

Volume 6 Issue 1

January 2015



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)





W H E R E W I S D O M S H A R E S

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS



THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

OAlster

getCITED



BASE
Bielefeld Academic Search Engine

ULRICHSWEB™
GLOBAL SERIALS DIRECTORY

arXiv.org

DOAJ DIRECTORY OF
OPEN ACCESS
JOURNALS

IET InspecDirect

INDEX COPERNICUS
INTERNATIONAL

WorldCat™
Window to the world's libraries

Microsoft™
Academic Search

EBSCO
HOST
Research
Databases

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 6 Issue 1 January 2015

ISSN 2156-5570 (Online)

ISSN 2158-107X (Print)

©2013 The Science and Information (SAI) Organization

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University

Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation

Associate Editors

Chao-Tung Yang

Department of Computer Science, Tunghai University, Taiwan

Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing

Elena SCUTELNICU

"Dunarea de Jos" University of Galati, Romania

Domain of Research: e-Learning, e-Learning Tools, Simulation

Krassen Stefanov

Professor at Sofia University St. Kliment Ohridski, Bulgaria

Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications

Maria-Angeles Grado-Caffaro

Scientific Consultant, Italy

Domain of Research: Electronics, Sensing and Sensor Networks

Mohd Helmy Abd Wahab

Universiti Tun Hussein Onn Malaysia

Domain of Research: Intelligent Systems, Data Mining, Databases

T. V. Prasad

Lingaya's University, India

Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics

Reviewer Board Members

- **Abassi Ryma**
Higher Institute of Communications Studies of Tunis
, Iset'com
- **Abbas Karimi**
Islamic Azad University Arak Branch
- **Abdelghni Lakehal**
Université Abdelmalek Essaadi Faculté
Polydisciplinaire de Larache Route de Rabat, Km 2 -
Larache BP. 745 - Larache 92004. Maroc.
- **Abdel-Hameed A. Badawy**
Arkansas Tech University
- **Abdur Rashid Khan**
Gomal University
- **Abeer Mohamed ELkorany**
Faculty of computers and information, Cairo
Univesity
- **ADEMOLA ADESINA**
University of the Western Cape
- **Aderemi A. Atayero**
Covenant University
- **Ahmed S.A AL-Jumaily**
Ahlia University
- **Ahmed Boutejdar**
- **Ahmed Nabih Zaki Rashed**
Menoufia University
- **Akbar Hossain**
- **Akram Belghith**
University Of California, San Diego
- **Albert Alexander S**
Kongu Engineering College
- **Alci-nia Zita Sampaio**
Technical University of Lisbon
- **Alexandre Bouënard**
Sensopia
- **Ali Ismail Awad**
Luleå University of Technology
- **Amitava Biswas**
Cisco Systems
- **Anand Nayyar**
KCL Institute of Management and Technology,
Jalandhar
- **Andi Wahju Rahardjo Emanuel**
Maranatha Christian University
- **Andrews Samraj**
Mahendra Engineering College
- **Anirban Sarkar**
National Institute of Technology, Durgapur
- **Antonio Formisano**
- **Anuranjan misra**
Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**
University Technology Malaysia(UTM)
- **Aree Ali Mohammed**
Directorate of IT/ University of Sulaimani
- **Aris Skander Skander**
Constantine 1 University
- **Ashok Matani**
Government College of Engg, Amravati
- **Ashraf Mohammed Iqbal**
Dalhousie University and Capital Health
- **Ashraf Hamdy Owis**
Cairo University
- **Asoke Nath**
St. Xaviers College(Autonomous), 30 Park Street,
Kolkata-700 016
- **Ayad Ghany Ismaeel**
Department of Information Systems Engineering-
Technical Engineering College-Erbil Polytechnic
University, Erbil-Kurdistan Region- IRAQ
- **Ayman EL-SAYED**
Computer Science and Eng. Dept., Faculty of
Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**
University of Ibadan
- **Badre Bossoufi**
University of Liege
- **BASANT KUMAR VERMA**
JNTU
- **Basil Hamed**
Islamic University of Gaza
- **Basil M Hamed**
Islamic University of Gaza
- **Bhanu Prasad Pinnamaneni**
Rajalakshmi Engineering College; Matrix Vision
GmbH
- **Bharti Waman Gawali**
Department of Computer Science & information T

- **Bilian Song**
LinkedIn
- **Brahim Raouyane**
FSAC
- **Bright Keswani**
Associate Professor and Head, Department of
Computer Applications, Suresh Gyan Vihar
University, Jaipur (Rajasthan) INDIA
- **Brij Gupta**
University of New Brunswick
- **C Venkateswarlu Venkateswarlu Sonagiri**
JNTU
- **Chandrashekhar Meshram**
Chhattisgarh Swami Vivekananda Technical
University
- **Chao Wang**
- **Chao-Tung Yang**
Department of Computer Science, Tunghai
University
- **Charlie Obimbo**
University of Guelph
- **Chien-Peng Ho**
Information and Communications Research
Laboratories, Industrial Technology Research
Institute of Taiwan
- **Chun-Kit (Ben) Ngan**
The Pennsylvania State University
- **Ciprian Dobre**
University Politehnica of Bucharest
- **Constantin Filote**
Stefan cel Mare University of Suceava
- **Constantin POPESCU**
Department of Mathematics and Computer
Science, University of Oradea
- **CORNELIA AURORA Gyorödi**
University of Oradea
- **Dana - PETCU**
West University of Timisoara
- **Deepak Garg**
Thapar University
- **Dheyaa Kadhim**
University of Baghdad
- **Dong-Han Ham**
Chonnam National University
- **Dr K Ramani**
K.S.Rangasamy College of Technology,
Tiruchengode
- **Dr. Harish Garg**
Thapar University Patiala
- **Dr. Sanskruti V Patel**
Charotar Univeristy of Science & Technology,
Changa, Gujarat, India
- **Dr. Santosh Kumar**
Graphic Era University, Dehradun (UK)
- **Dr. JOHN S MANOHAR**
VTU, Belgaum
- **Dragana Becejski-Vujaklija**
University of Belgrade, Faculty of organizational
sciences
- **Driss EL OUADGHIRI**
- **Duck Hee Lee**
Medical Engineering R&D Center/Asan Institute for
Life Sciences/Asan Medical Center
- **Elena Camossi**
Joint Research Centre
- **Elena SCUTELNICU**
Dunarea de Jos University of Galati
- **Eui Chul Lee**
Sangmyung University
- **Evgeny Nikulchev**
Moscow Technological Institute
- **Ezekiel Uzor OKIKE**
UNIVERSITY OF BOTSWANA, GABORONE
- **FANGYONG HOU**
School of IT, Deakin University
- **Faris Al-Salem**
GCET
- **Firkhan Ali Hamid Ali**
UTHM
- **Fokrul Alom Mazarbhuiya**
King Khalid University
- **Frank AYO Ibikunle**
Botswana Int'l University of Science & Technology
(BIUST), Botswana.
- **Fu-Chien Kao**
Da-Y eh University
- **Gamil Abdel Azim**
Suez Canal University
- **Ganesh Chandra Sahoo**
RMRIMS
- **Gaurav Kumar**
Manav Bharti University, Solan Himachal Pradesh,
- **George Mastorakis**
Technological Educational Institute of Crete
- **George D. Pecherle**

- University of Oradea
- **Georgios Galatas**
The University of Texas at Arlington
- **Gerard Dumancas**
Oklahoma Baptist University
- **Ghalem Belalem Belalem**
University of Oran 1, Ahmed Ben Bella
- **Giacomo Veneri**
University of Siena
- **Giri Babu**
Indian Space Research Organisation
- **Govindarajulu Salendra**
- **Grebenisan Gavril**
University of Oradea
- **Gufran Ahmad Ansari**
Qassim University
- **Gunaseelan Devaraj**
Jazan University, Kingdom of Saudi Arabia
- **GYÖRÖDI ROBERT STEFAN**
University of Oradea
- **Hadj Hamma Tadjine**
IAV GmbH
- **Hamid Mukhtar**
National University of Sciences and Technology
- **Hamid Alinejad-Rokny**
The University of New South Wales
- **Hamid Ali Abed AL-Asadi**
Department of Computer Science, Faculty of Education for Pure Science, Basra University
- **Hany Kamal Hassan**
EPF
- **Harco Leslie Hendric SPITS WARNARS**
Surya university
- **Hazem I. El Shekh Ahmed**
Pure mathematics
- **Hesham G. Ibrahim**
Faculty of Marine Resources, Al-Mergheb University
- **Himanshu Aggarwal**
Department of Computer Engineering
- **Hossam Faris**
- **Huda K. AL-Jobori**
Ahlia University
- **Iwan Setyawan**
Satya Wacana Christian University
- **JAMAIAH HAJI YAHAYA**
NORTHERN UNIVERSITY OF MALAYSIA (UUM)
- **James Patrick Henry Coleman**
Edge Hill University
- **Jatinderkumar Ramdass Saini**
Narmada College of Computer Application, Bharuch
- **Jayaram A M**
- **Ji Zhu**
University of Illinois at Urbana Champaign
- **Jia Uddin Jia**
Assistant Professor
- **Jim Jing-Yan Wang**
The State University of New York at Buffalo, Buffalo, NY
- **John P Sahlin**
George Washington University
- **JOSE LUIS PASTRANA**
University of Malaga
- **Jyoti Chaudhary**
high performance computing research lab
- **K V.L.N.Acharyulu**
Bapatla Engineering college
- **Ka-Chun Wong**
- **Kashif Nisar**
Universiti Utara Malaysia
- **Kayhan Zrar Ghafoor**
University Technology Malaysia
- **Khin Wee Lai**
Biomedical Engineering Department, University Malaya
- **KITIMAPORN CHOOCHOTE**
Prince of Songkla University, Phuket Campus
- **Kohei Arai**
Saga University
- **Krasimir Yankov Yordzhev**
South-West University, Faculty of Mathematics and Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov Stefanov**
Professor at Sofia University St. Kliment Ohridski
- **Labib Francis Gergis**
Misr Academy for Engineering and Technology
- **Lazar Stošic**
Collegefor professional studies educators Aleksinac, Serbia
- **Leandros A Maglaras**
University of Surrey
- **Leon Andretti Abdillah**
Bina Darma University
- **Lijian Sun**

- Chinese Academy of Surveying and
- **Ljubomir Jerinic**
University of Novi Sad, Faculty of Sciences,
Department of Mathematics and Computer Science
- **Lokesh Kumar Sharma**
Indian Council of Medical Research
- **Long Chen**
Qualcomm Incorporated
- **M. Reza Mashinchi**
Research Fellow
- **M. Tariq Bandy**
University of Kashmir
- **Manas deep**
Masters in Cyber Law & Information Security
- **Manju Kaushik**
- **Manoharan P.S.**
Associate Professor
- **Manoj Wadhwa**
Echelon Institute of Technology Faridabad
- **Manpreet Singh Manna**
Associate Professor, SLIET University, Govt. of India
- **Manuj Darbari**
BBD University
- **Marcellin Julius Antonio Nkenlifack**
University of Dschang
- **Maria-Angeles Grado-Caffaro**
Scientific Consultant
- **Marwan Alseid**
Applied Science Private University
- **Mazin S. Al-Hakeem**
LFU (Lebanese French University) - Erbil, IRAQ
- **MD RANA**
University of Sydney
- **Md. Zia Ur Rahman**
Narasaraopeta Engg. College, Narasaraopeta
- **Mehdi Bahrami**
University of California, Merced
- **Messaouda AZZOUZI**
Ziane Achour University of Djelfa
- **Milena Bogdanovic**
University of Nis, Teacher Training Faculty in Vranje
- **Miriampally Venkata Raghavendra**
Adama Science & Technology University, Ethiopia
- **Mirjana Popovic**
School of Electrical Engineering, Belgrade University
- **Miroslav Baca**
University of Zagreb, Faculty of organization and informatics / Center for biometrics
- **Mohamed Ali Mahjoub**
Preparatory Institute of Engineer of Monastir
- **Mohamed A. El-Sayed**
Faculty of Science, Fayoum University, Egypt.
- **Mohamed Najeh LAKHOUA**
ESTI, University of Carthage
- **Mohammad Ali Badamchizadeh**
University of Tabriz
- **Mohammad Hani Alomari**
Applied Science University
- **Mohammad Azzeh**
Applied Science university
- **Mohammad Jannati**
- **Mohammad Haghighat**
University of Miami
- **Mohammed Shamim Kaiser**
Institute of Information Technology
- **Mohammed Sadgal**
Cadi Ayyad University
- **Mohammed Abdulhameed Al-shabi**
Associate Professor
- **Mohammed Ali Hussain**
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**
Universiti Tun Hussein Onn Malaysia
- **Mona Elshinawy**
Howard University
- **Mostafa Mostafa Ezziyani**
FSTT
- **Mourad Amad**
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**
University Malaysia Pahang
- **Murthy Sree Rama Chandra Dasika**
Geethanjali College of Engineering & Technology
- **Mustapha OUJAOURA**
Faculty of Science and Technology Béni-Mellal
- **MUTHUKUMAR S SUBRAMANYAM**
DGCT, ANNA UNIVERSITY
- **N.Ch. Sriman Narayana Iyengar**
VIT University,
- **Nagy Ramadan Darwish**
Department of Computer and Information Sciences,
Institute of Statistical Studies and Researches, Cairo University.

