.
- CMOEA/D-DE-TDA with the dynamic and adaptive penalty functions fails partly if there are few infeasible solutions in the initial population.
- The comparison of CMOEA/D-DE-TDA with the three best performers in CEC 2009 special session and competition indicated that CMOEA/D-DE-TDA has attained the best (i.e., lowest) IGD-metric value with Eq. 5 for test instance CF1, with Eq. 7 for test instance CF4, and with Eq. 8 for test instances CF8 and CF9. The algorithm has also found the second best values for three test instances CF2, CF6, and CF10. Specifically, our algorithm overall found better statistics for test instances CF1, CF8, and CF9.

References

[1] M. A. Jan, "Decomposition Based Evolutionary Methods for Constrained Multiobjective Optimization," Ph.D. dissertation, University of Essex, 2011.

[2] K. Miettinen, *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, 1999.

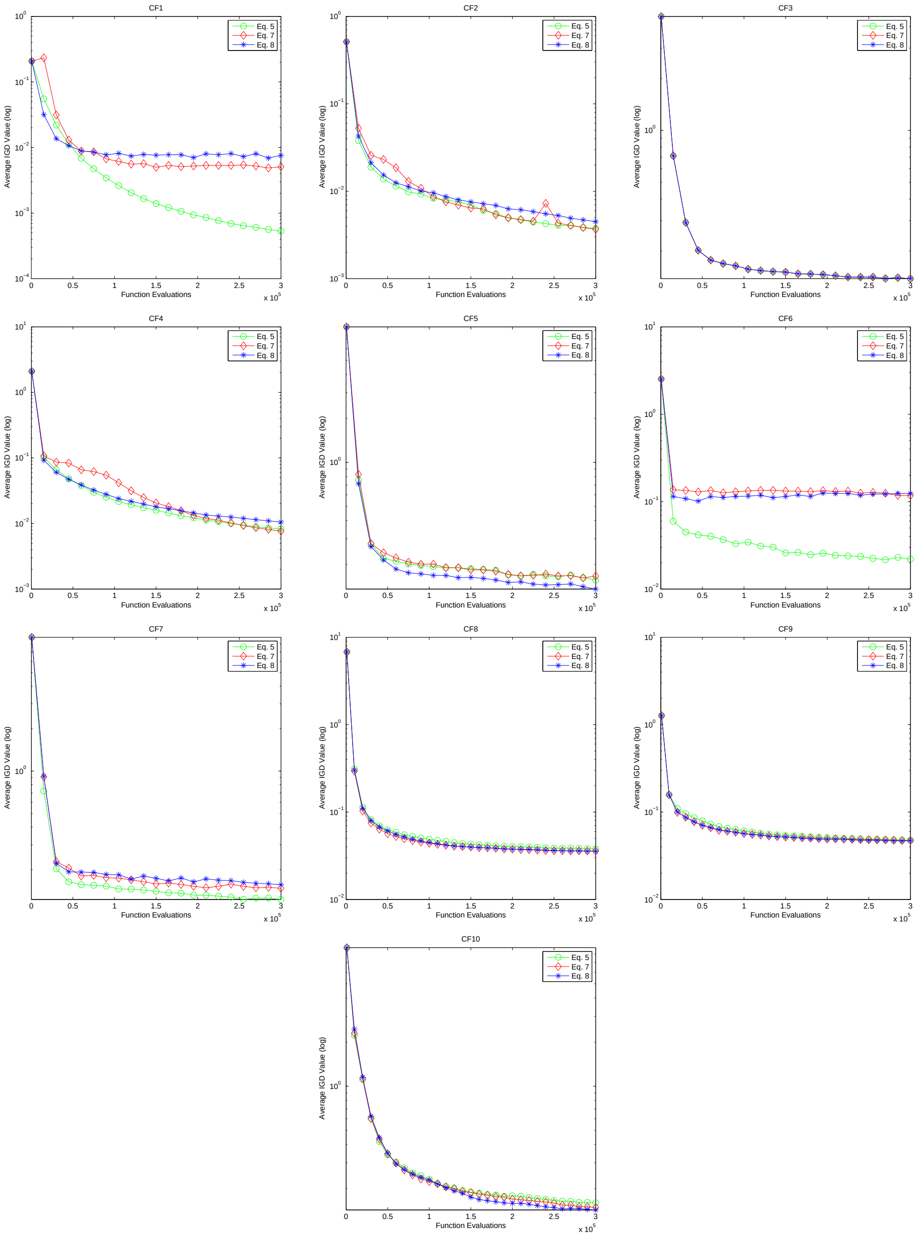


Fig. 6: Evolution of the IGD-metric values versus function evaluations when CMOEA/D-DE-TDA uses Eqs. 5, 7, and 8 for CF1-CF10.

[3] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons LTD, 2001.

[4] T. P.Runarsson and X. Yao, “Stochastic ranking for constrained evolutionary optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 3, pp. 284–294, September 2000.

[5] J. Joines and C. Houck, “On the use of nonstationary penalty functions to solve nonlinear constrained optimization problems with gas,” in *in Proc. 1st IEEE Conf. Evol. Program.*, E. D. Fogel, Ed., Orlando, FL, 1994, pp. 579–584.

[6] A. E. Smith and D. W. Coit, “Constraint handling techniquespenalty functions,” *In Back, D.B. Fogel.T Z.Michalewicz. eds. Handbook of Evolutionary Computation*, 1997.

[7] S. Kazarlis and V. Petridis, “Varying Fitness Functions in Genetic Algorithms: Studying the Rate of Increase of the Dynamic Penalty Terms,” in *Proceedings of the 5th Parallel Problem Solving from Nature (PPSN V)*, A. E. Eiben, T. Bäck, M. Schoenauer, and H.-P. Schwefel, Eds., Amsterdan, The Netherlands. Heidelberg, Germany: Springer-Verlag, September 1998, pp. 211–220, lecture Notes in Computer Science Vol. 1498.

[8] M. A. Jan and Q. Zhang, “MOEA/D for constrained multiobjective optimization: Some preliminary experimental results,” in *UK Workshop on Computational Intelligence (UKCI)*. IEEE, 2010, pp. 1–6.

[9] Q. Zhang and H. Li, “MOEA/D: A multiobjective evolutionary algorithm based on decomposition,” *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, 2007.

[10] H. Li and Q. Zhang, “Multiobjective Optimization Problems with Complicated Pareto Sets, MOEA/D and NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 284–302, April 2009.

[11] Q. Zhang, W. Liu, and H. Li, “The Performance of a New Version of MOEA/D on CEC09 Unconstrained Mop Test Instances,” in *Special*

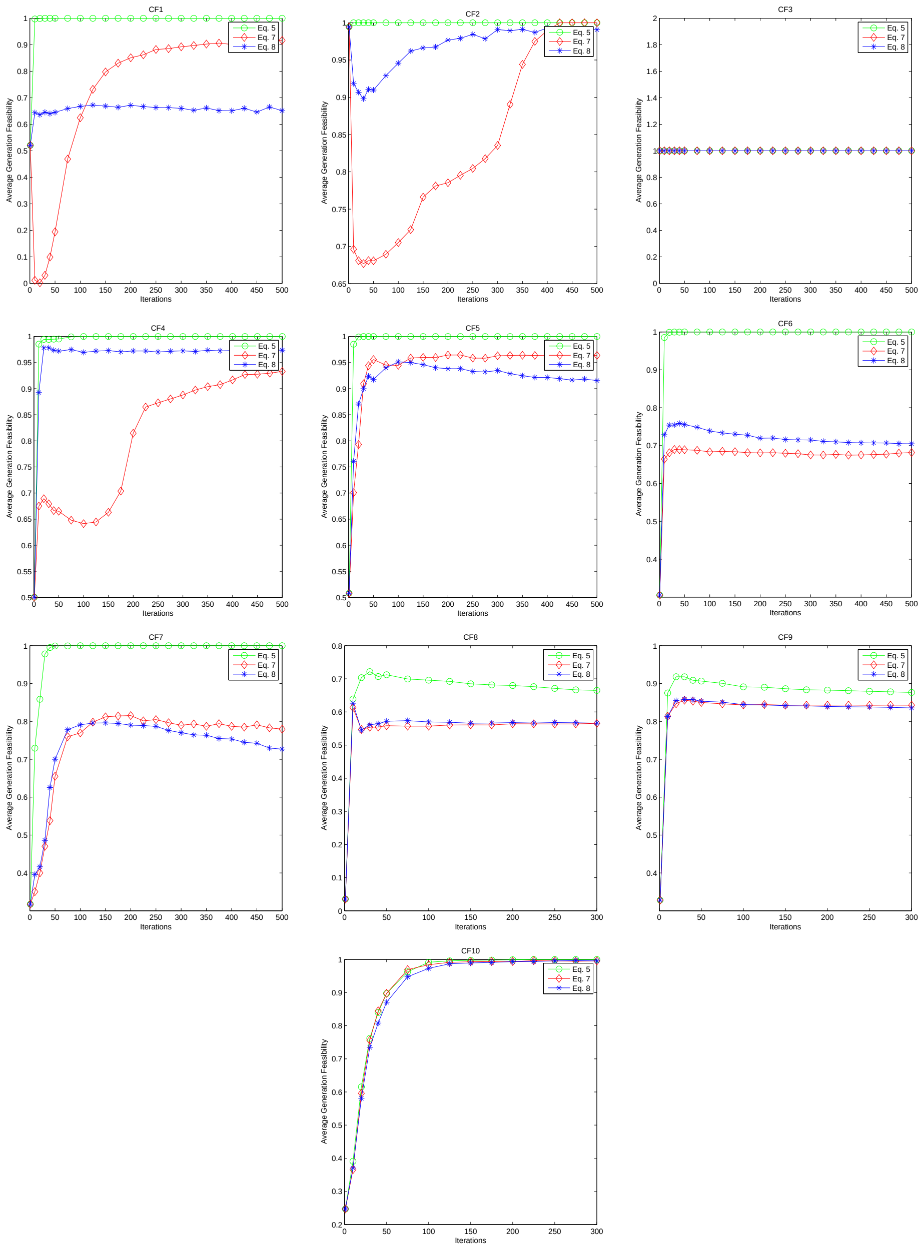


Fig. 7: Evolution of the generation feasibility versus generations when CMOEA/D-DE-TDA uses Eqs. 5, 7, and 8 for CF1-CF10.

TABLE III: COMPARISON BETWEEN CMOEA/D-DE-TDA WITH EQS. 5, 7, AND 8 (INDICATED BY JZ1, JZ2, JZ3, RESPECTIVELY), TSENG AND CHEN'S [12] (INDICATED BY TC), LIU AND LI'S [13] (INDICATED BY LL), AND LIU ET. AL'S [14] (INDICATED BY LI) ALGORITHMS IN TERMS OF THE IGD VALUES BASED ON 30 INDEPENDENT RUNS. THE RESULTS IN **BOLDFACE** INDICATE THE BEST VALUES AND THE RESULTS IN *ITALIC* INDICATE THE SECOND BEST RESULTS.

	best (lowest)						mean						st. dev.					
Test Instance	JZ1	JZ2	JZ3	TC	LL	LI	JZ1	JZ2	JZ3	TC	LL	LI	JZ1	JZ2	JZ3	TC	LL	LI
CF1	**0.0003**	0.0034	0.0035	0.0007	0.0071	*0.0005*	**0.0005**	0.0051	0.0076	0.0192	*0.0009*	0.0113	*0.0002*	0.0017	0.0027	0.0026	**0.0001**	0.0028
CF2	*0.0026*	0.0028	*0.0026*	0.0041	0.0027	**0.0016**	0.0038	*0.0037*	0.0045	0.0192	**0.0009**	0.0113	0.0014	*0.0011*	0.0020	0.0147	0.0026	**0.0005**
CF4	0.0056	**0.0054**	0.0062	0.0089	0.0090	*0.0055*	0.0083	*0.0076*	0.0106	0.0111	0.0142	**0.0070**	0.0018	0.0016	0.0041	**0.0014**	0.0033	*0.0015*
CF5	0.0326	0.0526	0.0328	*0.0176*	0.0588	**0.0079**	0.1560	0.1175	0.1356	*0.0208*	0.1097	**0.0158**	0.0932	0.0348	0.0921	**0.0024**	0.0307	*0.0067*
CF6	*0.0069*	0.0423	0.0387	0.0096	0.0090	**0.0062**	0.0220	0.1667	0.1241	0.0162	**0.0139**	*0.0150*	0.0160	0.1162	0.0303	*0.0060*	**0.0026**	0.0065
CF7	0.0374	0.0551	0.0556	*0.0187*	0.0535	**0.0104**	0.1227	0.1474	0.1567	*0.0247*	0.1045	**0.0191**	0.0787	0.0684	0.06590	**0.0047**	0.0351	*0.0061*
CF8	0.0344	*0.0336*	**0.0335**	0.6220	0.0473	0.0388	0.0380	*0.0359*	**0.0357**	1.0854	0.0607	0.0475	0.0017	*0.0016*	**0.0013**	0.2191	0.0130	0.0064
CF9	*0.0431*	0.0433	**0.0429**	0.0721	0.0460	0.1191	0.0475	*0.0473*	**0.0471**	0.0851	0.0505	0.1434	*0.0023*	**0.0022**	0.0025	0.0082	0.0034	0.0214
CF10	*0.1020*	0.1054	0.1062	0.1173	0.1055	**0.0984**	0.1602	0.1487	*0.1430*	**0.1376**	0.1974	0.1621	0.0464	0.0380	*0.0308*	**0.0092**	0.0760	0.0316

Session on Performance Assessment of Multiobjective Optimization Algorithms/CEC 09 MOEA Competition, Norway, 18-21 May 2009.