- **Najib A. Kofahi**
Yarmouk University
- **Natarajan Subramanyam**
PES Institute of Technology
- **Nazeeruddin - Mohammad**
Prince Mohammad Bin Fahd University
- **NEERAJ SHUKLA**
ITM UNiversity, Gurgaon, (Haryana) Inida
- **Nestor Velasco-Bermeo**
UPFIM, Mexican Society of Artificial Intelligence
- **Nidhi Arora**
M.C.A. Institute, Ganpat University
- **Ning Cai**
Northwest University for Nationalities
- **Noura Aknin**
University Abdelamlek Essaadi
- **Oliviu Matei**
Technical University of Cluj-Napoca
- **Om Prakash Sangwan**
- **Omaima Nazar Al-Allaf**
Asesstant Professor
- **Osama Omer**
Aswan University
- **Ousmane THIARE**
Associate Professor University Gaston Berger of
Saint-Louis SENEGAL
- **Paresh V Virparia**
Sardar Patel University
- **Poonam Garg**
Institute of Management Technology, Ghaziabad
- **Prabhat K Mahanti**
UNIVERSITY OF NEW BRUNSWICK
- **PROF DURGA PRASAD SHARMA (PHD)**
AMUIT, MOEFDRE & External Consultant (IT) &
Technology Tansfer Research under ILO & UNDP,
Academic Ambassador for Cloud Offering IBM-USA
- **Professor Ajantha Herath**
- **Qifeng Qiao**
University of Virginia
- **Rachid Saadane**
EE departement EHTP
- **Raed Kanaan**
Amman Arab University
- **Raghuraj Singh**
Harcourt Butler Technological Institute
- **Rahul Malik**
- **Raja Sarath Kumar Boddu**

- LENORA COLLEGE OF ENGINEERNG
- **Rajesh Kumar**
National University of Singapore
- **Rakesh Chandra Balabantaray**
IIIT Bhubaneswar
- **Rakesh Kumar Dr.**
Madan Mohan Malviya University of Technology
- **Rashad Abdullah Al-Jawfi**
Ibb university
- **Rashid Sheikh**
Shri Aurobindo Institute of Technology, Indore
- **Ravi Prakash**
University of Mumbai
- **Ravisankar Hari**
CENTRAL TOBACCO RESEARCH INSTITUE
- **Rawya Y. Rizk**
Port Said University
- **Reshmy Krishnan**
Muscat College affiliated to stirling University.U
- **Ricardo Ângelo Rosa Vardasca**
Faculty of Engineering of University of Porto
- **Ritaban Dutta**
ISSL, CSIRO, Tasmaniia, Australia
- **Ruchika Malhotra**
Delhi Technoogical University
- **SAADI Slami**
University of Djelfa
- **Sachin Kumar Agrawal**
University of Limerick
- **Sagarmay Deb**
Central Queensland Universiry, Australia
- **Said Ghoniemy**
Taif University
- **Sandeep Reddivari**
University of North Florida
- **Sasan Adibi**
Research In Motion (RIM)
- **Satyendra Prasad Singh**
Professor
- **Sebastian Marius Rosu**
Special Telecommunications Service
- **Seema Shah**
Vidyalankar Institute of Technology Mumbai,
- **Selem Charfi**
University of Pays and Pays de l'Adour
- **SENGOTTUVELAN P**
Anna University, Chennai

- **Senol Piskin**
Istanbul Technical University, Informatics Institute
- **Sérgio André Ferreira**
School of Education and Psychology, Portuguese Catholic University
- **Seyed Hamidreza Mohades Kasaei**
University of Isfahan,
- **Shafiqul Abidin**
Northern India Engineering College (Affiliated to GGS I P University), New Delhi
- **Shahanawaj Ahamad**
The University of Al-Kharj
- **Shaiful Bakri Ismail**
- **Shawki A. Al-Dubae**
Assistant Professor
- **Sherif E. Hussein**
Mansoura University
- **Shriram K Vasudevan**
Amrita University
- **Siddhartha Jonnalagadda**
Mayo Clinic
- **Sim-Hui Tee**
Multimedia University
- **Simon Uzezi Ewedafe**
Baze University
- **Siniša Opic**
University of Zagreb, Faculty of Teacher Education
- **Sivakumar Poruran**
SKP ENGINEERING COLLEGE
- **Slim BEN SAOUD**
National Institute of Applied Sciences and Technology
- **Sohail Jabbar**
Bahria University
- **Sri Devi Ravana**
University of Malaya
- **Sudarson Jena**
GITAM University, Hyderabad
- **Suhas J Manangi**
Microsoft
- **SUKUMAR SENTHILKUMAR**
Universiti Sains Malaysia
- **Sumazly Sulaiman**
Institute of Space Science (ANGKASA), Universiti Kebangsaan Malaysia
- **Sumit Goyal**
National Dairy Research Institute
- **Suresh Sankaranarayanan**
Institut Teknologi Brunei
- **Susarla Venkata Ananta Rama Sastry**
JNTUK, Kakinada
- **Suxing Liu**
Arkansas State University
- **Syed Asif Ali**
SMI University Karachi Pakistan
- **T C.Manjunath**
HKBK College of Engg
- **T V Narayana rao Rao**
SNIST
- **T. V. Prasad**
Lingaya's University
- **Taiwo Ayodele**
Infonetmedia/University of Portsmouth
- **Tarek Fouad Gharib**
Ain Shams University
- **Thabet Mohamed Slimani**
College of Computer Science and Information Technology
- **Totok R. Biyanto**
Engineering Physics, ITS Surabaya
- **Touati Youcef**
Computer sce Lab LIASD - University of Paris 8
- **Uchechukwu Awada**
Dalian University of Technology
- **Urmila N Shrawankar**
GHRCE, Nagpur, India
- **Vaka MOHAN**
TRR COLLEGE OF ENGINEERING
- **Vinayak K Bairagi**
AISSMS Institute of Information Technology, Pune
- **Vishnu Narayan Mishra**
SVNIT, Surat
- **Vitus S.W. Lam**
The University of Hong Kong
- **VUDA SREENIVASARAO**
PROFESSOR AND DEAN, St.Mary's Integrated Campus,Hyderabad.
- **Wei Wei**
Xi'an Univ. of Tech.
- **Xiaoqing Xiang**
AT&T Labs
- **Yi Fei Wang**
The University of British Columbia
- **Yihong Yuan**

University of California Santa Barbara

- **Yilun Shang**
Tongji University
- **Yu Qi**
Mesh Capital LLC
- **Zacchaeus Oni Omogbadegun**
Covenant University
- **Zairi Ismael Rizman**
Universiti Teknologi MARA
- **Zenzo Polite Ncube**
North West University

- **Zhao Zhang**
Department of EE, City University of Hong Kong
- **Zhixin Chen**
ILX Lightwave Corporation
- **Ziyue Xu**
National Institutes of Health, Bethesda, MD
- **Zlatko Stapic**
University of Zagreb, Faculty of Organization and Informatics Varazdin
- **Zuraini Ismail**
Universiti Teknologi Malaysia

CONTENTS

Paper 1: Cluster-Based Context-Aware Routing Protocol for Mobile Environments

Authors: Ahmed. A. A. Gad-ElRab, T. A. A. Alzohairy, Almohammady S. Alsharkawy

PAGE 1 – 10

Paper 2: Security Issues of a Recent RFID Multi Tagging Protocol

Authors: Mehmet Hilal Özcanhan, Sezer Baytar, Semih Utku, Gökhan Dalkılıç

PAGE 11 – 15

Paper 3: Model Driven Testing of Web Applications Using Domain Specific Language

Authors: Viet-Cuong Nguyen

PAGE 16 – 22

Paper 4: A Design of Pipelined Architecture for on-the-Fly Processing of Big Data Streams

Authors: Usamah Algemili, Simon Berkovich

PAGE 23 – 30

Paper 5: Review of Cross-Platforms for Mobile Learning Application Development

Authors: Nabil Litayem, Bhawna Dhupia, Sadia Rubab

PAGE 31 – 39

Paper 6: Fault-Tolerant Attitude Control System for a Spacecraft with Control Moment Gyros Using Multi-Objective Optimization

Authors: Ai Noumi, Misuzu Haruki, Takuya Kanzawa, Masaki Takahashi

PAGE 40 – 48

Paper 7: Golay Code Transformations for Ensemble Clustering in Application to Medical Diagnostics

Authors: Faisal Alsaby, Kholood Alnowaiser, Simon Berkovich

PAGE 49 – 53

Paper 8: A Monitoring Model for Hierarchical Architecture of Distributed Systems

Authors: Phuc Tran Nguyen Hong, Son Le Van

PAGE 54 – 62

Paper 9: Ipv6 Change Threats Behavior

Authors: Firas Najjar, Homam El-Taj

PAGE 63 – 68

Paper 10: Emotional Engagement and Active Learning in a Marketing Simulation: A Review and Exploratory Study

Authors: Kear Andrew, Bown Gerald Robin

PAGE 69 – 76

Paper 11: Modeling and Simulation Analysis of Power Frequency Electric Field of UHV AC Transmission Line

Authors: Chen Han, Yuchen Chen

PAGE 77 – 81

Paper 12: A Proposal of SNS to Improve Member's Motivation in Voluntary Community Using Gamification

Authors: Kohei Otake, Yoshihisa Shinozawa, Akito Sakurai, Makoto Oka, Tomofumi Uetake, Ryosuke Sumita

PAGE 82 – 88

Paper 13: Fast Vertical Mining Using Boolean Algebra

Authors: Hosny M. Ibrahim, M. H. Marghny, Noha M. A. Abdelaziz

PAGE 89 – 96

Paper 14: Investigation of Adherence Degree of Agile Requirements Engineering Practices in Non-Agile Software Development Organizations

Authors: Mennatallah H. Ibrahim, Nagy Ramadan Darwish

PAGE 97 – 103

Paper 15: A Review on Parameters Identification Methods for Asynchronous Motor

Authors: Xing Zhan, Guohui Zeng, Jin Liu, Qingzhen Wang, Sheng Ou

PAGE 104 – 109

Paper 16: An Intelligent Natural Language Conversational System for Academic Advising

Authors: Edward M. Latorre-Navarro, John G. Harris

PAGE 110 – 119

Paper 17: A Comparative Study of Thresholding Algorithms on Breast Area and Fibroglandular Tissue

Authors: Shofwatul 'Uyun, Sri Hartati, Agus Harjoko, Lina Choridah

PAGE 120 – 124

Paper 18: Using Heavy Clique Base Coarsening to Enhance Virtual Network Embedding

Authors: Ashraf A. Shahin

PAGE 125 – 132

Paper 19: Ontology Based SMS Controller for Smart Phones

Authors: Mohammed A. Balubaid, Umar Manzoor, Bassam Zafar, Abdullah Qureshi, Numairul Ghani

PAGE 133 – 139

Paper 20: Android Platform Malware Analysis

Authors: Khalid Alfalqi, Rubayyi Alghamdi, Mofareh Waqdan

PAGE 140 – 146

Paper 21: A Survey of Topic Modeling in Text Mining

Authors: Rubayyi Alghamdi, Khalid Alfalqi

PAGE 147 – 153

Paper 22: Effectiveness of Iphone's Touch ID: KSA Case Study

Authors: Ahmad A. Al-Daraiseh, Diana Al Omari, Hadeel Al Hamid, Nada Hamad, Rawan Althemali

PAGE 154 – 161

Paper 23: A Comparison of OSPFv3 and EIGRPv6 in a Small IPv6 Enterprise Network

Authors: Richard John Whitfield, Shao Ying Zhu

PAGE 162 – 167

Paper 24: Give a Dog ICT Devices: How Smartphone-Carrying Assistance Dogs May Help People with Dementia

Authors: Chika Oshima, Kiyoshi Yasuda, Toshiyuki Uno, Kimie Machishima, Koichi Nakayama

PAGE 168 – 176

Paper 25: Orientation Capture of a Walker's Leg Using Inexpensive Inertial Sensors with Optimized Complementary Filter Design

Authors: Sebastian Andersson, Liu Yan

PAGE 177 – 181

Paper 26: Technical Perspectives on Knowledge Management in Bioinformatics Workflow Systems

Authors: Walaa N. Ismail, M.Sabih Aksoy

PAGE 182 – 188

Paper 27: Public Transportation Management System based on GPSWiFi and Open Street Maps

Authors: Saed Tarapiah, Shadi Atalla

PAGE 189 – 194

Cluster-Based Context-Aware Routing Protocol for Mobile Environments

Ahmed. A. A. Gad-ElRab
Department of Mathematics
Faculty of Science
Al-Azhar University-Cairo,Egypt

T. A. A. Alzohairy
Department of Mathematics
Faculty of Science
Al-Azhar University-Cairo,Egypt

Almohammady S. Alsharkawy
Department of Mathematics
Faculty of Science
Al-Azhar University-Cairo,Egypt

Abstract—Mobile environment has many issues due to mobility, energy limitations and status changing over time. Routing method is an important issue and has a significant impact in mobile networks, whereas selecting the optimum routing path will reduce the wasting in network resources, reduce network overhead and increase network reliability and lifetime. To decide which path will achieve the networks objectives, we need to construct a new routing algorithm that uses context attributes of a mobile device such as available bandwidth, residual energy, connection number and mobility value. In this paper, we propose a new mobile nodes ranking scheme based on the combination of two multi-criteria decision making approaches, the analytic hierarchy process (AHP) and the technique for order performance by similarity to ideal solution (TOPSIS) in Fuzzy environments. The Fuzzy AHP is used to analyze the structure of the clusterhead selection problem and to determine weights of the criteria, while the Fuzzy TOPSIS method is used to obtain the final mobile node ranking value. By basing on node ranking, we propose a new cluster based routing algorithm select the optimal clusterheads and the best routing path. Our simulation results show that the proposed method increases the network accuracy and lifetime and reduces network overhead.

Keywords—Clustering; Context; FMCDM; Mobile and Routing

I. INTRODUCTION

Many mobile systems utilize the mobile device context such as current location, residual energy, time and user's activity to obtain the best services to the mobile user. The key objective of these systems is to significantly simplify computing devices usage by realizing the changing in entities status and the surrounding environments. Context-aware systems use the contextual information to clarify the current situation and adapt mobile systems to be suitable for both user and device requirements.

Exchanging data between mobile nodes in the network is one of the basic challenges in this environment. Utilizing a context in mobile devices is receiving considerable attention to meet these challenges. In context-aware systems, mobile applications can use the contextual information such as user's location, day time, nearby people and devices and user's activity in useful way to solve many mobile issues. One of the most important issues in mobile computing is how to evaluate mobile device. Evaluating mobile device using user information, device information and environmental information is very helpful in many mobile applications such as data management and routing data in mobile networks. So,

we can rely on the rank values to use the highest performance mobile devices to send data to other nodes in the network. This method will keep most of mobile resources as energy and will increase network lifetime. Using the context in routing data through the network paths will achieve a high accuracy mobile network, and will reduce network overhead. Routing data using cluster methods allows fast connection, topology management, better routing, improves network lifetime, routing delay, bandwidth consumption, and throughput.

The main objective of this research is to introduce a systematic evaluation model to help the actors in mobile computing for evaluating and selecting the optimal mobile node among a set of available alternatives (mobile nodes). Evaluating mobile node based on context is a multi-criteria decision making problem (DM), where many context attributes should be considered in the decision-making. *DM* processes involve a series of steps: identifying the problems, constructing the preferences, evaluating the alternatives and determining the best alternatives. *DM* is extremely intuitive when considering single criterion problems, since we only need to specify the alternative with the highest preference rating. However, when *DM* method evaluates alternatives with multiple criteria (context attributes), many problems will arise in the evaluation process such as criteria weights, preference dependence, and conflicts among criteria. These problems need to be overcome by more sophisticated methods. So, network clustering which is based on multi-criteria will achieve a high performance routing method in mobile environments.

Fuzzy decision making is a method to solve the complex *DM* problems in a fuzzy environment. This method can deal with the problem of evaluation and selection. In the real world, linguistic variable is used by human beings to make decisions. Classical *DM* method works only with exact and ordinary data without qualitative data. This research will use the linguistic variable to express reasonably situation that difficult to define such as available bandwidth, residual energy and device mobility factor, and then select the best alternatives for data management in the mobile environment using a cluster based routing protocol.

In this paper, we propose a new ranking scheme for mobile nodes that is based on the combination of two multi-criteria decision making approaches, the analytic hierarchy process (AHP) and the technique for order performance by similarity to ideal solution (TOPSIS) in Fuzzy environments. The Fuzzy AHP is used to analyze the structure of the clusterhead selection problem and to determine weights of the criteria,

while the Fuzzy TOPSIS method is used to obtain final mobile node ranking. Finally, based on the node ranking value, we propose a new cluster based routing algorithm for selecting optimal clusterheads and the best routing path.

The rest of the paper is organized as the following. Section 2 includes a detailed survey of the related work. Section 3 introduces Multi-Criteria Decision Making Approaches. Section 4 describes the proposed cluster based context-aware routing protocol (CBCA). Section 5 presents simulation and analysis of the experimental results. Finally, Section 6 concludes the paper.

II. RELATED WORK

Selecting the best clusterhead and discovering the efficient routing path are very important to achieve a high accuracy and reliable network in the mobile environments. So, many researchers have been worked to fulfill this purpose and many protocols have been introduced. In this section, we will review in briefly the previous proposed routing protocols schemes in mobile environments.

A. Routing Protocols in Mobile Environments

Routing is the process of transferring the packets between the networks or within the network from the source to the destination node. Routing is mainly done by specially configured nodes which are called routers and is often confused with the bridging techniques. By basing on network structure, routing methods are categorized as Proactive (Table Driven) Routing Protocols, Reactive (On Demand) Routing Protocols and Hybrid Routing Protocols [1]. In proactive routing such as *DSDV*, each node maintains one or more routing tables. Proactive protocols continuously learn the topology of the network by exchanging topological information among the network nodes. The differences among the protocols lie in their routing table structure, number of tables, updating frequency, use of control messages and the presence of a central node. In Reactive routing protocol such as *AODV*, *DSR* and *TORA* routes from source to destination doesn't exist. Whenever route is required each node discovers and maintains the route as and when required. In On-Demand routing protocol paths are explored only when needed. Hybrid routing protocols such as *ZRP* include the features of proactive and reactive routing protocols. Proactive tactic is used to discover and maintain routes to nearer nodes, while routes for far away nodes are discovered reactively. The author in [1] introduced a survey of routing algorithms for mobile networks.

Ad hoc On-demand Distance Vector Routing (*AODV*) [2] is very popular routing protocol that is based on classical distance vector routing algorithm. *AODV* is essentially a combination of both *DSR* [3] and *DSDV* [4]. It shares *DSR*'s on-demand characteristics hence discovers routes whenever it is needed via a similar route discovery process. However, *AODV* is loop-free due to the destination sequence numbers associated with routes. It creates routes only when they are needed, which reduces the periodic control message overhead which is associated with proactive routing protocols. *AODV* adapts traditional routing tables, one entry per destination which is in contrast to *DSR* that maintains multiple route cache entries for each destination. The initial design of *AODV* is

undertaken after the experience with *DSDV* routing algorithm. *AODV* also has other significant features. Whenever a route is available from source to destination, it does not add any overhead to the packets. However, route discovery process is only initiated when routes are not used and/or they expired and consequently discarded. This strategy reduces the effects of stale routes as well as the need for route maintenance for unused routes. *AODV* have the ability to provide unicast, multicast and broadcast communication. *AODV* uses a broadcast route discovery algorithm and uses the unicast in route reply message.

B. Cluster Based Routing Protocols

Mobile networks are characterized as dynamic topology, bandwidth and link capacity, nodes are energy constrained. Network cluster methods allow fast connection and topology management, better routing and also improve network lifetime, routing delay, bandwidth consumption and throughput. Each cluster in the network contains clusterhead (*CH*). The *CH* responsible to provide communication bridge between members and the other clusters. In the mobile environments, the topology changes dynamically. So, to achieve a high performance in the network, any clustering algorithm should operate with minimum overhead of cluster maintenance and try to preserve its structure as much as possible when nodes are moving and/or the topology is slowly changing. Many approaches for network clustering have been developed by researchers which focus on different performance metrics, most used weight metrics like average consumed power, residual energy, computing capabilities, distance with all neighbors, mobility and node trust value. Most of previous clustering approaches focus on some of this metrics to evaluate the network node. In clustered network each *CH* is responsible for the following jobs:

- Identify each node in the cluster (assign IDs).
- Calculate the path weight (cost) of sending/receiving data to all neighbor clusters.
- Communicate all mobile nodes in the cluster
- Define the path that will receive data from across the gateways, and report its gateway which is located on this path to receive data from the common cluster.
- Define routing table.
- Communicate with other clusters through the gateways.
- Send data to all cluster's member nodes.

Naeimi et. al. [5] has introduced taxonomy of *CH* selection. In this survey, the clusterhead selection is classified into self-organized schemes, assisted schemes and multi-factor evaluation schemes. Cluster based routing is a most convenient way to develop an efficient routing scheme in mobile environments. But it has to deal with several problems like, control overhead of cluster formation, maintenance, battery Power, stability of cluster, fairness, load balancing etc. So, authors in [6]. Summarize that to optimize the clusterhead election algorithm and to perform efficient cluster based routing in mobile environments, it is necessary to consider all metrics rather than focusing on particular metric.

Using clustering method in mobile network gives the network several advantages. These advantages introduce them as the most compatible routing protocols in these environments. We list some of these advantages as the following:

- Minimizing the total transmission energy.
- Balancing the energy-exhausting load among all nodes.
- Reducing the bandwidth demand and efficient use of limited channel bandwidth.
- Eliminating the redundant and highly route discovery process.
- Routing path limited to the clusterhead and gateways and thus generating small-size routing tables.
- Increasing the lifetime and scalability of the network.

To ensure that the selection of clusterhead achieve all network requirements and increase network lifetime. The selection of clusterhead must be a multi-criteria decision issue with complex inter-relation between factors. Barfunga et. al [7], introduced an Energy Efficient Cluster Based Routing Protocol. Also, Anitha et. al [8], proposed an enhanced cluster based routing protocol for mobile nodes, this protocol is aimed to prolonging the lifetime of the sensor networks by balancing the energy consumption of the nodes. Naeimi et. al. [5] has surveyed that there are two clusterhead selections which is based on Multi-Factor Evaluation Schemes, Analytical Hierarchy Process (AHP) [9] and Fuzzy Logic Controller (FLC). The AHP method is characterized by decomposing complex decision of CH selection into a hierarchy of more easily understood sub-problem using numerical values and the FLC characterized by smooth noise tolerance, adaptive modifiable rules, low cost and complexity, more flexible to variable range of applications. Therefore, using a multi-criteria decision making to evaluate mobile nodes according its context will increase network lifetime and decrease network overhead. In this paper, we will introduce a cluster based routing protocol that utilizes the context as a multi-criteria decision making problem to select the optimum clusterhead and select the best routing path. A detailed survey on cluster based routing protocols can be found in [10], [11], [12], [13].

III. MULTI-CRITERIA DECISION MAKING APPROACHES

A. Fuzzy Sets and Fuzzy Number

Zadeh (1965) introduced the Fuzzy Set Theory (FST) to deal with the uncertainty and ambiguous of data. A major contribution of FST is the capability of representing uncertain data. FST also allows mathematical operators and programming to be performed to the fuzzy domain. A Fuzzy Set (FS) is a class of objects with a continuum of grades of membership. Such a set is characterized by a membership function, which assigns to each object a grade of membership ranging "between" zero and one.

Fuzzy Set: A fuzzy set \tilde{A} . In a universe of discourse X is characterized by a membership function $\tilde{\mu}_A(x)$ which associates with each element x in X a real number in the

interval $[0, 1]$. The function value $\tilde{\mu}_A(x)$ is termed the grade of membership of x in \tilde{A} . L.A. Zadeh [14].

Triangular Fuzzy Number: A triangular fuzzy number \tilde{A} can be defined by a triplet (L, M, U) shown in Fig-1. The membership function $\tilde{\mu}_A(x)$ is defined in [15] as

$$\tilde{\mu}_A(x) = \begin{cases} 0 & x < L \\ \frac{x-L}{M-L} & L \leq x \leq M \\ \frac{x-U}{M-U} & M \leq x \leq U \\ 1 & x > U \end{cases} \quad (1)$$

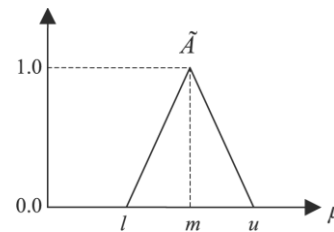


Fig. 1. Triangular Fuzzy Number (L, M, U)

A Symbol "~" will be placed above A if the A shows a FST. A Triangular Fuzzy Number (TFN) \tilde{A} TFN represented with three points as follows: (L, M, U) . L stand for the lower bound of the fuzzy number, and U stand for the upper bound. This representation is interpreted as membership functions and holds the following conditions.