[12] L. Tseng and C. Chen, "Multiple trajectory search for unconstrained/constrained multi-objective optimization," in *IEEE Congress on Evolutionary Computation, CEC2009*. IEEE, 2009, pp. 1951–1958.

[13] H. Liu and X. Li, "The multiobjective evolutionary algorithm based on determined weight and sub-regional search," in *IEEE Congress on Evolutionary Computation, CEC2009*. IEEE, pp. 1928–1934.

[14] M. Liu, X. Zou, Y. Chen, and Z. Wu, "Performance assessment of DMOEA-DD with CEC 2009 MOEA competition test instances," in *IEEE Congress on Evolutionary Computation, CEC2009*. IEEE, 2009, pp. 2913–2918.

[15] K. M. Miettinen, *Nonlinear Multiobjective Optimization.* Kluwer Academic Publishers, 1999.

[16] M. A. Jan and R. A. Khanum, "A study of two penalty-parameterless constraint handling techniques in the framework of MOEA/D," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 128–148, 2013.

[17] Q. Zhang, A. Zhou, S. Zhao, P. N. Suganthan, W. Liu, and S. Tiwari, "Multiobjective optimization test instances for the CEC 2009 special session and competition," University of Essex and Nanyang Technological University, Tech. Rep. CES-487, 2008.

Regression-Based Feature Selection on Large Scale Human Activity Recognition

Hussein Mazaar
Faculty of Computers & Info.
Cairo University, Egypt

Eid Emary
Faculty of Computers & Info.
Cairo University, Egypt

Hoda Onsi
Faculty of Computers & Info.
Cairo University, Egypt

***Abstract*—In this paper, we present an approach for regression-based feature selection in human activity recognition. Due to high dimensional features in human activity recognition, the model may have over-fitting and can't learn parameters well. Moreover, the features are redundant or irrelevant. The goal is to select important discriminating features to recognize the human activities in videos. R-Squared regression criterion can identify the best features based on the ability of a feature to explain the variations in the target class. The features are significantly reduced, nearly by 99.33%, resulting in better classification accuracy. Support Vector Machine with a linear kernel is used to classify the activities. The experiments are tested on UCF50 dataset. The results show that the proposed model significantly outperforms state-of-the-art methods.**

Keywords*—Action Bank; Template Matching; SpatioTemporal Orientation Energy; Correlation; R-Squared; Support Vector Machine; Logistic Regression; Linear Regression; Human Activity Recognition

I. Introduction

Human activity recognition is an active research area in artificial intelligence, human-computer interaction and computer vision. Applications of human activities include patient monitoring systems, surveillance systems, interfaces, virtual reality, motion analysis, robot navigation, robot recognition, video indexing, browsing, choreography,..etc. Human activities are conceptually partitioned based on their complexity into four different categories: gestures, actions or activities, group activities and interactions. Nowadays, digital cameras can record the most daily activities of people and this makes the video sources to be rich on the internet, and also brings the problem of video categorization and how a new input video is classified based on their activities classes. Generally speaking, the process of classification of input videos movies in the real world is impossible, also, the manual task is time-consuming. Many researchers engage a lot of attention to these problems. They tried to create a machine recognition model which the feature descriptors originated from the training videos are trained to automatically recognize the activities of the new videos [1], [2], [3].

Feature selection is a significant step in human activity recognition to identify the minimum number of features that improve the accuracy of the model. Moreover, the models with the smallest number of features can be simpler and faster in building and understanding. In general, the main types of feature selection are filters, wrappers, and embedded machine learning. The last type selects the features based on integration with machine learning.

Filters methods depend on the properties of the data to evaluate the features and are independent regarding learning methods, but they use statistical methods like information gain, correlation to calculate splitting criterion for decision tree. These statistical methods evaluate how well each feature partitions dataset. Wrapper methods measure the features based on the estimates or results of machine learning algorithms which integrate predictive estimates as feedback. One of the common methods is regularization, which uses in the optimization process of learning in predictive modeling as penalization. This approach penalizes the irrelevant features(coefficients) and selects the most important features to reduce the complexity (over-fitting) like LASSO, Ridge regressions. Feature selection in embedded methods performs in the training process of machine learning. It is efficient because no need for splitting data into training and validation sets. Also the approach is fast due to the re-training of a feature is not necessary. Wrapper methods provide better results than filters, but the computational cost is increased. Embedded methods have good results between performance and cost [4], [5].

The organization of this paper is structured as follows. In Section II, we discuss related work. Section III presents the Model framework. Section IV presents the feature detection based on spatiotemporal orientation energy and the detected features are described based on maximum pooling of template matching. Section V presents the feature selection process which mainly based on the R-squared regression model. Support vector machine is introduced in VI. Section VII shows the simulation results and the conclusion of the paper is summarized in section VIII.

II. Related Works

At the present time, local spatiotemporal features are the most public techniques of video representation. The techniques of local spatiotemporal features depend on detectors and descriptors. The detectors capture spatiotemporal interest point locations, like, Cuboids [6] and Harris3D [7]. The descriptors are extracted by HOG3D [8] or HOG/HOF [9]. Then prelearned codebooks are defined to quantify the extracted features. Bag of Visual Words (BoVW) [10] can model videos. The local descriptors are local and repeatable features which are suitable advantages in video representation. They describe appearance and motion information of a local cuboid nearly interest point. Due to simplicity and repeatability, the local descriptors are robust to deformation and intra-class variability. The drawback of local descriptors that They only display low level information, not high level motion, which makes the features lack discriminative power. Many recent researchers try to fix the issues by developing high level models like Silhouette [11], Space-time Shape [12], Motion Energy and History Image [13]. The recent approach is Actionbank [14]. A large combination of activity detectors are applied on input videos and the responses are used as rich representation for videos. The detectors are composed of global templates of activities which are discriminating and global. However, the global features are sensitive to deformation and intra-class variations.

III. The Proposed Framework

The proposed model of human action recognition is composed of four steps: feature detection, feature description, feature selection and classification (See Fig. 1). For each step, the algorithms are described in details in the following sections.

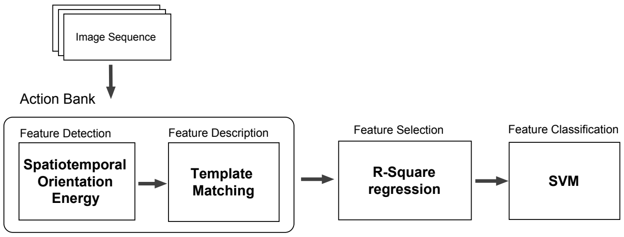


Fig. 1: The Proposed Framework of Human Action Recognition

IV. Feature Detection and Description

The videos are showed via high level features. The Action Bank [14] is the representation of videos. It is similarly close to object bank [15]. It represents the video as composed action detectors that each produces a correlation volume. The base element of feature is the template-based action detector. It is invariant/robust to variations in appearance, scale, viewpoint,and tempo.

A. Spatiotemporal Orientation Energy

Motion energies can represent an activity or video in various spationtemporal orientation. A composition of energies along various space-time orientations can capture the motion at a point during decomposition of video. These energies are the basis for low level activity representation. A decomposition of spatiotemporal orientation energies is performed using third derivatives of 3D Gaussian steerable filter which represents the strength of motion and used as local filter. Let $G^3_{\hat{\theta}}(x)$ denotes 3D Gaussian third derivatives, where $x = (x, y, t)$ indicates for location of spatiotemporal space and $\hat{\theta}$ denotes for unit vector of 3D directions. The spatiotemporal orientation energy is computed at every pixel as follows:

$$E_{\hat{\theta}}(x) = \sum_{x' \in \Omega(x)} (G^3_{\hat{\theta}} * V)^2 \tag{1}$$

where $\Omega(x)$ denotes for a local region around x, $V \equiv V(x)$ denotes for input video, and (*) indicates for convolution. Gaussian filters are separable filter that has some properties like estimation spatiotemporal orientation energy without executing convolution for all directions. The result of convolution is summed and squared over neighborhood space time Ω to get the energy measurement.

Marginalization for energy is a process to eliminate spatial orientation influence. Formally, the computation of energy with normal n̂ at frequency domain plane $E_{\hat{\theta}_i}(\hat{n})$ by a simple sum

$$E'_{\hat{n}}(x) = \sum_{i=0}^{N} E_{\hat{\theta}_i \hat{n}}(x) \tag{2}$$

where N denotes for is Gaussian derivatives order, $\hat{\theta}_i$ is one of $N + 1 = 4$ directions calculated from Eqn. 2.

Officially $\hat{\theta}_i$ is provided by,

$$\hat{\theta}_i = cos\left(\frac{\pi i}{4}\right) \hat{\theta}_a(\hat{n}) + sin\left(\frac{\pi i}{4}\right) \hat{\theta}_b(\hat{n}), \tag{3}$$

where $\hat{\theta}_a(\hat{n}) = \hat{n} \times \hat{e}_x / \|\hat{n} \times \hat{e}_x\|$, $\hat{\theta}_b(\hat{n}) = \hat{n} \times \hat{\theta}_a(\hat{n})$,ê is the unit vector along the spatial x axis in the Fourier domain and $0 \leq i \leq 3$. The implementation for detectors of action bank, seven raw spatiotemporal energies are defined with different velocities:static E_s, leftward E_l, rightward E_r, upward E_u, downward E_d, flicker E_f , and lack of structure E_o. The lack of structure energy is calculated as function of six other energies and has peaks when no strong response from other six energies. The goal of this energy is to eliminate the instabilities of small energy points and gets a saliency. The pure energies are extracted from energies with subtraction of background and noise and are normalized to avoid influence of illumination adjustment and contrast as follows:

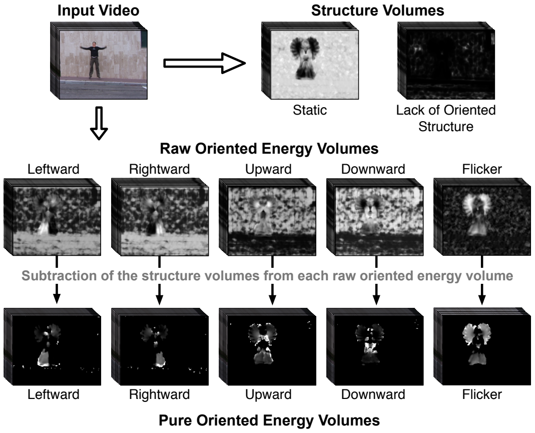

Input Video
Structure Volumes
Static
Lack of Oriented Structure
Raw Oriented Energy Volumes
Leftward
Rightward
Upward
Downward
Flicker
Subtraction of the structure volumes from each raw oriented energy volume
Leftward
Rightward
Upward
Downward
Flicker
Pure Oriented Energy Volumes

Fig. 2: A spatiotemporal orientation energy representation [14]

$$\hat{E}_i = max(E_i - E_o - E_s, 0), \; \forall i \in \{f, l, r, u, d\} \tag{4}$$

B. Template Matching

Detection an activity of small video called ”template video” in a large video called ”search video” is performed by scanning a 3D template video over all positions in spacetime. The similarity is determined by calculating each location among histogram of oriented energy of the template and search video. The ”action spotting” algorithm is the recent detector which is applied due to appropriate features of in-variance to activity localization, appearance variation, natural explanation like the decompose oriented energies and efficiency [16], [14]. The correlation between template video T and search video or query video is calculated by Bhattacharya coefficient m(.) as follows:

$$M(x) = \sum_{u} m(T(u), V(x-u)) \tag{5}$$

where M() denotes for the results of correlation and u denotes for ranges of template video. The correlation is efficiently performed in frequency domain and the output value is between 1 denoting full match or complete match and 0 denoting a complete mismatch which interprets volumetric max-pooling method.