- L to M is increasing function
- M to U is decreasing function
- $L \leq M \leq U$.

Fuzzy sets were originally proposed to deal with problems of subjective uncertainty. Subjective uncertainty results from using linguistic variables to represent the problem or the event, linguistic variable is a variable that is expressed by verbal words or sentences in a natural or artificial language. Linguistic variables are also employed as a way to measure the achievement of the performance value for each criterion. Since the linguistic variables can be defined by the corresponding membership function and the fuzzy interval. Linguistic variables were proposed in [16], For example, linguistic variables with triangular fuzzy numbers may take on effect values such as very high (very good), high (good), fair, low (bad), and very low (very bad). So, we can naturally manipulate the fuzzy numbers to deal with the FMADM problems. The membership function of linguistic variables represented in triangular fuzzy number showed in Fig-2.

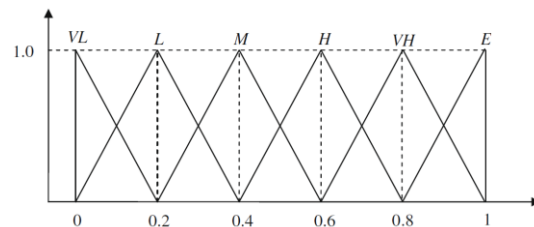


Fig. 2. Triangular fuzzy numbers of linguistic variables

B. Analytic Hierarchy Process (AHP)

Bernoulli (1738) proposed the concept of utility function to reflect human pursuit, such as maximum satisfaction, and von Neumann and Morgenstern (1947) presented the theory of game and economic behavior model, which expanded the studies on human economic behavior for multiple criteria decision making (MCDM) problems [16], an increasing amount of literature has been engaged in this field. The MCDM can be summarized in five main steps as follows:

- 1) Define the nature of the problem.
- 2) Construct a hierarchy system for its evaluation Fig-3.
- 3) Select the appropriate evaluation model.
- 4) Obtain the relative weights and performance score of each attribute with respect to each alternative.
- 5) Determine the best alternative according to the synthetic utility values, which are the aggregation value of relative weights, and performance scores corresponding to alternatives.
- 6) Outrank the alternatives referring to their synthetic fuzzy utility values from Step 5.

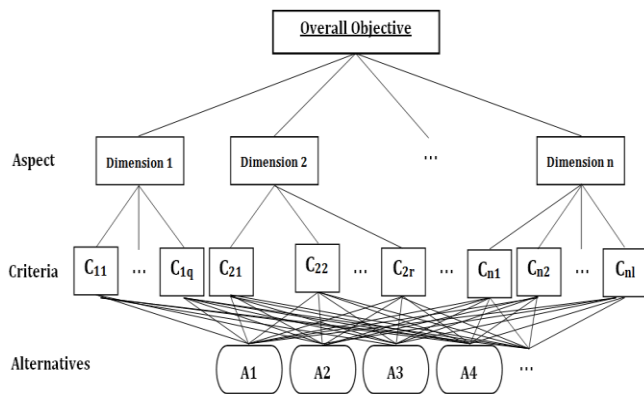


Fig. 3. Hierarchical system for MADM

The analytic hierarchy process (AHP) was proposed to derive the relative weights according to the appropriate hierarchical system. There are four methods, including the eigenvalue method, the geometric mean method, the linear programming method and the lambda-max method to derive the weights using the AHP. Only the eigenvalue method is employed to deal with crisp numbers and the other methods are adapted to handle the AHP under fuzzy numbers [16].

In AHP method, the pairwise comparisons for each level with respect to the goal of mobile evaluation are conducted using a nine-point scale. Each pairwise comparison represents an estimate of the priorities of the compared context attribute. The nine-point scale developed by Saaty (1980). Table I expresses preferences between options as equally, moderately, strongly, very strongly, or extremely preferred. These preferences are translated into pairwise weights of 1, 3, 5, 7, and 9, respectively, with 2, 4, 6, and 8 as intermediate values.

TABLE I. RATIO SCALE IN AHP (SAATY (1980))

Linguistic	Intensity Importance	Description
Equally important	1	Two factors contribute equally to the objective
Moderately more important	3	Experience and judgment slightly favor one over the other
Strongly more important	5	Experience and judgment strongly favor one over the other
Very strongly more important	7	Experience and judgment very strongly favor one over the other. Its importance is demonstrated in practice
Extremely more important	9	The evidence favoring one over the other is of the highest possible validity
Intermediate values	2, 4, 6, 8	When compromise is needed

C. Fuzzy AHP

The global weights for each candidate is determined and the candidates fuzzy priorities are calculated based on sub-factors using Linguistic variables, which are defined for the triangular fuzzy numbers, see Table II:

TABLE II. LINGUISTIC VALUES AND FUZZY NUMBERS

Linguistic values	Fuzzy numbers
Very low (VL)	(0, 0, 0.2)
Low (L)	(0, 0.2, 0.4)
Medium (M)	(0.2, 0.4, 0.6)
High (H)	(0.4, 0.6, 0.8)
Very high (VH)	(0.6, 0.8, 1)
Excellent (E)	(0.8, 1, 1)

The geometric mean method was first employed by Buckley (1985) to extend the AHP to consider the situation of using linguistic variables (Zadeh 1965). The degrees of the pairwise comparison of linguistic variables can be expressed using the fuzzy numbers see the following table. Table III.

TABLE III. THE PAIRWISE COMPARISON OF LINGUISTIC VARIABLES USING FUZZY

Intensity of fuzzy scale	Fuzzy numbers	Number user defined
$\tilde{1}$	(L,M,U)	(., 1 ,.)
$\tilde{3}$	(L,M,U)	(., 3 ,.)
$\tilde{5}$	(L,M,U)	(., 5 ,.)
$\tilde{7}$	(L,M,U)	(., 7 ,.)
$\tilde{9}$	(L,M,U)	(., 9 ,.)
$\tilde{2}, \tilde{4}, \tilde{6}, \tilde{8}$	(L,M,U)	(., . ,.)

From the information of the pairwise comparison, we can form the fuzzy positive reciprocal matrix as the following:

$$\tilde{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \quad (2)$$

Where $\tilde{a}_{ij} \odot \tilde{a}_{ji} \approx 1$ and $\tilde{a}_{ij} \cong \frac{w_i}{w_j}$

Then, the geometric mean method for finding the final fuzzy weights of each criterion can be formulated as the following:

$$\tilde{w}_i = \tilde{r}_i(\tilde{r}_1 \oplus \tilde{r}_2 \oplus \dots \oplus \tilde{r}_n)^{-1} \quad (3)$$

Where

$$\tilde{r}_i = (\tilde{a}_{i1} \odot \tilde{a}_{i2} \odot \dots \odot \tilde{a}_{in})^{1/n} \quad (4)$$

The fuzzy weights of each criterion can also be defuzzified by center of area (CoA) in order to obtain a crisp solution.

D. TOPSIS

The Technique for Order Preferences by Similarity to an Ideal Solution (TOPSIS) method was proposed by Hwang and Yoon (1981). The main idea came from the concept of the compromise solution to choose the best alternative which has the shortest Euclidean distance from the positive ideal solution (optimal solution) and farthest Euclidean distance from the negative ideal solution. The positive-ideal solution (PIS) is a solution that maximizes the benefit criteria and minimizes the cost criteria, whereas the negative ideal solution (NIS) maximizes the cost criteria and minimizes the benefit criteria. Then, choose the best one of sorting, which will be the best alternative. So, according to this technique we can evaluate mobile node based on the context.

E. Fuzzy TOPSIS

The use of numerical values (Crisp values) in the rating of alternatives may have limitations to deal with uncertainties and ambiguous. So, extensions of TOPSIS were developed to solve problems of decision making with uncertain data resulting in fuzzy TOPSIS. In practical applications, the triangular shape of the membership function is often used to represent fuzzy numbers. Fuzzy models using triangular fuzzy numbers proved to be very effective for solving decision-making problems where the available information is imprecise.

Given a set of alternatives, $A = \{A_k / k = 1, \dots, n\}$, and a set of criteria, $C = \{C_j | j = 1, \dots, m\}$, where $X = \{X_{kj} / k = 1, \dots, n; j = 1, \dots, m\}$ denotes the set of performance ratings and $w = \{w_j / j = 1, \dots, m\}$ is the set of weights, the information table $I = (A, C, X, W)$ can be represented as shown in TableIV. The first step of TOPSIS is to calculate normalized ratings by TableIV.

TABLE IV. TOPSIS INFORMATION TABLE I = (A, C, X, W)

Alternatives	C ₁	C ₂	...	C _m
A1	x ₁₁	x ₁₂	...	x _{1m}
A2	x ₂₁	x ₂₂	...	x _{2m}
⋮	⋮	⋮	⋮	⋮
An	a _{n1}	a _{n2}	...	a _{nn}
w	w ₁	w ₂	...	w _m

Now we will list the TOPSIS main steps as the following:

Step 1: Construct the normalized decision matrix

To transform the various attribute dimensions into non-dimensional attributes, which allows comparison across the

attributes. The first step of TOPSIS is to calculate normalized ratings using the following equation:

$$r_{kj}(x) = \frac{x_{kj}}{\sqrt{\sum_{k=1}^n x_{kj}^2}} \quad | k = 1, \dots, n; j = 1, \dots, m; \quad (5)$$

Step 2: Construct the weighted normalized decision matrix

- For benefit criteria (larger is better)

$$r_{kj}(x) = \frac{x_{kj} - x_j^-}{x_j^* - x_j^-} \quad (6)$$

Where $x_j^* = \max_k x_{kj}$ and $x_j^- = \min_k x_{kj}$ or setting x^* is the aspired/desired level and x^- is the worst level.

- For cost criteria (smaller is better)

$$r_{kj}(x) = \frac{x_j^- - x_{kj}}{x_j^- - x_j^*} \quad (7)$$

Step 3: Calculate weighted normalized ratings using

$$v_{kj}(x) = w_j r_{kj}(x) \quad | k = 1, \dots, n; j = 1, \dots, m; \quad (8)$$

Step 4: Obtain the positive ideal point (PIS) and the negative ideal point (NIS)

A^+ represents positive ideal point and A^- represents negative ideal point.

$$\begin{aligned} A^+ &= \{v_1^+, v_2^+, \dots, v_j^+, \dots, v_m^+\} \\ &= \{(\max_k v_{kj}(x), j \in J_1), (\min_k v_{kj}(x), j \in J_2) \\ &\quad | k = 1, \dots, n; \} \end{aligned} \quad (9)$$

$$\begin{aligned} A^- &= \{v_1^-, v_2^-, \dots, v_j^-, \dots, v_m^-\} \\ &= \{(\min_k v_{kj}(x), j \in J_1), (\max_k v_{kj}(x), j \in J_2) \\ &\quad | k = 1, \dots, n; \} \end{aligned} \quad (10)$$

Where J_1 and J_2 are the benefit and the cost attributes, respectively

Step 5: Calculate the separation from the PIS and the NIS between alternatives.

The separation values can be measured using the Euclidean distance, which is given as: D_k^* Positive Ideal Separation and D_k^- negative Ideal Separation.

$$D_k^* = \sqrt{\sum_{j=1}^m [v_{kj}(x) - v_j^+]^2} \quad , \quad k = 1, \dots, n; \quad (11)$$

$$D_k^- = \sqrt{\sum_{j=1}^m [v_{kj}(x) - v_j^-]^2} \quad , \quad k = 1, \dots, n; \quad (12)$$

Step 6: Calculate the Relative Closeness to the Ideal Solution

$$C_k^* = \frac{D_k^-}{D_k^* + D_k^-} \quad (13)$$

Where $C_k^* \in [0, 1] \quad \forall k = 1, \dots, n;$

Finally, the preferred orders can be obtained according to the similarities to the PIS (C_k^*) in descending order to choose the best alternatives.

IV. PROPOSED CLUSTER BASED ROUTING PROTOCOL

The proposed model of clusterhead selection problem combines of two *MCDM* approaches *FAHP* and *FTOPSIS* approaches to evaluate mobile node which is based on different context attributes as mobility, available bandwidth, residual energy and number of neighbors. The main difference between our proposed clusterhead selection method and other clustering algorithms that other methods rely on only one or two factors to complete the clustering process which is not sufficient to increase network lifetime and solve network overhead problem. Also, these methods ignore other factors which affect the network lifetime. This section will illustrate in details the new cluster based context-aware (CBCA) routing protocol. So, we offer a detailed explanation of the new protocol as the following.