Let N_a denotes for number of detectors for a given action bank and N_s denotes for scales of activity (run times), the output of correlation volumes are $N_a \times N_s$. The max-pooling technique in [17] is adapted as in Fig. 3 to be three levels in the octree which is $1^3+2^3+4^3$ or a 73 dimension vector [14]. For each activity, the total length of feature vector equals to $N_a \times N_s \times 73$.

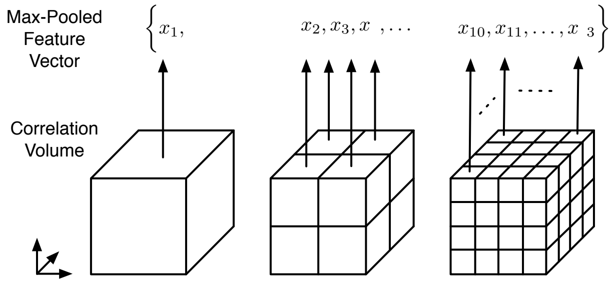


Fig. 3: Volumetric max-pooling technique [14]

V. Feature Selection

Feature selection is an important area in predictive modeling and statistics. Theory and practice of feature selection have shown that feature selection is an effective way in improving learning, enhancing recognition accuracy and decreasing complexity of human activity recognition. The objective of feature selection in supervised learning produces higher classification accuracy [18], [19], [20].

One of the most crucial issues in high-dimensional data is determining which features should be included in a model of human activity recognition. From a practical point of view, a model with less features may be more interpretative and less complexity. Statistically speaking, the model with less features is often more attractive. Also, some models are negatively affected by irrelevant features [21], [20].

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. It uses a forward stepwise least squares regression that maximizes the model R-squared value. In this model, the features assessment are fast provided as a preparatory step and the predictive models are rapidly simplified in development with huge data. Linear models can quickly identify input useful features for classifying the target classes. The R-Squared feature selection criterion has applied two steps processes as follow:

A. Squared Correlations

The squared correlation coefficient is the ratio of single input feature explains the variation in target class with elimination of other features in calculations. Also, It is called Coefficient of Determination (CoD) in statistics. The value ranges of squared correlation coefficient are between 0 (no relationship between the target class and input feature) and 1 (the variation of target class is totally explained with input feature). In human activity recognition, all input features are interval, so the squared correlation coefficient is calculated by a simple linear regression as follow:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{6}$$

where Y denotes response variable or target, X denotes for input feature, β_0 denotes for intercept parameter, β_1 denotes for slope parameter and ε indicates the error deviation of Y about $\beta_0 + \beta_1 X$ (See Fig. 4a).

The feature has a significant influence if it explains the target, so the simple linear regression model is compared to the baseline model (Fig. 4b). The baseline regression has a horizontal fitted regression line over any value in input feature with slope equals to 0 and the intercept equals to the mean of response target $\bar{Y}$.

Explained variability is the distinction between the regression line and baseline line. The regression sum of squares (SSR) is the amount of variability explained by your model. The comparison between the explained variability to unexplained variability determines the amount of variability explained by regression line rather than baseline line. The Fig. 4c shows a seemingly contradictory relationship between explained, unexplained and total variability. The regression sum of squares (SSR) is equal to

$$\sum (\hat{Y}_i - \bar{Y})^2 \tag{7}$$

Unexplained variability is the distinction between the between the actual values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by regression model. The error sum of squares is equal to

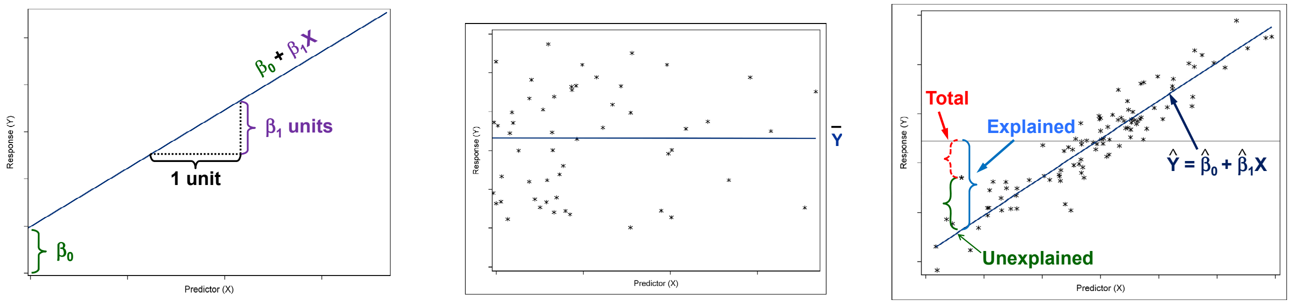
$$\sum (Y_i - \hat{Y}_i)^2 \tag{8}$$

Total variability is the distinction between the actual values and baseline regression line. The corrected total sum of squares (SST) is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to

$$\sum (Y_i - \bar{Y})^2 \tag{9}$$

R-Squared the proportion of variability observed in the data explained by the regression line. The R-Squared is equal to

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{10}$$



(a) Simple Linear Regression

(b) Baseline Regression

(c) Expained vs. Unexplained Variabilty

Fig. 4: Regression Model

B. Forward Stepwise Regression & Logistic Regression

This algorithm is applied after calculating the squared correlation coefficient for all input features in human activity recognition, the other important features are measured using a forward stepwise R-Squared regression. The sequential forward regression chooses The feature that has the highest squared correlation coefficient which explains the largest amount of variation in class target. At each iteration, the additional input feature is selected that gives the largest incremental increase in model of R-Squared. The stepwise algorithm ends when no other input feature can meet the Stop R-Squared criterion. The final logistic regression analysis is performed using the predicted values that are output from the forward stepwise selection as the independent input.

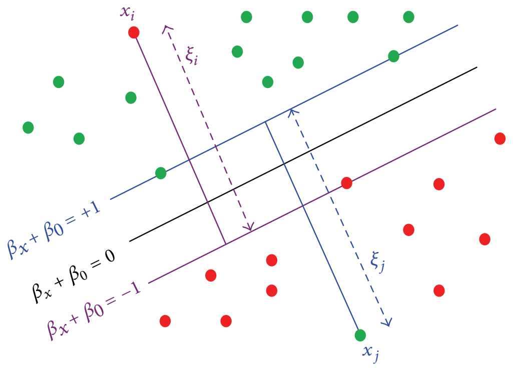


Fig. 5: Support Vector Machine with Slack Variables

VI. Support Vector Machine

The concluding stage of the recognition process is the classification of the extracted features into a predefined set of classes. The field of machine learning has many powerful classification models. Our goal in this stage is to contribute to this field by introducing a reliable, accurate and interaction-centric classifier.

The human activities recognition are formulated by multicalss classification problem. Each activity is represented by each class. The goal is assigning and classifying a video sequence to classes of activities. Many supervised learning methods are learned to activity recognizer. Support Vector Machine (SVM) is one of the superior machine learning in human activity recognition and high dimensional data because the prime generalization strength and highly accurate results. SVM can avoid over-fitting in neural networks based on risk minimization theory. Also, SVM can handle a high dimensional space by creating a maximal hyperplane to separate non-overlapping classes. Two parallel hyperplanes are proceeded in SVM and the goal of SVM seeks to find the maximal distance between the parallel byperplanes (Fig. 5). The better the classification, the larger the distance between byperplanes and vice vera.

Formally, Let the data set of training is $\mathbf{D} = \{\{x_i, y_i\}_{i=1}^{n} \mid x_i \in \Re^d, y_i \in \{-1, +1\}\}$ with n observations in a d-dimensional space and y_i denotes for classes, SVM can handle non-separable observation by slack variable ξ_i for observation x_i which indicates how much the observation violates the soft margin constraints. The values of slack variable have three type: $\xi_i = 0$ denotes the observation away with at least $\frac{1}{\|W\|}$ from the hyperplane, $0 \leq \xi_i \leq 1$ denotes the observation between margins and when $\xi_i \geq 1$ then the observation is wrongly classified and appears on the wrong side. This approach achieves best performance for SVM. The quadratic programming can determine the optimal generalized separating hyperplane as follow:

$$\underset{w,b,\xi_i}{\arg\min} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(\xi_i)^k \tag{11}$$

Subject to $y_i(w^T x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0 \; \forall x_i \in \mathbf{D}$.

The parameter C is a constant called ”regularization constant” to control the misclassification cost which governs the trade-off among maximal margins and minimal loss. The term $\sum_{i=1}^{n}(\xi_i)^k$ denotes for loss. The constant k controls the loss which becomes hinge loss when k is 1 and quadratic loss when k is 2. Dual formulation is recommended to solve SVM due to computational purposes. This solution uses Lagrangian method and is optimized with Lagrange multiplier α. The weight vector for predicting decision is $\beta = \sum_i \alpha_i x_i y_i; 0 \leq \alpha_i \leq C$. The instances $x_i \; with \; \alpha_i > 0$ are called support vectors, as they uniquely define the maximum margin hyperplane.

VII. Simulation Results

The experiments are conducted using UCF50 action dataset [22]. UCF50 is an activity recognition data set with 50 activities classes, composing of real Youtube videos. The large variations in cluttered background, camera motion, object scale, object appearance and pose, illumination conditions and viewpoint make the dataset to be very challenging. The total videos in UCF50 are 6680. The videos in UCF50 are grouped into 25 groups. For each group, the video clips have similar features, such as the same person, similar viewpoint, similar background, and so on. The classes or activities are visually shown in Fig. 6. The experiments are implemented on computer with CPU i7, 2.6 GHz, 16 RAM, Matlab 2013b and R-Studio. Initially speaking, The features in UCF50 dataset are extracted using the spatiotemporal orientation energy, then the extracted values are described in vectors using template matching as action bank. The length of feature vector is 14746 and the number of observations is 6680. The R-Squared model is implemented to select the features that describe the variations in target. The features that explain the target class are selected and the other features are redundant or irrelevant. The minimum R-squared in our implementation is 0.005. It specifies the lower bound for the individual R-square value of a feature in order to be eligible for the model selection process. The number of selected features for each action is described in Fig. 7. The average number of features using R-Squared is 99 which is 0.67% from the original data. About 99.33% of features can't improve the performance of the model, but these features degrade negatively the recognition due to the large number of features which are redundant or irrelevant. The irrelevant features can make an over-fitting in the model.

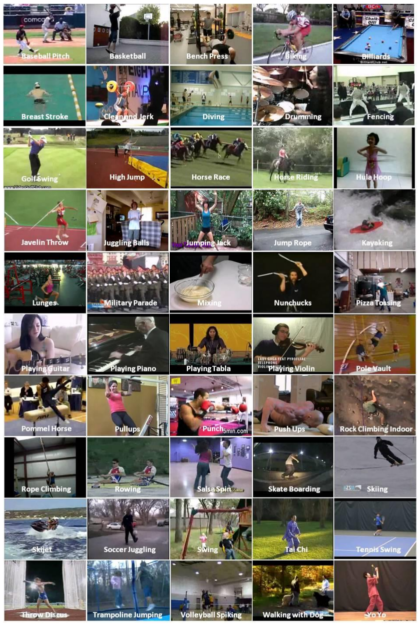


Fig. 6: UCF50 Dataset

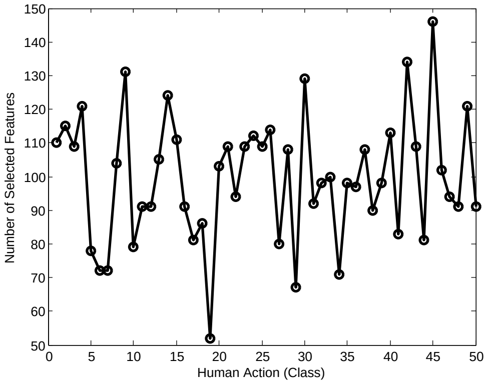


Fig. 7: Number of Selected Features using R-Squared Feature Selection for each Action

TABLE I: Sorted Accuracies score per class(action) in UCF50 dataset

2/3 train, 1/3 test		5-fold video-wise cross validation		5-fold group-wise cross validation	
Billiards	0.99	JumpingJack	0.9887	BreastStroke	0.956
PushUps	0.987	Billiards	0.98	Billiards	0.95
JumpRope	0.971	BreastStroke	0.965	CleanAndJerk	0.923
Biking	0.97	PlayingViolin	0.965	HighJump	0.891
BreastStroke	0.97	CleanAndJerk	0.964	JumpingJack	0.89
BaseballPitch	0.967	JugglingBalls	0.9549	PommelHorse	0.887
JumpingJack	0.965	PushUps	0.9528	PushUps	0.885
Mixing	0.96	Mixing	0.95	PlayingGuitar	0.881
PlayingViolin	0.958	PommelHorse	0.943	ThrowDiscus	0.877
YoYo	0.945	BaseballPitch	0.94	GolfSwing	0.875
Swing	0.938	MilitaryParade	0.937	PlayingPiano	0.873
Drumming	0.937	SalsaSpin	0.936	Mixing	0.872
RockClimbingIndoor	0.936	HighJump	0.9349	JumpRope	0.871
Fencing	0.935	PullUps	0.9333	MilitaryParade	0.864
PlayingPiano	0.933	Rowing	0.93	HorseRiding	0.863
HulaHoop	0.9303	Kayaking	0.9299	TaiChi	0.86
PommelHorse	0.93	GolfSwing	0.9295	BaseballPitch	0.857
PullUps	0.928	Nunchucks	0.9233	Fencing	0.854
HorseRace	0.925	RopeClimbing	0.9192	HorseRace	0.853
VolleyballSpiking	0.925	PlayingPiano	0.919	SkateBoarding	0.846
MilitaryParade	0.921	JumpRope	0.9189	BenchPress	0.845
Diving	0.92	RockClimbingIndoor	0.9189	PlayingViolin	0.845
Lunges	0.919	PlayingGuitar	0.9156	Skijet	0.84
HighJump	0.918	Diving	0.915	VolleyballSpiking	0.836
Kayaking	0.917	JavelinThrow	0.9102	PoleVault	0.835
Rowing	0.917	HorseRace	0.9094	Diving	0.831
JugglingBalls	0.916	Fencing	0.909	Punch	0.831
BenchPress	0.91	Biking	0.9069	RockClimbingIndoor	0.83
Skiing	0.91	Punch	0.906	SalsaSpin	0.827
Punch	0.909	HorseRiding	0.9056	Biking	0.823
HorseRiding	0.904	VolleyballSpiking	0.905	JugglingBalls	0.821
SalsaSpin	0.903	Swing	0.894	YoYo	0.818
ThrowDiscus	0.902	Drumming	0.891	JavelinThrow	0.813
CleanAndJerk	0.9	Lunges	0.89	Swing	0.807
RopeClimbing	0.9	Skijet	0.89	Basketball	0.789
GolfSwing	0.898	ThrowDiscus	0.8816	Drumming	0.781
JavelinThrow	0.891	SkateBoarding	0.879	PlayingTabla	0.781
Nunchucks	0.883	BenchPress	0.875	WalkingWithDog	0.778
SkateBoarding	0.88	TaiChi	0.875	Rowing	0.774
TrampolineJumping	0.877	Basketball	0.8723	PullUps	0.765
TaiChi	0.872	PoleVault	0.8687	Lunges	0.763
PlayingGuitar	0.863	PlayingTabla	0.8669	SoccerJuggling	0.756
PlayingTabla	0.861	Skiing	0.864	Nunchucks	0.753
Basketball	0.855	YoYo	0.859	RopeClimbing	0.753
SoccerJuggling	0.851	HulaHoop	0.84	HulaHoop	0.742
Skijet	0.843	PizzaTossing	0.8377	TennisSwing	0.739
PizzaTossing	0.836	SoccerJuggling	0.83	Kayaking	0.736
TennisSwing	0.802	TennisSwing	0.826	PizzaTossing	0.731
PoleVault	0.799	WalkingWithDog	0.796	Skiing	0.731
WalkingWithDog	0.742	TrampolineJumping	0.794	TrampolineJumping	0.721

The UCF50 features data are evaluated using 5-fold group-wise cross-validation, 5-fold video-wise cross-validation and $\frac{1}{3}(34\%)$ testing data. In our model, One-vs-rest SVM is applied to classify the actions using Linear kernel. The penalty is 1 and the maximum iterations is 25. For each action, positive video clips are labeled as 1 and negative videos are as labeled -1. For each action, R-Squared and SVM are applied. The accuracies are sorted for each action using 5-fold group-wise cross-validation, 5-fold video-wise cross-validation and 1/3 testing data in Table I. The accuracies are visually shown in Fig. 8.

The overall accuracy using our approach is 82.64% for 5-fold group-wise cross-validation, 90.49% for 5-fold video-wise cross-validation and 90.8% for 34% testing data. The comparisons to available related works are described in Table II.

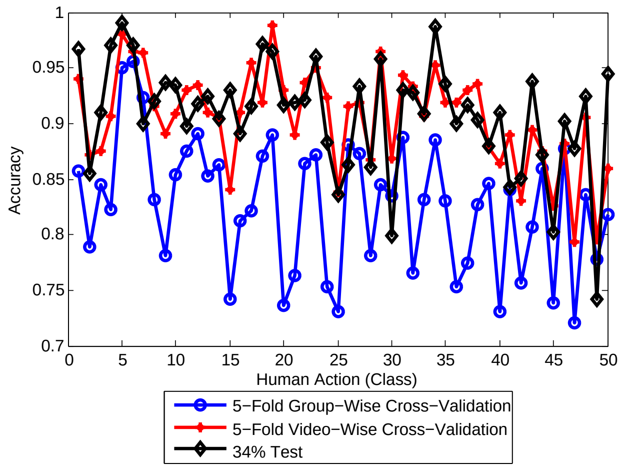


Fig. 8: Accuracy using R-Squared Feature Selection for each Human Action

TABLE II: Comparison with the Literature Results on UCF50 Dataset

Author	Experimental Setup	Accuracy
Our Method	5-fold group-wise cross validation	82.64%
Our Method	5-fold video-wise cross validation	90.49%
Our Method	2/3 training and 1/3 testing for each class	90.8%
Reddy and Shah [22]	Leave One Group Out Cross validation (25 cross-validations)	76.9%
Sadanand and Corso [14]	video-wise cross validation	76.4%
Sadanand and Corso [14]	group-wise cross validation	57.90%
Todorovic [23]	2/3 training and 1/3 testing for each class	81.03%
Solmaz et al. [24]	Leave One Group Out Cross validation(25 cross-validations)	73.70%
Kliper-Gross et al. [25]	Leave One Group Out Cross validation (25 cross-validations)	72.60%

VIII. Conclusions

Human activity recognition based on spatiotemporal orientation energy and activity template is simple and advanced discrimination techniques in detection and extraction features based on multiple activity detectors. The features in human activity recognition often more than the number of observations, so the feature selection is a major step before classification to avoid irrelevant or redundant features and over-fitting problems. R-Squared model is applied to get the best important discriminative features that explain the target. Also, R-Squared can handle a huge data in rapidly simplified manner. The model can significantly improve the performance/accuracy of human activities and reduce the features.

In the future, We will plan to apply the regression-based feature selection in human activity recognition based on different feature extraction methods that have large amount of features.

References

[1] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, April 2011.

[2] R. Gao, "Dynamic feature description in human action recognition," Master's thesis, Leiden Institute of Advanced Computer Science, Leiden University, 2009.

[3] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[4] A. G. Janecek, W. N. Gansterer, M. A. Demel, and G. F. Ecker, "On the relationship between feature selection and classification accuracy," in *JMLR: Workshop and Conference Proceedings*, pp. 90–105, 2008.

[5] E. Tuv, A. Borisov, G. Runger, K. Torkkola, I. Guyon, and A. R. Saffari, "Feature selection with ensembles, artificial variables, and redundancy elimination," *JMLR*, 2009.

[6] P. Doll'ar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS 05)*, pp. 65 – 72, October 2005.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space time shapes," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 1395 – 1402, 2005.

[8] A. Klaser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference (BMVC)*, pp. 995 – 1004, 2008.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1 – 8, 2008.

[10] X. Wang, L. Wang, and Y. Qiao, "A comparative study of encoding, pooling and normalization methods for action recognition," in *Computer Vision ACCV 2012* (K. Lee, Y. Matsushita, J. Rehg, and Z. Hu, eds.), vol. 7726 of *Lecture Notes in Computer Science*, pp. 572–585, Springer Berlin Heidelberg, 2013.

[11] K. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I–77–I–84 vol.1, June 2003.

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 29, pp. 2247 – 2253, Dec 2007.

[13] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates,," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[14] S. Sadanand and J. Corso, "Action bank: A high-level representation of activity in video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1234–1241, June 2012.

[15] L. Li, H. Su, Y. Lim, and F. Li, "Object bank: An object-level image representation for high-level visual recognition," *International Journal of Computer Vision*, vol. 107, no. 1, pp. 20–39, 2014.

[16] K. Derpanis, M. Sizintsev, K. Cannons, and R. Wildes, "Efficient action spotting based on a spacetime oriented structure representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1990–1997, June 2010.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, 2006.

[18] M. Ramaswami and R. Bhaskaran, "A study on feature selection techniques in educational data mining," *Journal of Computing*, vol. 1, 2009.

[19] S. Garca, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Springer International Publishing, 2015.

[20] H. Mazaar, E. Emary, and H. Onsi, "Evaluation of feature selection on human activity recognition," in *IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS15)*, pp. 105–113, 2015.

[21] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer-Verlag New York, 2013.

[22] K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[23] S. Todorovic, "Human activities as stochastic kronecker graphs," in *Computer Vision ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona,

Y. Sato, and C. Schmid, eds.), vol. 7573 of *Lecture Notes in Computer Science*, pp. 130–143, Springer Berlin Heidelberg, 2012.

[24] B. Solmaz, S. Assari, and M. Shah, “Classifying web videos using a global video descriptor,” *Machine Vision and Applications*, vol. 24, no. 7, pp. 1473–1485, 2013.

[25] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, “Motion interchange patterns for action recognition in unconstrained videos,” in *European Conference on Computer Vision (ECCV)*, Oct. 2012.

Reflected Adaptive Differential Evolution with Two External Archives for Large-Scale Global Optimization

Rashida Adeeb Khanum
Jinnah College for Women
University of Peshawar
Khyber Pakhtunkhwa, Pakistan.
Email: adeeb_maths@yahoo.com

Nasser Tairan
College of Computer Science,
King Khalid University
Abha, Saudi Arabia
Email: nmtairan@kku.edu.sa

Muhammad Asif Jan
Department of Mathematics
Kohat University of Science & Technology
Khyber Pakhtunkhwa, Pakistan
Email: majan@kust.edu.pk

Wali Khan Mashwani
Department of Mathematics
Kohat University of Science & Technology
Khyber Pakhtunkhwa, Pakistan
Email: mashwanigr8@gmail.com

Abdel Salhi
Department of Mathematical Sciences
University of Essex
Colchester, CO4 3SQ, U.K.
Email: as@essex.ac.uk

***Abstract*—JADE is an adaptive scheme of nature inspired algorithm, Differential Evolution (DE). It performed considerably improved on a set of well-studied benchmark test problems. In this paper, we evaluate the performance of new JADE with two external archives to deal with unconstrained continuous large-scale global optimization problems labeled as Reflected Adaptive Differential Evolution with Two External Archives (RJADE/TA). The only archive of JADE stores failed solutions. In contrast, the proposed second archive stores superior solutions at regular intervals of the optimization process to avoid premature convergence towards local optima. The superior solutions which are sent to the archive are reflected by new potential solutions. At the end of the search process, the best solution is selected from the second archive and the current population. The performance of RJADE/TA algorithm is then extensively evaluated on two test beds. At first on 28 latest benchmark functions constructed for the 2013 Congress on Evolutionary Computation special session. Secondly on ten benchmark problems from CEC2010 Special Session and Competition on Large-Scale Global Optimization. Experimental results demonstrated a very competitive performance of the algorithm.**

***Keywords*—Adaptive differential evolution; large scale global optimization; archives.**

I. INTRODUCTION

Optimization deals with finding the optimal solution for single or multi-objective functions [1]. An unconstrained single objective optimization problem can be stated as follows:

$$\text{Minimize } f(\mathbf{x}), \tag{1}$$

where $f(\mathbf{x})$ denotes the objective function, and $\mathbf{x} = (x_1, x_2, ..., x_n)^T$ is an n-dimensional real vector.