A. Setup Phase

Setup phase consists of three levels to obtain the final results of ranking mobile nodes. In the first level: The Context-aware Middleware responsible for collecting the contextual information that will represent the criteria. The second level: Obtain the best weight for each criteria using *FAHP*. Finally, the third level: Responsible for using *FTOPSIS* to evaluate the alternatives and determine the final mobile node rank value.

Each mobile node collects the needed context for the evaluation operation. According to Gad-EIRab [17], the context-aware middleware is committed to support each mobile node by the contextual information which is needed in this operation. The context-aware middleware will retrieve context from its source and assign a new value to the mobile node. The source node broadcasts a hello message to all nodes in the network to inform each node to share its context with other nodes. This message contains a specified context attributes (mobility value, bandwidth, energy level and number of its neighbor nodes). These attributes will be used to evaluate each mobile node. All nodes that receive the hello message from source node will replay and send its context values to the source node or any other specified node in the network.

Now it's a time to start Device Dependent Context Rank (*DDCR*) operation. From [17], Device Dependent Context represents any contextual information that characterizes the device such as processing capabilities, energy level, available bandwidth, input sensors, visualization capabilities etc. So, each node will be evaluated based on its device dependent context using *FAHP* and *FTOPSIS* approaches. The *FAHP* method will compute the weight of each context attribute which will contribute in computing *DDCR*, note that computing weight is a pre-calculated by the application for one time only at the first of establishing the network. Fig-4 illustrates the decision hierarchy of mobile nodes evaluation process. The *FTOPSIS* will compute the final evaluation value of each node based on weight value from *FAHP* method and the received context from all nodes in the network. The source starts to normalize rating values which are received from each node and calculates weighted normalized ratings for all criteria. Then, it calculates positive and negative ideal point (A^+ , A^-). After that, the source node will determine Positive Ideal

Separation and Negative Ideal Separation D^+ , D^- for each node. Finally, it will determine the global context evaluation. At this point, the source obtains a rank value for each node in the network. After that, the source node broadcasts a message to inform each node by its own rank value. Then cluster formation phase will start.

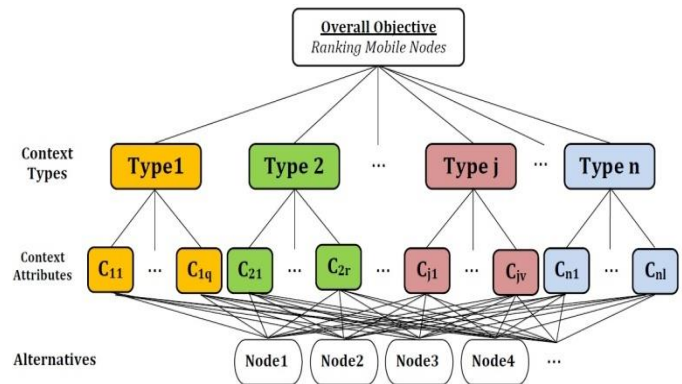


Fig. 4. Hierarchical system for Mobile Nodes Ranking

B. Cluster Formation Phase

In cluster formation phase, the network is partitioned into a number of clusters, each cluster contains one clusterhead *CH*. The *CH* is a node which has the largest *DDCR* value in the cluster, the remaining nodes in the cluster called member nodes. To start cluster formation phase, we have a number on evaluated nodes and each one knows its *DDCR* value, each node will start to check if any of its neighbor have rank value larger than its rank. If there is one or more; then the node will choose the largest of them and this selected node will be a parent of this node, and this node will be a member node in this cluster. If a node has the largest *DDCR* among all neighbors, this node will become *CH* of its cluster. Each selected *CH* sends an advertisement message to all nodes that belong to the cluster, this message is a small message containing the *CH* ID and the IDs of all nodes in its cluster. If there is a node has the largest *DDCR* from all its neighbor and does not select as a *CH*; then this node will select the nearest and the largest neighbor to be its parent. Until now, we have a number of clusters and each cluster contains one node which is called *CH*, this *CH* has the largest *DDCR* in the cluster and each cluster consists of K hops. A new problem arises after forming clusters which is draining of *CH* resources. The *CH* is responsible for many jobs in its cluster, these jobs consume the *CH* resources. So, to decrease this drains of *CH*s, each *CH* will elect number of nodes in its cluster to help it to accomplish some jobs such as communicating with other clusters, do some calculations and data collection, etc. The selected nodes are called *ViceCH* and any *ViceCH* falls on the first hop of its *CH*, and has the highest *DDCR* between neighbor, has the lowest connectivity, does not belong to any path to source node or other *CH*s. The *CH* of any cluster may have a *ViceCH* or not. The gateway nodes of any cluster will be authorized from the *CH* by handling any operation from any other clusters, such as calculating path weight between two clusters or other jobs. Fig-5 illustrates the cluster formation phase.

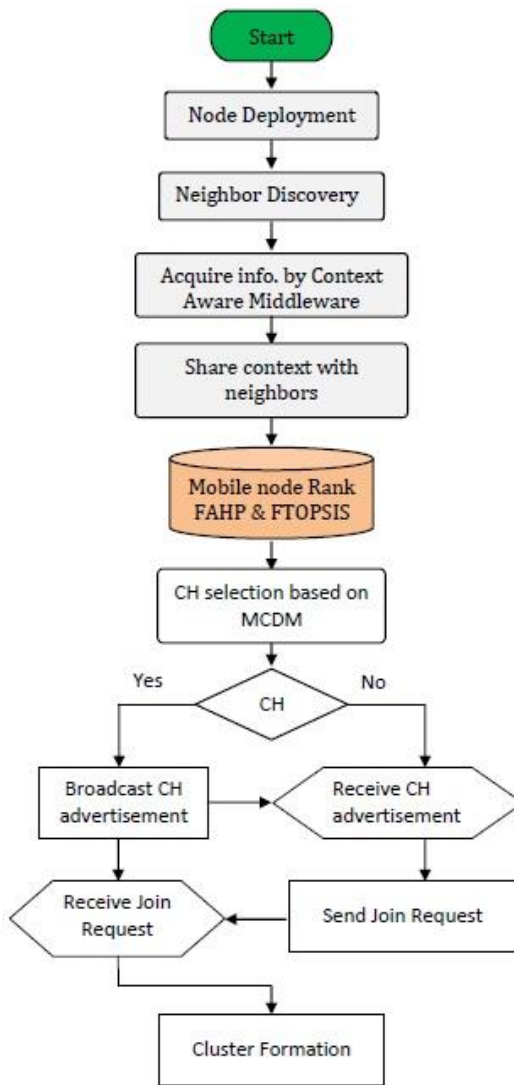


Fig. 5. Flow Diagram of Proposed Clustering Method

C. Cluster Maintenance Phase

Cluster maintenance phase will solve the problem of load balancing in network clusters. Each *CH* will calculate the average *DDCR* of its cluster. If the average *DDCR* is smaller than any of its neighbor clusters, then the cluster will discard the extra low *DDCR* nodes for the benefit of the largest neighbor cluster if it's possible. In the discrediting operation, each common node in the cluster knows the average *DDCR* value and the average *DDCR* value of its neighbor clusters; then the common node will join to the largest average *DDCR* value of its neighbor if its *DDCR* is very small.

If any member node separated from its cluster, it will attend to the nearest and the largest *DDCR* neighbor cluster. If the gateway node separated from its cluster, the *CH* will change its receiving path using the presented scenario in the previous discussion. If the separated node is a *CH*, its member nodes will try to join to the nearest cluster using the proposed scenario. If a new node (unranked node) adds to the network, it will join to the nearest cluster without computing *DDCR*.

D. Determining Cluster's Gateways

The gateway node is defined as a common node between two different clusters, if there is a routing path between two clusters, the sent message will pass through the common node that is called gateway. Each gateway knows the number of nodes in each common cluster and all paths to the common clusters. The gateway that is common to more than two clusters is worthies to handle communication of the common clusters. If the gateway node is connected to more than one node in the other cluster, it will select the highest *DDCR* and nearest gateway to make a path to the other *CH*. The *CH* is worthier to send the message to other cluster if it was directly connected to another cluster (the node is *CH* and gateway at the same time). If there is more than one gateway to the same cluster, the *CH* will select the largest *DDCR* node to be a gateway. Each *CH* sends data first to all used gateways after finishing the *CH* distributes data to all member nodes which sorted according to *DDCR* values.

E. Discovering Routing Path

In this phase, the network will discover the best routing path from source to destination, determining routing path begins by computing all paths weights (cost) between source node and destination node. The routing path cost will be derived from the average *DDCR* of each path A_r and its number of hops (Delay cost) D , as the following:

$$R_{c_i} = \frac{1}{D_i} w_1 + A_{r_i} w_2, \quad i = 1, \dots, p; \quad (14)$$

Such that w_1 represents the weight of delay cost D , w_2 represents the weight of A_r of each path, and p is the number of discovered paths. The network application will commit to determine the values of w_1 and w_2 such that:

$$w_1 + w_2 = 1 \quad (15)$$

In this paper we use values of w_1 and w_2 as, $w_1 = 0.5$ and $w_2 = 0.5$. The routing path that has *maximum* R_c will be selected to be the path between source and destination nodes.

V. SIMULATION RESULTS

In this section, we present the simulation results of the comparison between the proposed cluster based context-aware CBCA routing protocol and the standard routing protocol AODV [2]. We implemented the proposed protocol using the OMNET++ simulator [18].

A. Performance Metrics

We used many ways to study the proposed algorithm. The performance of CBCA protocol evaluated according to the following metrics:

1) **Average Packet Delivery Ratio:** It is the ratio of the number of successfully received packets to the total number of packets sent.

2) **Average end-to-end delay:** The end-to-end delay is averaged over all surviving data packets from the sources to the destinations.

3) **Control overhead:** The control overhead is defined as the total number of routing control packets normalized by the total number of received data packets.

Due to continuous changes in the topology of the mobile network. We generated different network scenarios for number of nodes, bandwidth and number of messages. Also, we used Random Waypoint mobility to model a mobility of nodes. Table V shows our simulation parameters.

TABLE V. SIMULATION PARAMETERS

Parameter	Value
Network Area	1000m x 1000m
Number of Nodes	25 - 250
Initial Energy	0.5 J
Mobility Type	RandomWPMobility
Radio Transmission Range	250m
Bandwidth	3 - 30 Kbps

As we mentioned in section 3, we can obtain the weight value of each criteria using FAHP method. This can be done through pairwise comparisons by asking how much the importance of a criterion compared to another criterion. By using this method, we can deduce all required weight values. So, a mobility weight was 0.074, bandwidth weight was 0.486, energy weight was 0.324 and connection number weight was 0.191. The simulation will use these values to obtain final mobile nodes rank value.

B. Simulation Results and Analysis

In this section, we will discuss the routing protocol simulation results and compare the proposed CBCA protocol and AODV protocol based on the above mentioned performance metrics.

Fig-6, compares the percentage of packet delivery ratio (PDR) for CBCA and AODV. As shown in Fig-6 PDR decreases as the number of nodes increases. We can see that the packet delivery ratio of CBCA protocol is clearly higher than the AODV protocol and our algorithm can scale up to larger network.

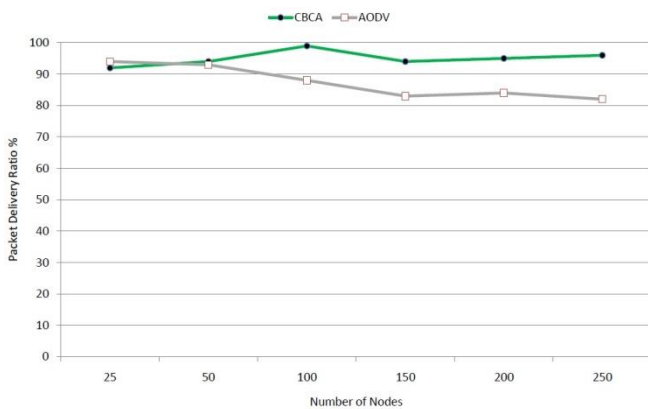


Fig. 6. Packet Delivery Ratio vs. Number of Nodes

The comparison of the end-to-end delay is shown in Fig-7. We can see that as the number of nodes increases, the average end-to-end delay increases, because more connections and congestions appear in higher density network. It can also be

concluded that the average end-to-end delay for CBCA protocol is better than the AODV protocol.