DE [2] is a most popular bio-inspired scheme for finding the global optimum $\mathbf{x}^*$ of problem (1). The heuristic is essentially an evolutionary one and relies on the usual genetic operators of mutation and crossover. DE is easy to understand and implement, has a few parameters to control, and is robust. There is no doubt that DE is a remarkable optimizer for many optimization problems. However, it has few drawbacks like, stagnation, premature convergence, and loss of diversity. Since it is a global optimizer, so its local search ability is not that good. More details can be found in [3].

To enhance the performance of DE, many modifications to the classic DE have been suggested and various variants of DE are proposed. A novel work is done by Wang et al. [4], in which they utilized orthogonal crossover instead of binomial and exponential crossover. A group of researchers have introduced new variants like opposition based DE [5], centroid dependent initialization ciJADE [6], cluster-based population initialization (CBPI) [7] jDE [8], genDE [9], Individual dependent Mechanism (IDE) [10] etc. Control parameters adaptation and self-adaptation have devised in [11], [12], jDErpo [13] SaDE [14], JADE [15], [16], EPSDE [17], IDE [18], SHADE [19]L-SHADE [20] [21], EWMA-DECrF [22]. Cooperative coevolution have been brought into DE for large scale optimization [23]. Some researchers applied it to problems from the discrete domain [24], [25], while others are taking the advantage of its global searching in the continuous domains [4], [26]–[28].

In another experiment, adaptive variant of DE, the so-called JADE [15], is proposed for numerical optimization. It has shown performance improvement over the state-of-the-art algorithms, jDE [8], SaDE [29] and DE/rand/1/bin [2] according to the reported results in [15] and [30]. However, JADE is not reliable; on some problems. For instance, it finds the global optima in some runs, but it can also be trapped in local optima [30]. To improve the reliability of JADE, in this paper, we introduce two new strategies in JADE and thus propose Reflected Adaptive Differential Evolution with Two External Archives (RJADE/TA).

The rest of this paper is organized as follows. Section II describes the basic DE and JADE algorithm. Section III presents proposed RJADE/TA. Section IV gives the experimental results and finally Section V concludes this paper and discusses future research directions.

II. Differential Evolution and JADE

A. Differential Evolution

The four main schemes of differential evolution (DE) are detailed as follows.

1) Parent Selection: For each member $\mathbf{x}_i$, $i = 1, 2, ..., N_p$, of the current generation G three other members, $\mathbf{x}_{r_1}$, $\mathbf{x}_{r_2}$ and $\mathbf{x}_{r_3}$ are randomly selected, where r_1, r_2 and r_3 are randomly chosen indices such that r_1, r_2 and $r_3 \in \{1, 2, ..., N_p\}$ and $i \neq r_1 \neq r_2 \neq r_3$. Thus, for each individual, $\mathbf{x}_i$, a mating pool of four individuals is formed in which it breeds against three individuals and produces an offspring.

2) Mutation: After selection, mutation is applied to produce a mutant vector $\mathbf{v}_i$, by adding a scaled difference of the two already chosen vectors to the third chosen vector. i.e.,

$$\mathbf{v}_i = \mathbf{x}_{r_1} + F(\mathbf{x}_{r_2} - \mathbf{x}_{r_3}), \tag{2}$$

where $F \in (0, 2)$ [31] is the scaling factor.

3) Crossover: After mutation, the parameters of the parent vector $\mathbf{x}_i$ and mutant vector $\mathbf{v}_i$ are mixed by a crossover operator and a trial member $\mathbf{u}_i$ is generated as follows:

$$u_{i,j} = \begin{cases} v_{i,j} & \text{if } rand_j(0,1) \leq CR; \\ x_i & \text{otherwise,} \end{cases} \tag{3}$$

where $j \in \{1, 2, ..., n\}$.

4) Survivor Selection: At the end, the trial vector generated in (3) is compared with its parent on the basis of its objective function value. The fittest will propagate to the next generation. i.e.,

$$\mathbf{x}_{i+1} = \begin{cases} \mathbf{u}_i, & \text{if } f(\mathbf{u}_i) \leq f(\mathbf{x}_i); \\ \mathbf{x}_i, & \text{otherwise.} \end{cases} \tag{4}$$

B. JADE

Before presenting the new algorithm, we give the details of the DE's version JADE, upon which the devised algorithm in this paper is based. JADE [15] is an adaptive version of DE. It improves the performance of DE, by implementing a new mutation strategy DE/current-to-p best with/without external archive, and adaptively controlling the parameters F and CR. JADE adopts the crossover and selection scheme of classic DE as described in Equation (3) and Equation (4). DE/current-to-$p$best strategy incorporates not only the best solution information, but also the information of other good solutions. Specifically, any solution from the top $p\%$ population can be randomly selected in DE/current-to-p best to play the role of the single best solution in DE/current-to-best [15]. Where p is the percentage of top good solutions and the default value for it is 5% of N_p. Other suggested values of p are between 5% and 20%, inclusive. JADE modifies classic DE in three aspects.

1) DE/current/to-$p$best strategy: JADE utilizes two mutation strategies, one with external archive, and the other without it. These strategies are the improvement of DE/current-to-best/1 strategy. They can be expressed as follows [15]:

$$\begin{aligned} \mathbf{v}_i &= \mathbf{x}_i + F_i(\mathbf{x}^p_{best} - \mathbf{x}_i) + F_i(\mathbf{x}_{r_1} - \tilde{\mathbf{x}}_{r_2}), & (5) \\ \mathbf{v}_i &= \mathbf{x}_i + F_i(\mathbf{x}^p_{best} - \mathbf{x}_i) + F_i(\mathbf{x}_{r_1} - \mathbf{x}_{r_2}), & (6) \end{aligned}$$

where $\mathbf{x}^p_{best}$ is a vector chosen randomly from the top $p\%$ individuals and $\mathbf{x}_i$, $\mathbf{x}_{r_1}$ and $\mathbf{x}_{r_2}$ are chosen from the current population P, while $\tilde{\mathbf{x}}_{r_2}$ is chosen randomly from $P \cup A$. Where A denotes the archive of JADE, which records the inferior parent solutions found during the current generation.

2) Control Parameter Adaptation: For each individual $\mathbf{x}_i$, control parameter F_i and the crossover probability, CR_i are generated independently from Cauchy and Normal distributions, respectively as follows [15]:

$$\begin{aligned} F_i &= rand(\mu F, 0.1) & (7) \\ CR_i &= rand(\mu CR, 0.1), & (8) \end{aligned}$$

where $rand$ is a uniform random number from $[0, 1]$, and μCR and μF are the means of the Normal and Cauchy distributions with standard deviation 0.1. Cauchy distribution is more helpful than the Normal distribution to diversify the mutation factors and thus prevent premature convergence, which often occurs in mutation strategies if the mutation factors are highly concentrated around a certain value [15]. The standard deviation is chosen to be relatively small (0.1) because otherwise the adaptation does not function efficiently; e.g., in the case of an infinite standard deviation, the truncated Normal distribution gets independent of the value of μCR [15]. CR_i and F_i given in Equations (7) and (8) are then truncated to $(0, 1]$ and $[0, 1]$, respectively. Initially, both μF and μCR are set to 0.5 as suggested in [15]. They are expressed as below [15]:

$$\mu F = (1 - c)\mu F + c \cdot mean_L(S_F) \tag{9}$$

$$\mu CR = (1 - c)\mu CR + c \cdot mean_A(S_{CR}). \tag{10}$$

Here $mean_L$ denotes the Lehmer mean, $mean_A$ denotes the arithmetic mean, and S_F is the set of successful F_i's, while S_{CR} is the set of successful CR_i's at generation G. The Lehmer mean is helpful to propagate larger mutation factors, which in turn improves the progress rate. To the contrary, an arithmetic mean of S_F tends to be smaller than the optimal value of the mutation factor and thus it might cause premature convergence at the end. The parameter c in Equations (9) and (10) is a constant which controls the rate of parameter adaptation and is chosen between 0 and 1. The life span of a successful CR_i or F_i is roughly $\frac{1}{c}$ generations; i.e., after $\frac{1}{c}$ generations, the old value of μCR or μF is reduced by a factor of $(1 - c)^{\frac{1}{c}}$, when c is close to zero, if $c = 0$ no parameter adaptation takes place.

3) Optional External Archive: At each generation, the failed parents are sent to the archive. The Euclidian distance of the archive members from the current population is utilized in the mutation operation in order to diversify the population and avert the premature convergence. If the archive size exceeds N_p, some solutions are randomly deleted from it to keep its size equal to N_p.

III. REVIEW

Almost two decades have been passed when DE was proposed in 1995 to cope with non-differentiable, non-convex and non-linear problems defined in the continuous parameter space [32]. Since then, DE and its uncountable and diversified variants have emerged as one of the most competitive and versatile family of the evolutionary computing optimizers and have been prosperously applied to solve numerous real-world problems from diverse discipline of science and technology [33]. Extensive literature on DE is available, which is evident from the recent surveys on DE [34], [32]. However, this section attempts to review some of the relevant methods. The hybridization of DE with local search strategies is a popular area of research among the practitioners. Many hybrid algorithms have shown significant performance improvement.

In [35] Sequential Quadratic Programming (SQP) is merged in DE algorithm. This new hybrid applies the DE algorithm until function evaluations reach 30% of the maximum function evaluations. It then applies SQP for the first time to the best point thus obtained. Afterwards, SQP is applied after each 100 generations to the best solution of the current search. In this work, the population size keeps reducing dynamically and the process terminates with minimum population size.

In another experiment DE is combined with simplex method and this method is know as NSDE [36]. The authors applied nonlinear simplex method with uniform random numbers to initialize DE population. Initially, N_p individuals are generated uniformly and then next N_p are generated from these N_p points by application of Nelder-Mead Simplex (NMS). Now from $2N_p$ population, the fittest N_p are selected as DE's initial population and the rest of DE is unchanged in this algorithm. Their algorithm only modify DE in the population step.

Further, differential evolution algorithm with localization around the best point (DELB) is proposed in [37]. In DELB the initial evolutionary steps are the same as DE except that the mutation scale factor F is chosen from $[-1, -0.4] \cup [0.4, 1]$ randomly for each mutant vector, DELB modifies the selection of DE by introducing reflection and contraction. The trial vector is compared with the current best and the parent vector. If the parent is worse than the trial vector it is replaced by a new concentrated or reflected vector. In DELB, the trial vector can be replaced by its parent vector, or reflected vector or contracted vector, while in classic DE only the trial vector replaces the parent.

Inspired by the above techniques, a new variant RJADE/TA of DE family is presented, which records the best individuals of the optimization process at regular intervals. Besides, it utilizes an reflection strategy of local search for replacing the archived solutions. The detail of RJADE/TA is presented in the following section.

IV. PROPOSED REFLECTED ADAPTIVE DIFFERENTIAL EVOLUTION WITH TWO EXTERNAL ARCHIVES

This section proposes a new DE algorithm, RJADE/TA, which modifies JADE in two aspects, first it introduces a second external archive into JADE, which stores superior solutions of the search at regular intervals of the optimization process. Second, these superior solutions are then reflected by new significant/potantial solutions in the current population. RJADE/TA adopts the same crossover and mutation operations as described in JADE [15]. We have done some modification to the Pseudo-code of JADE; this addition can be seen in lines 26 to 31 of Algorithm 1. Further in the last line the best solution is selected from PUA_2, the rest of the code remains the same.

Algorithm 1 Pseudo-code of RJADE/TA

```
1: Population size = $N_p$; $FES$ = Number of function evaluations; $\kappa$ = interval between second archive updates;
2: Uniformly and randomly sample $N_p$ solutions, $\mathbf{x}_{r_1,G}, \mathbf{x}_{r_2,G}, \ldots, \mathbf{x}_{r_{N_p},G}$ from the search space to form the initial population $P$;
3: Initialize the archives $A = \emptyset$; $A_2 = \emptyset$;
4: Set $\mu CR = 0.5$; $\mu F = 0.5$; $p = 5\%$; $c = 0.1$;
5: Set $S_{CR} = \emptyset$; $S_F = \emptyset$;
6: Evaluate these individuals; Set $FES = N_p$;
7: while $FES < n * 10000$ do
8:     Generate $CR_i = rand(\mu CR, 0.1)$;
9:     Generate $F_i = rand(\mu F, 0.1)$;
10:    Select $\mathbf{x}^p_{best,G}$ randomly from $100p\%$ population;
11:    Select $\mathbf{x}_{r_1,G} \neq \mathbf{x}_{i,G}$ randomly from $P$;
12:    Select $\tilde{\mathbf{x}}_{r_2,G} \neq \mathbf{x}_{r_2,G}$ randomly from $P \cup A$;
13:    Generate mutant $\mathbf{v}_i = \mathbf{x}_{i,G} + F_i(\mathbf{x}^p_{best,G} - \mathbf{x}_{i,G}) + F_i(\mathbf{x}_{r_1,G} - \tilde{\mathbf{x}}_{r_2,G})$;
14:    for $j = 1$ to $n$ do
15:        if $j < j_{rand}$ or $rand(0,1) < CR_i$ then
16:            $\mathbf{u}_{j,i,G} = \mathbf{v}_{j,i,G}$;
17:        else
18:            $\mathbf{x}_{j,i,G} = \mathbf{x}_{j,i,G}$;
19:        end if
20:    end for
21:    Select the best between $\mathbf{x}_{i,G}$ and $\mathbf{u}_{i,G}$;
22:    if $\mathbf{u}_{i,G}$ is better then
23:        $\mathbf{x}_{i,G} \rightarrow A$;,$CR_i \rightarrow S_{CR}$, $F_i \rightarrow S_F$;
24:    end if
25:    Delete individuals randomly from A if size $A > N_p$;
26:    Update second archive $A_2$ by sending best point of the search to it;
27:    if $Gen = \kappa$ then
28:        $\mathbf{x}_{best,G} \rightarrow A_2$; and reflect it as
29:        Compute the centroid of $P - \mathbf{x}_{best,G}$ as $\mathbf{x}_{c,G} = \frac{1}{N_p-1}\sum_{i=2}^{N_p} \mathbf{x}_{i,G}$
30:        Generate reflection point as $\mathbf{x}_{r,G} = \mathbf{x}_{c,G} + 1(\mathbf{x}_{c,G} - \mathbf{x}_{best,G})$
31:    end if
32:    $\mu CR = (1 - c) \cdot \mu CR + c \cdot mean_A(S_{CR})$;
33:    $\mu F = (1 - c) \cdot \mu F + c \cdot mean_L(S_F)$;
34: end while
35: Output: the solution vector with the smallest objective function value from $PUA_2$ in the search.
```

A. Best Solution's Reflection

Early convergence of the algorithms may be achieved due to best solution. Thus to avoid premature convergence, stagnation and local optima RJADE/TA reflects the best solution,

$\mathbf{x}_{best,G}$ of the search process and send it to the archive A_2. To implement the reflection mechanism [38] in RJADE/TA, first the center of mass of the current population P except the best solution $\mathbf{x}_{best,G}$ is computed as:

$$\mathbf{x}_{c,G} = \frac{1}{N_p - 1} \sum_{i=2}^{N_p} \mathbf{x}_{i,G} \tag{11}$$

where $\mathbf{x}_{c,G}$ denotes the center of mass of $N_p - 1$ individuals, since one candidate solution will be archived, this operation can be seen in Algorithm 1 (line 29). Once the center of mass of $N_p - 1$ individuals is calculated, then the best individual $\mathbf{x}_{best,G}$ (the solution with minimum objective value) of P is reflected through the center of mass $\mathbf{x}_{c,G}$ as follows:

$$\mathbf{x}_{r,G} = \mathbf{x}_{c,G} + 1 \cdot (\mathbf{x}_{c,G} - \mathbf{x}_{best,G}). \tag{12}$$

Where $\mathbf{x}_{r,G}$ is the mirror image or reflection [38] of $\mathbf{x}_{best,G}$ through the centroid $\mathbf{x}_{c,G}$, this newly produced solution is known as reflected solution. The coefficient of reflection is ”1” as suggested in [38].

The reflected solution replaces $\mathbf{x}_{best,G}$ in the population P and the best solution $\mathbf{x}_{best,G}$ by itself is transferred to the second archive A_2.

B. **Second External Archive in RJADE/TA**

When the search procedure reaches its 50% function evaluations the first archive A_2 update is made. After which A_2 is updated at regular interval of generations κ. As mentioned earlier that JADE has archive A, which stores inferior solutions, if the archive size exceeds N_p; some solutions are removed from it. In contrast the proposed second archive A_2 records the best solution of the search after each κ generations. In other words the best solution of the current population, after κ generations is removed from the search procedure and is kept passive in archive A_2 during the optimization. The objective of sending the best solution from the current optimization process is that the best solution information may cause difficulties such as premature convergence due to the resultant reduced population diversity [15]. Best solution some times mislead the search to local optima or stagnation.

The second archive A_2 is initialized as 0 and is updated with a best solution in each κ generations (see Algorithm 1). The interval between two reflections is κ, this is kept 1000 here. If we reflect the best solution at each generation, there will be one extra evaluation at each generation, which may be a wastage of computational energy. Furthermore, if we store best solution at each generation then the best solution of current generation and the previous will be not much different from each other. Which again will be wastage of computation. That is why we selected κ a 1000. There are few differences in A and A_2 which are given below.

1) A_2 stores best solution of the current population, while A records the recently explored inferior solutions.
2) The size of A is kept N_p, if this size exceeds, some solutions are randomly deleted from A, however in the new archive A_2 the size may exceeds N_p. It keeps the record of all best solutions, no solution is removed from it.
3) A_2 records the best solution (only one solution) of the current generation, this may be a parent solution or a child solution. In contrast, A keeps the inferior parents solutions (more than one) only, it does not record inferior child solution.
4) A_2 is initialized as 0 and is updated after κ generations (1000 say). On the other hand A is updated at the end of each generation.
5) The recorded inferior parents of A are later on utilized in mutation. Where in A_2 the stored best solution is reflected with a new solution; which is sent to the current population. Once a solution is kept in A_2, it remains inactive during the optimization. When the search procedures are terminated, then the second archive’s solution contribute towards the selection of optimal solution.

V. Numerical Experiments and Results

A. Experimental Setup

Experimental validations for the proposed RJADE/TA are conducted on a set of 28 new and complex test functions [39] provided by CEC 2013 special session and a 1000 dimensional functions designed for CEC 2010 competition on large scale global optimization problems [40].

B. CEC 2013 Test Suite

In the CEC 2013 test suite, the previously proposed composition functions of CEC 2005 [2] are enhanced and additional test functions are considered for real parameter single objective optimization. Three types of problems are developed:

- Functions 1-5 are unimodal;
- 6-20 are multimodal functions.
- 21-28 are composit functions, which are designed by combining various problems into a complex landscape.

C. Parameter Settings for CEC 2013 Test Suite

We performed our experiments following the guidelines of the CEC2013 competition [39]. For all the problems, the initialization range is $[-100; 100]$. For all of the problems the number of dimensions are $n = 10$ and 30, and the maximum number of objective function evaluations are $10000 \times n$ per run. When the difference between the values of the best solution found and the optimal (known) solution is 10^{-8} or less, the error is set to 0. The population size is set to 100.

D. Results on CEC 2013 functions

The experimental statistics(best, mean, median, worst and standard deviation) obtained by our algorithm in 51 runs, on 28 functions with dimensions $n = 10$ of the CEC 2013 test functions are summarized in Table I. In Table II, the Mean values of function error values($f(x) - f(x^*)$) obtained by RJADE/TA are presented for $n = 10$. These values are compared with state of the art algorithms, jDE, jDEsoo [41] a new version of jDE, SPSRDEMMS [42] and jDErpo [13]. Among these SPSRDEMMS and jDErpo were specially developed for CEC 2013 competition.

In Table II the **-** shows that the corresponding algorithm loesses against our RJADE/TA algorithm. The **+** indicates that the particular algorithm wins against our algorithm, and **=** reveals that both the algorithms performs equivalently. The outstanding performance of RJADE/TA is clearly visible from Table II, where many negative **-** signs made this fact evident. It is very clear from the Table that our RJADE/TA algorithm performed significantly better than jDE and jDEsoo algorithms on 15 out of 28 functions, on 4 functions both got similar results. On the other hand jDE and jDEsoo showed better performance on only 9 functions. As compared with SPSRDEMMS, our algorithm found better solutions for 16 out of 28 functions and SPSRDEMMS showed good results on 12 functions. Furthermore, jDErpo and RJADE/TA performed better than each other on 12 functions.

Table III shows the comparison of RJADE/TA against jDE, jDEsoo, SPSRDEMMS and jDErpo for $n = 30$. It is interesting to note that the performance of RJADE/TA increased with the increase in dimension. It found better results for 20 out of 28 function against jDE and jDEsoo. jDE only solved 5 out of 28 problems for 30 dimensions, and jDEsoo got good results on 3 out of 28 functions. SPSRDEMMS and jDErpo performed inferior on 16 functions, and superior on 8 functions only, which can be seen from Table III.

Tables II and III showed the comparison of RJADE/TA against each of the particular algorithms. Here we present the overall percentage of all the algorithms, jDE, jDErpo, SPSRDEMMS, jDErpo and RJADE/TA on 30 dimensional problems. Table IV demonstrates that RJADE/TA performance percentage is 50% while jDErpo is 37%, the remaining three algorithms in comparison performed less than or equal to 25%. This percentage validity is even more clearly visible from the bar graph 1. Each bar shows the number of test problems optimized by particular algorithm. The last bar representing RJADE/TA.

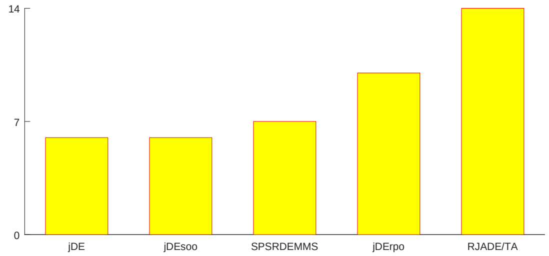


Fig. 1: Comparison of RJADE/TA and other up to date algorithms with dimension $n = 30$

E. CEC2010 Test Instances

Here we evaluate RJADE/TA on ten complex optimization problems used in CEC2010 special session and competition on large scale global optimization [40]. Since separability provides a measure of the complexity of various problems, in [40] a test suite for high dimensional problems is devised which is based on separability and non separability of the functions. Here, three kinds of high-dimensional problems are considered:

- Separable functions;
- Partially-separable functions, in which only a small number of variables are dependent and the rest are independent;
- Partially-separable functions that consist of multiple independent subcomponents, each of which is m-non-separable; and

This test suite provided an enhanced platform for evaluating the performance of algorithms on high-dimensional problems in various scenarios [40]. Below we list only those test functions (F1-F10) which are used in this work.

1) Separable Functions (3)
   - F1: Shifted Elliptic Function
   - F2: Shifted Rastrigin's Function
   - F3: Shifted Ackley's Function
2) Single-group m-nonseparable Functions (5)
   - F4: Single-group Shifted and m-rotated Elliptic Function
   - F5: Single-group Shifted and m-rotated Rastrigin's Function
   - F6: Single-group Shifted and m-rotated Ackley's Function
   - F7: Single-group Shifted m-dimensional Schwefel's Problem 1.2
   - F8: Single-group Shifted m-dimensional Rosenbrock's Function
3) $\frac{n}{2m}$-group m-nonseparable Functions (2)
   - F9: $\frac{n}{2m}$-group Shifted and m-rotated Elliptic Function
   - F10: $\frac{n}{2m}$-group Shifted and m-rotated Rastrigin's Function

The parameter m controls the number of variables in each group and hence defining the degree of separability.

F. Parameter Settings for CEC2010 instances

For this experiment the population size N_p is chosen 50 and the dimension n is set to 1000. The maximum function evaluations are chosen $3 \times 10^{+06}$. The value to reach is set to 10^{-2}. RJADE/TA and JADE were run 25 independent times for all test instances as suggested in the original paper [40]. All these experiments were conducted in MATLAB software.

G. Comparison of RJADE/TA with JADE 0n CEC 2010 instances

The best, median, mean and standard deviation of function error values obtained in 25 runs of the proposed algorithm, RJADE/TA are presented in Table V. These statistics were requested in [40] as well. The best results are typed as bold.

As can be seen from Table V, overall RJADE/TA performed well as compared with JADE in obtaining the "best" solution for five out of ten test instances, F3, F4, F5, F7 and F8. For F6 both algorithms got the same accuracy. Here F3 is separable and all others are single-group m-nonseparable functions. Surly it is due to the additional second archive of RJADE/TA which provides more chance to the population for searching the region and discouraging early convergence. For

TABLE I: EXPERIMENTAL RESULTS OF RJADE/TA ON 28 TEST FUNCTIONS OVER 51 RUNS WITH DIMENSION n = 10.