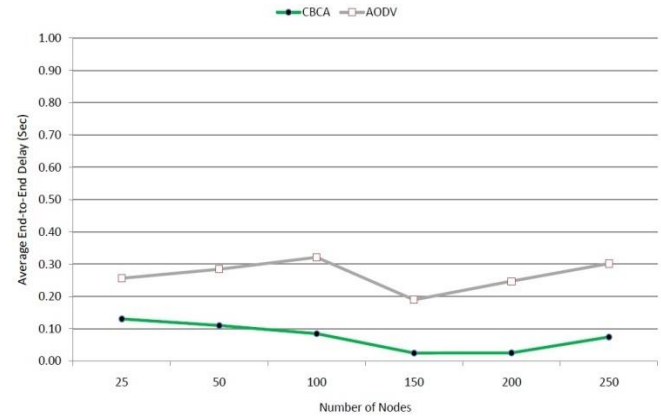


Fig. 7. Average End-to-End Delay vs. Number of Nodes

As shown in Fig-6 and Fig-7, the CBCA protocol enhances the PDR and end-to-end delay because in the CBCA, the route discovery phase and routing packets based on the high performance nodes in the network (the CHs and the Gateways). The CHs and the Gateways work with higher bandwidth and they have high number of connection in the network. These features minimize the delivery time from source node to destination and maximize packets delivery ratio in all network.

Fig-8, compares the packets overhead for CBCA and AODV. As shown in Fig-8, the packets overhead increases as the number of nodes increases. Also, we can ensure that the control overhead is less for CBCA when it is compared to AODV. So, CBCA protocol is more efficient in larger network.

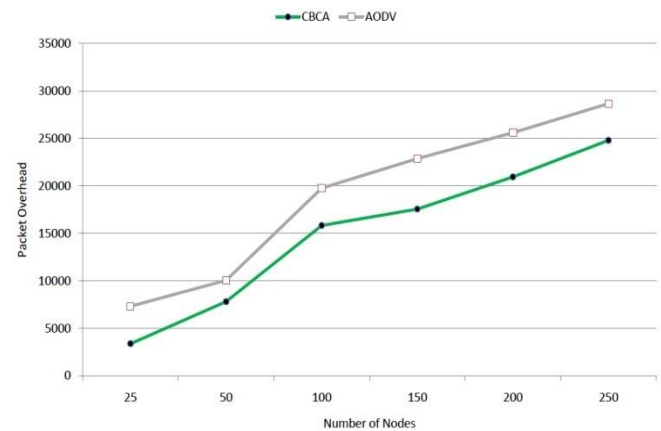


Fig. 8. Packet Overhead vs. Number of Nodes

Fig-9 shows the comparison of packet delivery ratio for CBCA and AODV in different bandwidth. As shown in Fig-9 PDR increases as the bandwidth of nodes increases. So, the number of packet drops also decreases. However, the proposed CBCA protocol achieves good delivery ratio, compared to AODV, which means that our approach has better performance.

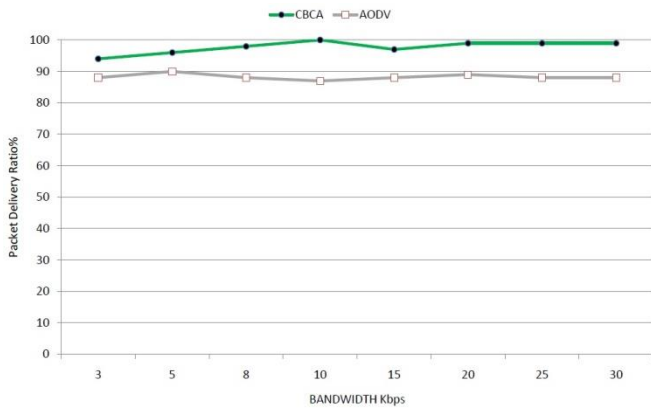


Fig. 9. Packet Delivery Ratio vs. Node Bandwidth

As shown in Fig-10, the average end-to-end delay of the proposed CBCA protocol is lower than the AODV protocol. This is because CBCA routing protocol takes into account node bandwidth and it needs smaller route discovery time than AODV.

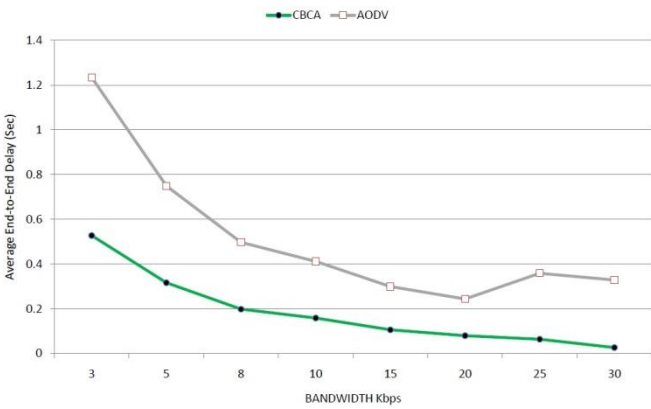


Fig. 10. Average End-to-End Delay vs. Node Bandwidth

Fig-11, shows a comparison between Packet Delivery Ratio and Number of Messages. As shown in Fig-11, PDR decreases as the number of messages increases. We can see that the packet delivery ratio of CBCA protocol is clearly higher than the AODV protocol.

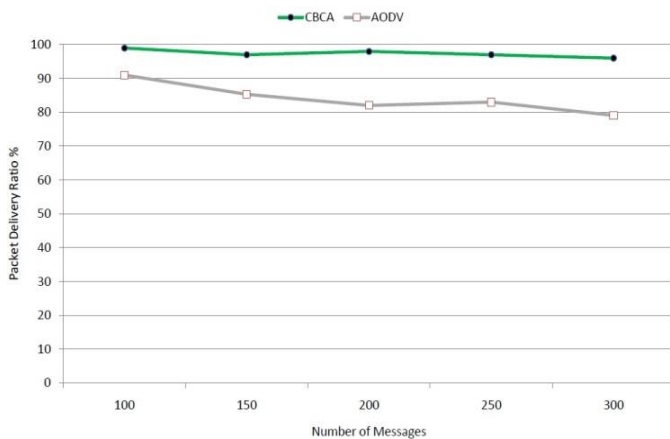


Fig. 11. Packet Delivery Ratio vs. Number of Messages

Fig-12, shows the comparison of the end-to-end delay between CBCA protocol and AODV protocol. We can see that as the number of messages increases, the average end-to-end delay increases. This is because the increasing in the number of messages leads to network congestion, which increases the postponement of sending the messages.

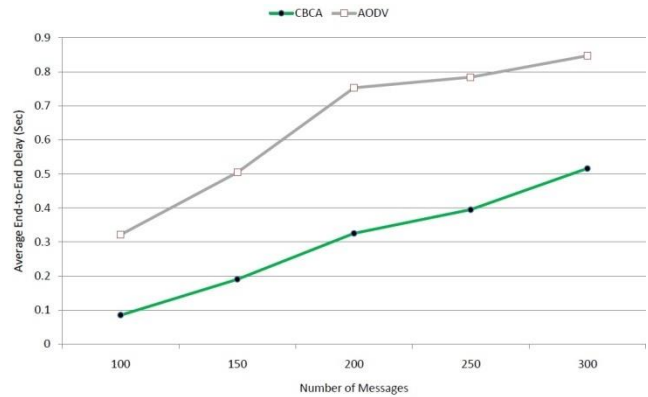


Fig. 12. Average End-to-End Delay vs. Number of Messages

Fig-13 shows the comparison of the packet overhead for CBCA and AODV with various number of messages. For both CBCA and AODV there is increasing in packet overhead with the increases of message number. CBCA routing provides smaller packet overhead than AODV. We can observe that CBCA has small increasing rate. This is because in the cluster formation process, a lot of control packets are exchanged. Also, the proposed CBCA protocol uses small number of nodes in route discovery phase.

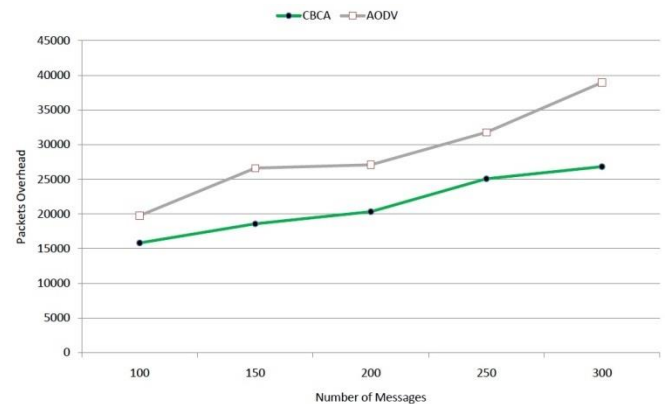


Fig. 13. Packet Overhead vs. Number of Messages

Fig-14 shows the average delay time in different weights of hops number w_l as we explained in equation (14) in route path selection process. As shown in this figure, When the weight of hops number was 0.3 to 0.8 given us the best results in average delay time.

VI. CONCLUSION

In this paper, we introduced the most relevant routing protocols types in mobile environments. Also, we discussed cluster based routing protocols issues. In addition, we proposed a new context based routing protocol in mobile environment. The new CBCA protocol is based on ranking network's nodes

according to its Device Dependent Context *DDCR*. The evaluation process is relying on two *MCDM* approaches

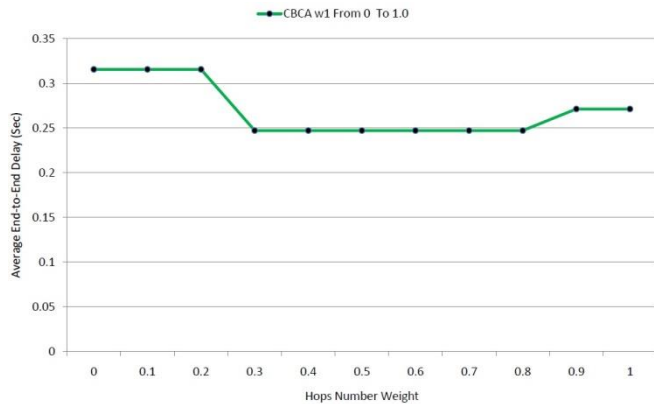


Fig. 14. Average End-to-End Delay vs. Hops Number Weight

FAHP and *FTOPSIS* to determine the mobile node ranking value. Then, we proposed a network clustering schema based on the previous ranking process. Finally, we introduced a routing discovery model to obtain the optimum routing path between source and destination, the best routing path that has the maximum average rank and minimum hops number. The performance of the new protocol has been evaluated through extensive simulation with various network sizes, bandwidth rate and number of messages. The simulation results demonstrate that there is a significant improvement in packet delivery ratio and the average end to end delay over traditional routing protocol AODV, and better performance than other routing algorithms in literature as well. So, the proposed CBCA protocol can increase the network lifetime and decrease network overhead. Which achieve the reliability and accuracy to the network in these environments.

Our future work involves using CBCA protocol to develop a new adaptive broadcasting protocol in mobile environments, based on user dependent context to reduce the network overhead and deliver the information to the user who will utilize and take care of the shared information.

REFERENCES

[1] Mamta Dhanda; Shikha Chaudhry. Survey of routing protocols for mobile ad hoc networks. International Journal of Advanced Research in Computer Science and Software Engineering, 3(4):1026–1031, 2013.
[2] Charles E Perkins and Elizabeth M Royer. Ad-hoc on-demand distance vector routing. In Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA'99. Second IEEE Workshop on, pages 90–100. IEEE, 1999.