Func	Best	Worst	Median	Mean	Std Dev
1	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$
2	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$
3	$2.2737E - 013$	$9.3924E + 02$	$5.1956E + 01$	$1.2108E + 02$	$1.8941E + 02$
4	$0.0000E + 00$	$5.9114E + 03$	$0.0000E + 00$	$1.1591E + 02$	$8.2776E + 02$
5	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$	$0.0000E + 00$
6	$0.0000E + 00$	$9.8124E + 00$	$9.8124E + 00$	$7.8884E + 00$	$3.9346E + 00$
7	$1.9695E - 03$	$1.2013E + 00$	$7.5908E - 02$	$1.5927E - 01$	$2.1904E - 01$
8	$2.0201E + 01$	$2.0511E + 01$	$2.0358E + 01$	$2.0366E + 01$	$6.7627E - 02$
9	$3.4033E + 00$	$6.0808E + 00$	$4.4493E + 00$	$4.4593E + 00$	$6.0360E - 01$
10	$5.6843E - 014$	$6.8717E - 02$	$3.6112E - 02$	$3.5342E - 02$	$1.4363E - 02$
11	$5.6843E - 014$	$1.1937E - 012$	$2.2737E - 013$	$2.4298E - 013$	$1.9922E - 013$
12	$3.5441E + 00$	$1.1542E + 01$	$6.7460E + 00$	$6.7571E + 00$	$1.6197E + 00$
13	$3.9347E + 00$	$1.1345E + 01$	$7.9523E + 00$	$7.7246E + 00$	$1.9071E + 00$
14	$1.0282E - 04$	$1.2604E - 01$	$2.1817E - 03$	$1.1994E - 02$	$2.5730E - 02$
15	$3.9803E + 02$	$9.2821E + 02$	$6.5814E + 02$	$6.6660E + 02$	$1.2744E + 02$
16	$6.2944E - 01$	$1.4778E + 00$	$1.1505E + 00$	$1.1336E + 00$	$1.8774E - 01$
17	$1.0122E + 01$	$1.0122E + 01$	$1.0122E + 01$	$1.0122E + 01$	$4.5729E - 06$
18	$1.7593E + 01$	$3.1133E + 01$	$2.2134E + 01$	$2.2715E + 01$	$2.8525E + 00$
19	$2.9479E - 01$	$5.2259E - 01$	$4.5204E - 01$	$4.4224E - 01$	$5.3887E - 02$
20	$1.2860E + 00$	$3.4877E + 00$	$2.5708E + 00$	$2.5317E + 00$	$3.7190E - 01$
21	$2.0000E + 02$	$4.0019E + 02$	$4.0019E + 02$	$3.9627E + 02$	$2.8033E + 01$
22	$1.9796E - 02$	$1.1123E + 02$	$1.7431E + 01$	$2.7022E + 01$	$2.6637E + 01$
23	$2.8879E + 02$	$1.0544E + 03$	$6.9580E + 02$	$7.0015E + 02$	$1.5859E + 02$
24	$1.3524E + 02$	$2.1472E + 02$	$2.0279E + 02$	$2.0217E + 02$	$1.2455E + 01$
25	$2.0003E + 02$	$2.1188E + 02$	$2.0091E + 02$	$2.0314E + 02$	$3.6775E + 00$
26	$1.0514E + 02$	$2.0002E + 02$	$1.1187E + 02$	$1.2670E + 02$	$3.4574E + 01$
27	$3.0001E + 02$	$4.0438E + 02$	$3.0019E + 02$	$3.0351E + 02$	$1.6372E + 01$
28	$1.0000E + 02$	$3.0000E + 02$	$3.0000E + 02$	$2.8824E + 02$	$4.7525E + 01$

TABLE II: COMPARISON OF RJADE/TA WITH OTHER ALGORITHMS ON THE MEAN OF THE FUNCTION ERROR VALUES AT EXECUTION TERMINATION OVER 51 RUNS, ON 28 TEST FUNCTIONS WITH n=10.

Func	jDE	jDEsoo	SPSRDEMMS	jDErpo	RJADE/TA
1	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$
2	$7.6534e - 05$-	$1.7180e + 03$-	$6.8886e + 02$-	$0.0000e + 00$=	$0.0000e + 00$
3	$1.3797e + 00$+	$1.6071e + 00$+	$5.9735e + 00$+	$3.7193e - 05$+	$1.2108e + 02$
4	$3.6639e - 08$+	$1.2429e - 01$+	$3.8803e - 02$+	$0.0000e + 00$+	$1.1591e + 02$
5	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$
6	$8.6581e + 00$-	$8.4982e + 04$-	$8.6580e + 00$-	$5.3872e + 00$+	$7.8884e + 00$
7	$2.7229e - 03$+	$9.4791e - 01$-	$1.8732e - 01$-	$1.6463e - 03$+	$1.5927e - 01$
8	$2.0351e + 01$+	$2.0348e + 01$+	$2.0348e + 01$+	$2.0343e + 01$+	$2.0366e + 01$
9	$2.6082e + 00$+	$2.7464e + 00$+	$2.7311e + 00$+	$6.4768e - 01$+	$4.4593e + 00$
10	$4.5263e - 02$-	$7.0960e - 02$-	$1.0346e - 01$+	$6.4469e - 02$-	$3.5342e - 02$
11	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$=	$0.0000e + 00$
12	$1.2304e + 01$-	$6.1144e + 00$+	$7.5821e + 00$+	$1.3410e + 01$-	$7.7246e + 00$
13	$1.3409e + 01$-	$7.8102e + 00$-	$1.1042e + 01$-	$1.4381e + 01$-	$6.7571e + 00$
14	$0.0000e + 00$+	$5.0208e - 02$-	$8.3273e - 02$-	$1.9367e + 01$-	$1.1994e - 02$
15	$1.1650e + 03$-	$8.4017e + 02$-	$8.3072e + 02$-	$1.1778e + 03$-	$6.6660e + 02$
16	$1.0715e + 00$+	$1.0991e + 00$+	$1.1871e + 00$-	$1.0598e + 00$+	$1.1336e + 00$
17	$1.0122e + 01$=	$9.9240e + 00$+	$1.0127e + 01$	$1.0997e + 01$-	$1.0122e + 01$
18	$3.2862e + 01$-	$2.7716e + 01$-	$2.2949e + 01$-	$3.2577e + 01$-	$2.2715e + 01$
19	$4.3817e - 01$+	$3.1993e - 01$+	$3.1854e - 01$+	$7.4560e - 01$-	$4.4224e - 01$
20	$3.0270e + 00$-	$2.7178e + 00$-	$2.5112e + 00$+	$2.5460e + 00$-	$2.5317e + 00$
21	$3.7272e + 02$+	$3.5113e + 02$+	$3.9234e + 02$-	$3.7272e + 02$+	$3.9627e + 02$
22	$7.9231e + 01$-	$9.1879e + 01$-	$6.6219e + 01$-	$9.7978e + 01$-	$2.7022e + 01$
23	$1.1134e + 03$-	$8.1116e + 02$-	$9.4740e + 02$-	$1.1507e + 03$+	$7.0015e + 02$
24	$2.0580e + 02$-	$2.0851e + 02$-	$2.0442e + 02$-	$1.8865e + 02$+	$2.0217e + 02$
25	$2.0471e + 02$-	$2.0955e + 02$-	$2.0473e + 02$-	$1.9885e + 02$+	$2.0314e + 02$
26	$1.8491e + 02$-	$1.9301e + 02$-	$1.6886e + 02$-	$1.1732e + 02$+	$1.2670e + 02$
27	$4.7470e + 02$-	$4.9412e + 02$-	$4.7300e + 02$-	$3.0000e + 02$+	$3.0351e + 02$
28	$2.9216e + 02$-	$2.8824e + 02$=	$2.8431e + 02$+	$2.9608e + 02$-	$2.8824e + 02$
-	15	15	16	12	
+	9	9	9	12	
=	4	4	3	4	

TABLE III: COMPARISON OF RJADE/TA WITH OTHER ALGORITHMS ON THE MEAN OF THE FUNCTION ERROR VALUES AT EXECUTION TERMINATION OVER 51 RUNS, ON 28 TEST FUNCTIONS WITH n=30.

Func	jDE	jDEsoo	SPSRDEMMS	jDErpo	RJADE/TA
1	0.0000e + 00=	0.0000e + 00=	0.0000e + 00=	0.0000e + 00=	0.0000e + 00
2	1.1925e + 06-	1.2914e + 05-	1.0157e + 05-	2.8378e + 04-	7.4009e + 03
3	5.6216e + 06-	9.8414e + 06-	1.0951e + 07-	8.5740e + 01+	2.4293e + 05
4	9.3584e + 03-	1.9720e + 04-	2.4061e + 00+	1.7214e + 02+	5.1627e + 03
5	0.0000e + 00=	1.2606e − 08=	0.0000e + 00=	0.0000e + 00=	0.0000e + 00
6	1.4157e + 01-	7.9292e + 00-	1.7463e + 01-	7.5852e + 00-	1.0356e + 00
7	2.6171e + 01-	9.8167e + 00-	1.1038e + 01-	1.1163e + 00+	4.2514e + 00
8	2.0934e + 01+	2.0946e + 01-	2.0950e + 01-	2.0940e + 01-	2.0937e + 01
9	1.8151e + 01+	2.0971e + 01-	2.4903e + 01+	3.0923e + 01-	2.7961e + 01
10	3.8212e − 02-	7.9055e − 02-	5.3974e − 02-	9.2759e − 03+	3.7380e − 02
11	3.6609e + 01-	0.0000e + 00=	0.0000e + 00=	3.2858e + 01-	0.0000e + 00
12	1.7135e + 02-	4.2835e + 01-	4.2650e + 01-	1.7995e + 02-	3.6994e + 01
13	1.8086e + 0-2	7.0750e + 01-	7.9763e + 01-	1.8151e + 02-	5.7309e + 01
14	3.0639e + 03-	1.3327e + 00-	3.2550e + 00-	1.5120e + 03-	1.1223e + 00
15	7.2978e + 03-	4.8340e + 03-	4.4226e + 03-	7.1440e + 03-	4.1938e + 03
16	2.4646e + 00-	2.2791e + 00-	2.2801e + 00-	2.4687e + 00-	2.1305e + 00
17	7.8765e + 01-	3.0434e + 01=	3.0440e + 01-	7.3585e + 01-	3.0434e + 01
18	2.1731e + 02-	1.2341e + 02-	8.9310e + 01+	2.1298e + 02-	1.0213e + 02
19	7.0078e + 00-	1.0956e + 00+	1.1639e + 00+	7.5022e + 00-	2.0825e + 00
20	1.2564e + 01-	1.1639e + 01-	1.1236e + 01-	1.2268e + 01-	1.0858e + 01
21	2.7818e + 02+	2.9396e + 02-	2.8466e + 02+	2.8637e + 02+	2.9336e + 02
22	3.1346e + 03-	5.1621e + 01+	7.6606e + 01+	1.7779e + 03-	1.3131e + 02
23	7.2920e + 03-	4.6061e + 03-	4.7713e + 03-	7.2374e + 03-	4.2998e + 03
24	2.5511e + 02-	2.4818e + 02-	2.5330e + 02-	2.0102e + 02+	2.1616e + 02
25	2.5213e + 02+	2.6037e + 02+	2.6408e + 02+	2.5354e + 02+	2.7921e + 02
26	2.0015e + 02+	2.5758e + 02-	2.0001e + 02+	2.0000e + 02+	2.2275e + 02
27	7.8688e + 02-	7.2161e + 02-	8.8779e + 02-	3.7724e + 02-	7.1060e + 02
28	3.0000e + 02=	3.0000e + 02=	3.0000e + 02=	3.0000e + 02=	3.0000e + 02
-	20	20	16	16	
+	5	3	8	8	
=	3	5	4	3	

TABLE IV: %age comparison of RJADE/TA with other algorithms

Optimizer	jDE	jDEsoo	SPSRDEMMS	jDErpo	RJADE/TA
No. of probs. optimized	7	6	6	10	14
%age	25%	21%	21%	36%	**50%**

TABLE V: EXPERIMENTAL RESULTS OF JADE, AND RJADE/TA ON 10 TEST INSTANCES OF 1000 VARIABLES WITH $3 \cdot 10^{+06} FES$.Best, Median, Mean AND the Std Dev OF THE FUNCTION ERROR VALUES OBTAINED OVER 25 RUNS.

Test	Best		Mean		Median		Std Dev	
Instance	RJADE/TA	JADE	RJADE/TA	JADE	RJADE/TA	JADE	RJADE/TA	JADE
F1	3.12E + 05	**1.44E + 05**	1.26E + 06	**1.18E + 06**	**9.70E + 05**	1.05E + 06	**8.07E + 05**	9.24E + 05
F2	1.44E + 03	**1.09E + 01**	1.63E + 03	**7.28E + 01**	1.61E + 03	**5.87E + 01**	1.30E + 02	**4.60E + 01**
F3	**7.27E − 03**	9.49E − 01	**5.47E − 01**	1.20E + 00	**4.32E − 01**	1.20E + 00	5.43E − 01	**1.47E − 01**
F4	**6.29E + 10**	9.08E + 10	**6.60E + 12**	8.12E + 12	1.80E + 11	**1.56E + 11**	**1.03E + 13**	1.09E + 13
F5	**4.93E + 007**	5.65E + 007	**6.91E + 007**	7.62E + 007	**6.71E + 007**	7.71E + 007	1.52E + 007	**1.28E + 007**
F6	1.97E + 01 ≈	1.97E + 01	**1.98E + 01**	3.52E + 04	1.98E + 01 ≈	1.98E + 01	**2.80E − 02**	1.76E + 05
F7	**4.61E + 05**	4.58E + 05	**1.19E + 09**	6.29E + 07	7.80E + 05	**8.53E + 05**	**2.47E + 09**	1.36E + 08
F8	**7.10E + 04**	9.00E + 04	4.07E + 07	**2.60E + 07**	**4.11E + 06**	6.97E + 06	6.00E + 07	**3.49E + 07**
F9	3.98E + 07	**3.65E + 07**	5.03E + 07	**4.83E + 07**	5.02E + 07	**4.51E + 07**	**1.04E + 07**	7.52E + 06
F10	5.04E + 03	**3.34E + 03**	5.35E + 03	**3.67E + 03**	5.35E + 03	**3.69E + 03**	1.66E + 02	**1.59E + 02**

the remaining four test instances, F1, F3, F9 and F10 JADE got better solutions than RJADE/TA; here F1 and F3 are separable and two functions F9 and F10 are partially-separable functions that consist of multiple independent subcomponents. Furthermore, the failure on F10 ($\frac{n}{2m}$-group nonseparable) could be its complexity, as it is the sum of ten rotated Rastrigins functions applied to groups of m (50 here) decision variables each and one non-rotated Rastrigins function applied to the remaining 500 decision variables. The failure on F9 can be due to its complex nature like F10.

Considering “Mean”, “Median” and Standard deviation, we see that RJADE/TA’s is more suitable to solve single-group m-nonseparable functions, F3-F8, which is visible from Table V. Hence in general, the analysis of above experimental results lead us to the conclusion that RJADE/TA in much much better than JADE in optimizing problems from the category of single-group m-nonseparable functions.

VI. Conclusions

The current DE variant JADE with one optional external archive some times exhibit poor reliability [30]. Moreover, best solutions some times mislead the search to a local optima. In this paper, we have attempted to introduce a second archive A_2 into JADE for overcoming this shortcoming for large scale global optimization problems. This archive stores the best solution, which is removed from the current population after regular intervals. The removal of best solution is compensated by a new potential solution in the population. Thus we have proposed an approach RJADE/TA to add A_2 to JADE algorithm and add new good divers solutions to the population to make a systematic and rational search in the region defined for the search process. RJADE/TA takes the advantages of both archives, A with inferior solutions and A_2 with superior solutions. It is easy to implement and does not introduce any complicated structures.

The performance of the developed RJADE/TA has been demonstrated by taking advantage of 28 complex competition test functions from CEC 2013 and 10 functions from CEC2010. On CEC2013 test suit RJADE/TA was compared with jDE, jDEsoo, jDErpo and SPSRDEMMS algorithms on 10 and 30 dimensions. The superior performance of RJADE/TA was demonstrated on 10 and 30 dimensions. Moreover, we have compared RJADE/TA with classical JADE with 1000 dimensions. RJADE/TA notably outperformed JADE and is very competitive in solving single-group m-nonseparable functions. In this paper, our aim was to analyze the behavior of algorithm if the best solution is removed from it.

In future JADE with second Archive only can be explored. The experiments may be carried out at other higher dimensional problems. This may be extended to constrained optimization.

References

[1] B. D. Bunday, *Basic optimisation methods*. Arnold, London, 1984.

[2] R. Storn and K. V. Price, “Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces,” *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[3] R. Storn and K. Price, “Home page of differential evolution,” 2003.

[4] Y. Wang, Z. Cai, and Q. Zhang, “Differential evolution with composite trial vector generation strategies and control parameters,” *IEEE Transactions on Evolutionary Computation*, vol. 15, no. 1, pp. 55–66, 2011.

[5] S. Rahnamayan, T. H. R., and M. M. A. Salama, “Opposition-based differential evolution (ode),” *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 1, pp. 64–79, 2008.

[6] R. A. Khanum and M. A. Jan, “Centroid-based initialized jade for global optimization,” in *Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd*. IEEE, 2011, pp. 115–120.

[7] I. Poikolainen, F. Neri, and F. Caraffini, “Cluster based population initialization for differential evolution frameworks,” *Information Sciences*, vol. 297, pp. 216–235, 2015.

[8] J. Brest, , S. Greiner, B. Boskovic, V. Zumer, and M. M. S, “Self-adaptive control parameters in differential evolution: A comparative study on numerical benchmark problems,” pp. 446–657, 2006.

[9] N. Noman and H. Iba, “Acceleratinging differential evolution using an adaptive local search,” *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 3, pp. 107–125, 2008.

[10] L. Tang, Y. Dong, and J. Liu, “Differential evolution with an individual-dependent mechanism,” *Evolutionary Computation, IEEE Transactions on*, vol. 19, no. 4, pp. 560–574, 2015.

[11] J. Brest, A. Zamuda, B. Boskovic, S. Greiner, and V. Zumer, *Advances in Differential Evolution*, 2008, vol. SCI143, ch. An Analysis of the control parameters’ Adaptation in Differential Evolution.

[12] Z. Yang, K. Tang, and X. Yao, “Self-adaptive differential evolution with neighborhood search,” in *IEEE Congress on Evolutionary Computation (CEC 2008)*. Singapore: IEEE Press, 2008.

[13] J. Brest, A. Zamuda, I. Fister, and B. Boskovic, “Some improvements of the self-adaptive jde algorithm,” in *Differential Evolution (SDE), 2014 IEEE Symposium on*. IEEE, 2014, pp. 1–8.

[14] A. K. Qin and P. N. Suganthan, “Self adaptive differential evolution algorithm for numerical optimization,” in *IEEE Congress on Evolutionary Computation (2005)*, vol. 2, 2005, pp. 1785–1791.

[15] J. Zhang and A. C. Sanderson, “Jade: adaptive differential evolution with optional external archive,” *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 5, pp. 945–958, 2009.

[16] C. Zhang and L. Gao, “An effective improvement of jade for real-parameter optimization,” in *Advanced Computational Intelligence ICACI, 2013 Sixth International Conference on*. IEEE, 2013, pp. 58–63.

[17] R. Mallipeddi, P. N. Suganthan, Q. K. Pan, and M. F. Tasgetiren, “Differential evolution algorithm with ensemble of parameters and mutation strategies,” *Applied Soft Computing*, vol. 11, pp. 1979–1696, 2010.

[18] M. Ali, M. Pant, and V. P. Singh, “An improved differential evolution algorithm for real parameter optimization problems,” *International Journal of Recent Trends in Engineering*, vol. 1, 2009.

[19] R. Tanabe and F. Alex, “Success history based parameter adaptation for differential evolution,” in *Evolutionary Computation (CEC), 2013 IEEE Congress on*. IEEE, 2013, pp. 71–78.

[20] R. Tanabe and A. S. Fukunaga, “Improving the search performance of shade using linear population size reduction,” in *Evolutionary Computation (CEC), 2014 IEEE Congress on*. IEEE, 2014, pp. 1658–1665.

[21] S.-M. Guo, J. S.-H. Tsai, C.-C. Yang, and P.-H. Hsu, “A self-optimization approach for l-shade incorporated with eigenvector-based crossover and successful-parent-selecting framework on cec 2015 benchmark set,” in *Evolutionary Computation (CEC), 2015 IEEE Congress on*. IEEE, 2015, pp. 1003–1010.

[22] J. Aalto and J. Lampinen, “A mutation and crossover adaptation mechanism for differential evolution algorithm,” in *Evolutionary Computation (CEC), 2014 IEEE Congress on*. IEEE, 2014, pp. 451–458.

[23] Zheng, K. Tang, and X. Yao, “Large scale evolutionary optimization using cooperative coevolution,” *Journal of Information Sciences*, vol. 178, no. 15, pp. 2985–2999, 2008.

[24] C. Deng, B. Zhao, A. Y. Deng, and C. Liang, “Hybrid-coding binary differential evolution algorithm with application to 0-1 knapsack problems,” pp. 317–320, 2008.

[25] C. S. Deng, B. Y. Zhao, and C. Y. Liang, “Hybrid binary differential

evolution algorithm for 0-1 knapsack problem," *Computer Engineering and Design*, vol. 31, no. 8, pp. 1795–1798, 2010.

[26] C. Segura, C. A. C. Coello, E. Segredo, and C. León, "An analysis of the automatic adaptation of the crossover rate in differential evolution," in *Evolutionary Computation (CEC), 2014 IEEE Congress on*. IEEE, 2014, pp. 459–466.

[27] A. Qin, K. Tang, H. Pan, and S. Xia, "Self-adaptive differential evolution with local search chains for real-parameter single-objective optimization," in *Evolutionary Computation (CEC), 2014 IEEE Congress on*. IEEE, 2014, pp. 467–474.

[28] F. Wei, Y. Wang, and T. Zong, "Variable grouping based differential evolution using an auxiliary function for large scale global optimization," in *Evolutionary Computation (CEC), 2014 IEEE Congress on*. IEEE, 2014, pp. 1293–1298.

[29] A. K. Qin, V. L. Huang, and P. N. Sughanthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE transactions on Evolutionary Computation*, vol. 13, no. 5, pp. 398–417, Oct,2009.

[30] F. Peng, K. Tang, G. Chen, and X. Yao, "Multi-start JADE with knowledge transfer for numerical optimization," in *IEEE Congress on Evolutionary Computation (2007)*, 2009.

[31] S. Das and P. N. Suganthan, "Tutorial:differential evolution, foundations, prospectives and applications," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, Paris, France, April, 2011, pp. 1–59.

[32] S. Das, S. S. Mullick, and P. Suganthan, "Recent advances in differential evolution an updated survey," *Swarm and Evolutionary Computation*, pp. 1–62, 2016.

[33] Swagatam and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *Evolutionary Computation, IEEE Transactions on*, no. 99, pp. 1–28, 2011.

[34] F. Neri and V. Tirronen, "Recent advances in differential evolution: a survey and experimental analysis," *Artificial Intelligence Review*, vol. 33, no. 1, pp. 61–106, 2010.

[35] J. Brest, A. Zamuda, B. Bošković, S. Greiner, M. S. Maučce, and V. Žumer, "Self-adaptative differential evolution with sqp local search," in *3rd international conference on Bio-inspired optimization methods and their applications(BIOMA)*, 2008, pp. 59–66.

[36] M. Ali, M. Pant, and A. Abraham, "Simplex differential evolution," *Acta Polytechnica Hungarica*, vol. 6, no. 5, 2009.

[37] P. Kaelo and M. M. Ali, "A numerical study of some modified differential evolution algorithms," *European journal of operational research*, vol. 169, no. 3, pp. 1176–1184, 2006.

[38] J.A.Nelder and R. Mead, "A simplex method for function minimization," *Comput.J*, vol. 7, no. 4, pp. 308–313, 1965.

[39] J. Liang, B. Qu, P. Suganthan, and A. G. Hernández-Díaz, "Problem definitions and evaluation criteria for the cec 2013 special session on real-parameter optimization," 2013.

[40] K. Tang, Xiodongo, P. N. Suganthan, Z. Yang, and T. Weise, "Benchmark functions for the CEC2010 special session and competition on large scale global optimization," Nature Inspired Computation and Application Laboratory (NICAL), University of Science and Technology of China, Tech. Rep., 2010.

[41] J. Brest, B. Boskovic, A. Zamuda, I. Fister, and E. Mezura-Montes, "Real parameter single objective optimization using self-adaptive differential evolution algorithm with more strategies," in *Evolutionary Computation (CEC), IEEE Congress on*. IEEE, 2013, pp. 377–383.

[42] A. Zamuda, J. Brest, and E. Mezura-Montes, "Structured population size reduction differential evolution with multiple mutation strategies on cec 2013 real parameter optimization," in *Evolutionary Computation (CEC), IEEE Congress on*. IEEE, 2013, pp. 1925–1931.