[3] David B Johnson and David A Maltz. Dynamic source routing in ad hoc wireless networks. In Mobile computing, pages 153–181. Springer, 1996.
[4] Charles E Perkins and Pravin Bhagwat. Highly dynamic destination-sequenced distance-vector routing (dsvd) for mobile computers. In ACM SIGCOMM Computer Communication Review, volume 24, pages 234–244. ACM, 1994.
[5] Soroush Naeimi, Hamidreza Ghafghazi, Chee-Onn Chow, and Hiroshi Ishii. A survey on the taxonomy of cluster-based routing protocols for homogeneous wireless sensor networks. Sensors, 12(6):7350–7409, 2012.
[6] S. Mehta, P. Sharma, and K. Kotecha. A survey on various cluster head election algorithms for manet. In Engineering (NUICON), 2011 Nirma University International Conference on, pages 1–6. IEEE, Dec 2011.
[7] S.P. Barfunga, P. Rai, and H.K.D. Sarma. Energy efficient cluster based routing protocol for wireless sensor networks. In Computer and Communication Engineering (ICCCE), 2012 International Conference on, pages 603–607, July 2012.
[8] R.U. Anitha and P. Kamalakkannan. Enhanced cluster based routing protocol for mobile nodes in wireless sensor network. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on, pages 187–193, Feb 2013.
[9] Yaoyao Yin, Juwei Shi, Yinong Li, and Ping Zhang. Cluster head selection using analytical hierarchy process for wireless sensor networks. In Personal, Indoor and Mobile Radio Communications, 2006 IEEE 17th International Symposium on, pages 1–5, Sept 2006.
[10] M. Patil and R.C. Biradar. A survey on routing protocols in wireless sensor networks. In Networks (ICON), 2012 18th IEEE International Conference on, pages 86–91, Dec 2012.
[11] D. Sivakumar, B. Suseela, and R. Varadharajan. A survey of routing algorithms for manet. In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on, pages 625–640, March 2012.
[12] P. Kumari, M.P. Singh, and P. Kumar. Survey of clustering algorithms using fuzzy logic in wireless sensor network. In Energy Efficient Technologies for Sustainability (ICEETS), 2013 International Conference on, pages 924–928, April 2013.
[13] H. Kiwan and Y.L. Morgan. Hierarchical networks: Routing and clustering (a concise survey). In Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on, pages 1–4, May 2013.
[14] Zadeh Lotfi A. Fuzzy sets. Information and control, 8(3):338–353, 1965.
[15] Kaufmann ; Arnold, Gupta ; Madan M, and Kaufmann ; A. Introduction to fuzzy arithmetic: theory and applications. Van Nostrand Reinhold Company New York, 1985.
[16] G. H Tzeng and Jih-Jeng Huang. Multiple Attribute Decision Making: Methods and Applications. CRC Press, 2011.
[17] Ahmed. A. A. Gad-ElRab; T. A. A. Alzohairy; Almohammady S. Alsharkawy. Probabilistic context retrieval time-based algorithms in dynamic environments. International Journal of Current Engineering and Technology, 4(3):1633–1642, 2014.
[18] OMNET++ OpenSim Ltd. <http://www.omnetpp.org>.

Security Issues of a Recent RFID Multi Tagging Protocol

Mehmet Hilal Özcanhan
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey

Sezer Baytar
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey

Semih Utku
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey

Gökhan Dalkılıç
Department of Computer Engineering
Dokuz Eylul University
Izmir, Turkey

Abstract—RFID is now a widespread method used for identifying people and objects. But, not all communication protocols can provide the same rigorous confidentiality to RFID technology. In return, unsafe protocols put individuals and organizations into jeopardy. In this paper, a scheme that uses multiple low cost tags for identifying a single object is studied. Through algebraic analysis on chronologically ordered messages, the proposed multi tag arrangement is shown to fail to provide the claimed security. The weaknesses are discussed and previously proven precautions are recommended to increase the security of the protocol, and thus the safety of its users.

Keywords—Authentication; EPC Gen 2; ISO 18000-6; NFC; RFID; UHF tag

I. INTRODUCTION

Radio Frequency Identification (RFID) is the second widespread tool used in object identification and tracking, after paper barcodes. But, barcodes require a line of sight and can identify only one object at a time. Meanwhile, RFID does not require line of sight and as many as hundreds of objects can be identified within a second [1]. Therefore, it is not surprising to see RFID gradually replacing traditional barcodes in one of the biggest chain stores of the U.S.A. [2]. RFID has also proven itself in analysis of animal behavior [3], anti-counterfeiting [4], business automation [5], asset management [6], and recently in healthcare [7]. Indications are such that RFID will be one of the leading identification tools, in the near future.

Simply, RFID is a set-up of an electronic identification sticker (tag), a reader and a server. The tag has an integrated circuit with a unique identification number (ID) in its memory. An antenna attached to the integrated circuit is used to energize it through electromagnetism. The reader supplies the required electromagnetic energy to activate the tag. After activating the tag, the reader requests the ID of the tag [8]. A tag energized through the reader's electromagnetic field is called a passive tag. Other battery operated tags are called active tags and are not within the scope of the present work. In this study, a special type of passive tags - the low cost Ultra-High Frequency (UHF) tags - that are preferred due to their

long reading distance are focused on. Unfortunately, their limited resources cause UHF tags to lack strong security primitives. Capturing the Electronic Product Code (EPC, i.e. the ID) of some tags is very easy [9]. It is possible to track an item with an exposed ID, anywhere it goes on earth [1]. Therefore, it is necessary to look for a standard beyond the security supported in the ISO 18000-6 [10] and EPC Global Class 1 Generation 2 version 2 (Gen-2) [11] standards of the UHF tags. But, it should be noted that high security levels increase the cost of the tags. Therefore, the common goal of the researchers is to obtain a method with a balanced cost – security ratio.

In the rest of this paper, Section 2 summarizes previous work. Section 3 demonstrates weaknesses of a latest proposal. Section 4 contains authentication and security analysis of the proposal and four correction recommendations. In Section 5, the main conclusions and future work are presented.

II. RELATED WORK

Being pervasive yet insecure, early UHF tags have triggered many authentication proposals to be made. The proposals have been categorized according to the functions used for obscuring the tag ID [12]. The proposed protocols are grouped under four categories:

- Ultra-lightweight: Support only bitwise operation functions like AND, OR, XOR (\oplus), Shift, Rotate etc.
- Lightweight: Support random number generation and simple functions like cyclic redundancy check (CRC), but not hash functions.
- Simple: Support random number generation and one-way hash functions.
- Fully-fledged: Support conventional cryptographic functions.

Lately, researchers tried to stretch the boundaries between the neighboring categories. The categorization arguments gradually subsided and the attention was turned towards implementation of “lightweight” versions of hash and

cryptographic functions [13]. But, most proposals involve the authentication of a single tag, identifying a single object. There are of course the grouping proof protocols of multiple tags [14], but still each object is identified by a single tag.

Recently, identifying an object with multiple tags based on an ultra-lightweight authentication protocol has been proposed [15]. The proposal will be named Dhal and Gupta's Multi-Tag Authentication Protocol (DGMTAP). DGMTAP places multiple tags on an object as in Figure 1, each with an individual secret shared with the server. As always, the ultimate security goal is preventing the capture of the ID or the shared secret of the tag. The authors claim that DGMTAP resists known RFID attacks of listening adversaries.

Using the notation of Figure 1, m number of objects are marked by n number of tags. Each tag's index IN_j , shared secret key SK_j (2b bits long), old and new ID_j^{old} , ID_j^{new} are in the server's database. The index provides fast access to the tag record. The protocol assumes that the reader-server channel is secure, but the tag-reader channel is not. Therefore, the attackers can only use r_r (2b bits long), IN_j , M_j , $P1_j$, $P2_j$ that go between the reader and the tags. The equations and functions (Figure 1) used in the protocol are public; therefore available to malicious users as well. The mutual authentication of the server and the tag proceeds as follows: The reader triggers an identification session by sending a request and a random number (nonce) to a tag. Nonces are used for message freshness. No other secret or data is shared with the reader.

The server has all the information of the tags in an indexed database, as shown in Figure 1. One or multiple tags receiving the request, prepare their version of message M_j (equation 1), using their own secret SK_j . Next, M_j is sent to the reader preceded by the tag's index IN_j . The reader acts as a mediator to relay the replies of the tags together with its nonce, to the server. Using the index of the tag, the server finds the shared secret key SK_j of the tag and uses it with r_r in equation 2 to extract $(ID_j' - r_j || r_j)$. The apostrophe sign indicates that this is the received value. From here, the concatenated tag nonce r_j is obtained. With r_j , the server calculates $(ID_j^{new} - r_j)$ and checks if it equals the received $(ID_j' - r_j)$ value. If it is a match, the tag is authenticated and the object is identified. If not, the server checks if $(ID_j^{old} - r_j)$ equals $(ID_j' - r_j)$ value. If it is a match, the tag is authenticated and the object is identified. If not, the tag is rejected. After tag authentication is complete, a new tag $ID_j^{new'}$ is calculated and sent to the tag via the reader, hidden in messages $P1_j$, $P2_j$. The reader merely relays the messages together with the tag index. Tags check the index to decide if the broadcast is intended for itself. If it is, the tag carries out the XOR operation on $P1_j$ (equation 6). Next, the tag obtains tag $ID_j^{new'}$ by adding its own nonce to the result of equation 6 (equation 7). Using $ID_j^{new'}$, the tag analyses message $P2_j$ to verify if the sent ID_j' matches its present ID_j (equations 8 and 9). If it is a match, authentication of the server is complete and the tag saves the new tag $ID_j^{new'}$. The tag finishes and does not acknowledge the server about the completion of mutual authentication.

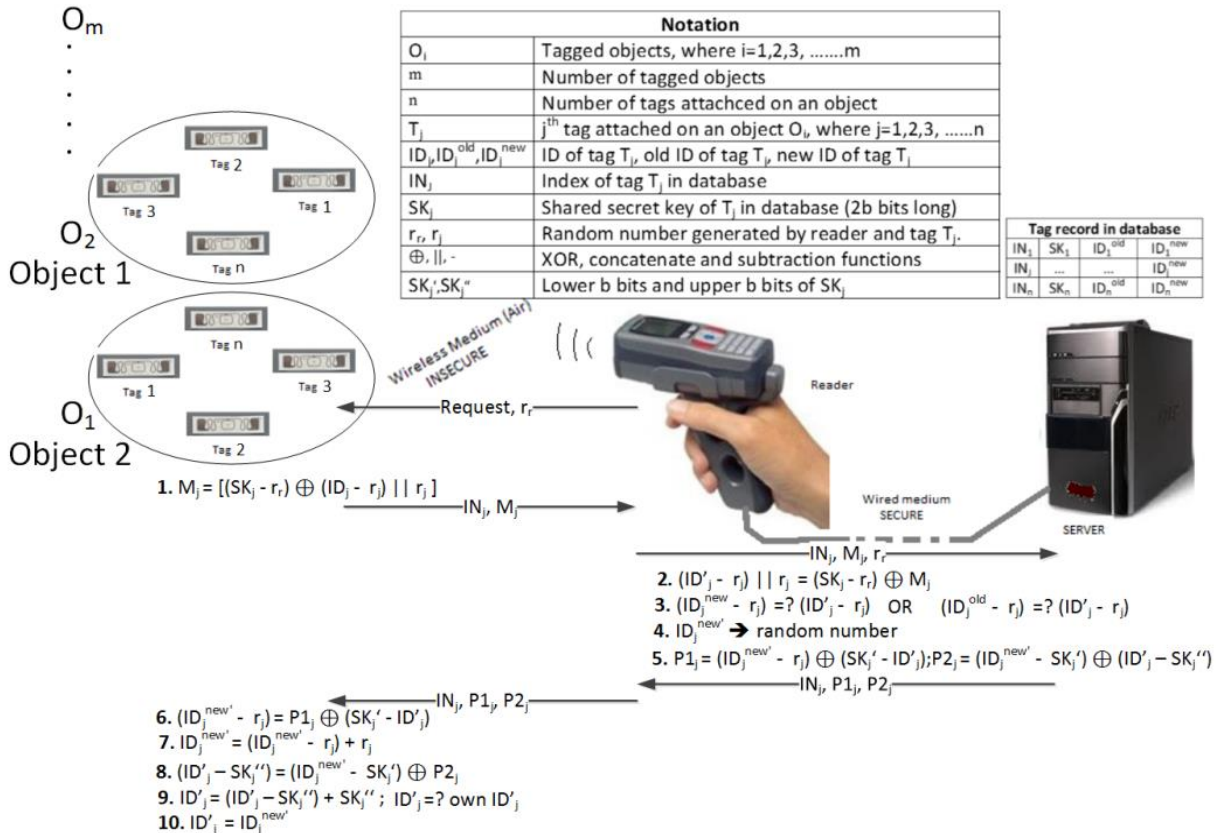


Fig. 1. DGMTAP Scheme (15)

III. ANALYZING DGMTAP

The presence of malicious wireless equipment users and dishonest readers is a common assumption, in radio frequency communications [14]. Adversaries are encouraged especially if a reply to every challenge is guaranteed. Due to the nature of RFID technology, every request is replied by a tag. Therefore, challenging from a distance and recording the replies of a tag is very popular among RFID hackers [16]. The replies are accumulated and analyzed, at a later time. In DGMTAP, although the presence of dishonest readers is assumed and no secrets are shared with the reader; the identity or the nonce (r_i) of the reader are not checked. The absence of the checks opens the way to a serious attack on DGMTAP. As a result of the attack, it becomes obvious that the claimed security properties of the protocol do not exist. Here is the attack scenario in detail:

An attacker challenges the tags of an object using the same bogus nonce $r_c = 0$ twice, and saves the replies. Observe that neither the tag nor the server checks for a zero r_c value. Denoting the first and second challenges with superscripts 1 and 2, respectively, from equation 1 of Figure 1:

$$M_j^1 = [(SK_j - r_c) \oplus ((ID_j - r_j^1) \parallel r_j^1)] \quad (1)$$

$$M_j^2 = [(SK_j - r_c) \oplus ((ID_j - r_j^2) \parallel r_j^2)] \quad (2)$$

XORing equations (1) and (2):

$$M_j^1 \oplus M_j^2 = (ID_j - r_j^1) \parallel r_j^1 \oplus (ID_j - r_j^2) \parallel r_j^2 \quad (3)$$

Because $(SK_j - r_c) \oplus (SK_j - r_c) = 0$ and $A \oplus 0 = A$. Equation (3) is an XOR operation which can be divided into XORing the lower and upper bits:

$$\text{Upper bits of } (M_j^1 \oplus M_j^2) = (ID_j - r_j^1) \oplus (ID_j - r_j^2) \quad (4)$$

$$\text{Lower bits of } (M_j^1 \oplus M_j^2) = r_j^1 \oplus r_j^2 \quad (5)$$

In mathematics, the XOR function is known as the modulo 2 addition without carry [17]. Therefore, the XOR operation can be approximated to addition. The trivial justification is left to the reader, while the XOR operations on the right hand side of equations (4) and (5) are approximated to addition:

$$UoM = (ID_j - r_j^1) + (ID_j - r_j^2) \quad (6)$$

$$LoM = r_j^1 + r_j^2 \quad (7)$$

Where LoM denotes the Lower bits of $(M_j^1 \oplus M_j^2)$ and UoM denotes the Upper bits of $(M_j^1 \oplus M_j^2)$. Adding equations (6) and (7):

$$LoM + UoM = 2 \times ID_j \quad (8)$$

The ID_j of the tag is obtained using equation (8), since M_j^1 and M_j^2 are passed in cleartext, during the message exchange. Now the attacker has the index IN_j and the ID_j of the tag. Next, the attacker uses the same dishonest reader to send the saved messages M_j^1 and M_j^2 to the server. Observe that, the server never checks the identity or the legitimacy of a reader. The attacker does not allow the replies of the server to reach the tag, but just plays M_j^1 and M_j^2 and saves the replies. The server believes that the tag used ID_j' , because it has not updated in the previous authentication session.

Therefore, the server uses the same ID_j' value in its database, for preparing its replies. As a result of the two sessions with the server, the following replies are received by the reader:

$$P1_j^1 = (ID_j^{new1} - r_j^1) \oplus (SK_j' - ID_j') \quad (9)$$

$$P2_j^1 = (ID_j^{new1} - SK_j') \oplus (ID_j' - SK_j'') \quad (10)$$

$$P1_j^2 = (ID_j^{new2} - r_j^2) \oplus ((SK_j' - ID_j') \oplus (ID_j' - SK_j'')) \quad (11)$$

$$P2_j^2 = (ID_j^{new2} - SK_j') \oplus (ID_j' - SK_j'') \quad (12)$$

XORing (9) and (11), then (10) and (11) yields:

$$P1_j^1 \oplus P1_j^2 = (ID_j^{new1} - r_j^1) \oplus (ID_j^{new2} - r_j^2) \quad (13)$$

$$P2_j^1 \oplus P2_j^2 = (ID_j^{new1} - SK_j') \oplus (ID_j^{new2} - SK_j') \quad (14)$$

Approximating the XOR operations in equations (13) and (14) to addition and subtracting (14) from (13) gives:

$$P1_j^1 + P1_j^2 - P2_j^1 - P2_j^2 = 2 \times SK_j' - (r_j^1 + r_j^2) \quad (15)$$

Using equation (7) and rearranging equation (15):

$$2 \times SK_j' = P2_j^1 + P2_j^2 - P1_j^1 - P1_j^2 - LoM \quad (16)$$

All of the terms on the right hand side of equation (16) are cleartext messages saved by the attacker. Therefore, now the lower b bits (notation table of Figure 1) of the shared secret SK_j are captured. The captured values $(ID_j$ and $SK_j')$ can now be used to break down the whole DGMTAP protocol. The attacker returns to equation (1) for a bitwise analysis and since $r_c = 0$, equation (1) reduces to:

$$M_j^1 = SK_j \oplus ((ID_j - r_j^1) \parallel r_j^1) \quad (17)$$

Separating the upper and lower b bits of the XOR operation, equation (17) can be broken into two equations:

$$UoM_j^1 = SK_j'' \oplus (ID_j - r_j^1) \quad (18)$$

$$LoM_j^1 = SK_j' \oplus r_j^1 \quad (19)$$

From equation (19), the value of r_j^1 is captured, because SK_j' was already exposed. Substituting the captured r_j^1 value in (18), the value of SK_j'' is also obtained. Now, the whole $2b$ bits of the shared secret SK_j are in the hands of the attacker. Inserting SK_j in equation 2, the second tag nonce r_j^2 is isolated. Now, by inserting the captured r_j^1 , SK_j' , ID_j' values in (9) and r_j^2 , SK_j' , ID_j' values in (11); both ID_j^{new1} and ID_j^{new2} are calculated. The tag's record in the database is now completely exposed. The capture of the full record of a tag is called a full-disclosure attack [9] and it has serious ramifications for the user of the tag.

IV. DISCUSSIONS

Authentication protocol proposals are as good as their claims. In other words, when the security of a proposed protocol is proven to be short of what it claims to be, it is immediately abandoned. As demonstrated, full record of DGMTAP tag can be exposed. An exposed RFID tag is not different than a barcode paper sticker on a commodity. The consequences of such a security breach are more critical than just revealing the secret identification of an object, as it will become apparent next.

A. Authentication Analysis

The authors of DGMTAP make four critical errors in their security analysis. First, since the reader - server channel is assumed to be secure, the backend server does not check the authenticity of the reader. The price paid is the giveaway of the two replies to the two bogus messages, in the full disclosure attack demonstrated, in the previous section. Secondly, the number of server replies with the old tag ID is not counted. Thus, blocking the replies of the server can go unnoticed. Hence, the server can be tricked to send multiple replies, using the same tag ID. The adversary simply accumulates the replies and exposes the repeated ID. Third error is the server's failure to check the nonce (r_r) of the reader. As observed in the attack above, a zero valued nonce facilitates the analysis of the DGMTAP messages. Finally, although multiple tags are used to identify an object, each tag's authentication does not add up to a more secure protocol, as in a grouping proof protocol [14]. As demonstrated in our full disclosure attack, the secrets of each tag can be exposed by carrying out the same analysis individually on each tag.

B. Security Analysis

Proposed protocols are normally expected to provide the basic security properties like message confidentiality, message integrity and privacy. Failing to do so, opens the way to the following known attacks.

1) *Eavesdropping*: Eavesdropping on messages going through air cannot be prevented and contrary to authors' claims, the secrets of DGMTAP tags are not secured enough to go through the air.

2) *Man-In-The-Middle Attack*: There is no need for this type of attack on DGMTAP, since the secrets can be obtained otherwise. But, after full acquisition of tag secrets, false messages can be formed and the server can be fooled by a man in the middle, using an unchecked dishonest reader.

3) *Replay Attack*: It has been demonstrated that replaying the same zero-valued reader's nonce, resulted in a full disclosure attack on DGMTAP.

4) *Location Tracing*: As the present and next identity values of a tag are exposed, by analyzing the exchange between a tag and a reader, an attacker can find out which object a tag belongs to. By recording the locations of the identified objects, tracing an object becomes easy.

5) *Forward Security*: This property cannot be provided by DGMTAP, because all coming identification values ID_j^{new} of the tag can be calculated, once the shared secret and the present identification ID_j are captured.

6) *Backward Security*: DGMTAP cannot provide this property, because by inserting the constant value of SK_j and the captured present identification ID_j in the saved message exchanges, all of the old ID_j values can be calculated.

7) *Synchronization Attack*: This attack is also possible, because a dishonest tag can be created with the captured secrets. The dishonest tag can communicate with the server because it can formulate M_j messages. The server is tricked to update ID_j twice. The authentic tag has no knowledge of the clandestine session between the server and the dishonest tag.

Hence, while the identity value in the authentic tag is unchanged, that value has been dropped out of the server's database. Consequently, the server will fail to recognize the authentic tag when it tries to authenticate with the server, because now it has no match in the database.

8) *Physical Attack*: This type of attack is in another category. Its prevention requires hardware sophistication such as secure memory and memory fuse architectures, which are beyond the scope of this work.

C. Some Recommendations for Correcting DGMTAP

DGMTAP can be improved easily by a number of precautions. First, the server should authenticate the reader and bind its use to a well-proven user. The user must have a secret login password and a unique feature of the reader; like the CPU ID, must be used. A detailed example can be found in work [18]. Such safety precautions eliminate the danger of malicious attacks via dishonest readers. Secondly, the server must check the reader nonce r_r , before evaluating any tag messages. "If $r_r == 0 \rightarrow$ abort" operation would suffice. Such a check eliminates the danger of simplifying the decryption of exchanged messages. Third, a further XOR operation after the concatenation operation in equation (1) can complicate the algebraic analysis of DGMTAP. Concatenation by itself is a weak operation, which can be easily reversed by breaking up a message at the point where it was concatenated. Therefore, concatenation should not be the last operation in an equation. Finally, a grouping proof protocol covering the tags attached on the same object can improve the security, as advised in work [14]. Grouping proof protocols usually challenge the first tag in the group (tag 1), next challenge tag 2 with the reply of tag 1, next challenge tag 3 with the reply of tag 2 and so on. At the end, the replies of the tags are packed and encrypted with the reader's user password. The server receives the resultant data package and verifies the reply of each tag. Any disagreement in the verification causes a fault in the authentication of the chain. Hence, the authentication of the object(s) is dependent on a more sophisticated protocol. DGMTAP has the multi tag basis for a grouping proof protocol, but does not use it.

V. CONCLUSION

A protocol attempting to bring security to RFID identification by introducing multiple tags per object has been analyzed. Full disclosure of the sensitive tag secrets was possible through an algebraic attack on the exchanged messages. The attack demonstrated that merely multiplying tags for identification can result in the breakdown of the claimed protocol's security features. Four recommendations have been made for improving the security of the analyzed protocol. But, it is best to start with the previous work, recommending lightweight cryptography for RFID tags [13].

Future work must try to comply with the new RFID standards aimed at popular UHF RFID tags [11]. Such intentions lead the research into introducing the Advanced Encryption Standard and Elliptic Curve Cryptography for secure channel initiation, in low cost RFID tags. Strong cryptographic tools are needed even in low cost tags, because

the captured messages are analyzed using computationally powerful computers.

ACKNOWLEDGMENT

This study was supported by TÜBİTAK (The Scientific and Technical Research Council of Turkey) (Project Number: 113S419)

REFERENCES

- [1] S. L. Ting, S. K. Kwok, A. H. Tsang and W. B. Lee, "Critical elements and lessons learnt from the implementation of an RFID-enabled healthcare management system in a medical organization," *J. Med. Syst.*, vol. 35(4), pp. 657-669, 2011.
- [2] M. L. Songini, "Wal-Mart details its RFID journey," *ComputerWorld* (April 22, 2007), <http://www.computerworld.com/article/2562768/enterprise-resource-planning/wal-mart-details-its-rfid-journey.html> (Accessed on 19 December 2014).
- [3] J. S. L. Ting, S. K. Kwok, W. B. Lee, A. H. C. Tsang and B. C. F. Cheung, "Design and development of an RFID-based behavioral awareness system for animal care management," *Annual Journal of IIE*, vol. 27, pp.47-56, 2007.
- [4] S. K. Kwok, A. H. C. Tsang, J. S. L. Ting, W. B. Lee and B. C. F. Cheung, "An intelligent RFID-based electronic anti-counterfeit system (InRECS) for the manufacturing industry," *Proceedings of Seventeenth International Federation of Automatic Control (IFAC) World Congress 2008*, pp. 5482-5487.
- [5] E. W. T. Ngai, T. C. E. Cheng, S. Au and K. H. Lai, "Mobile commerce integrated with RFID technology in a container depot," *Decis. Support Syst.*, vol. 43(1), pp. 62-76, 2007.
- [6] T. Tsuji, S. Kouno, J. Noguchi, M. Iguchi, N. Misu and M. Kawamura, "Asset management solution based on RFID," *NEC J. of Adv. Tech.*, vol. 1 (3), pp. 188-193, 2004.
- [7] W. Yao, C. H. Chu and Z. Li, "The adoption and implementation of RFID technologies in healthcare: a literature review," *J. Med. Syst.*, vol. 36(6), pp. 3507-3525, 2012.
- [8] M.H. Özcanhan, G. Dalkılıç and S. Utku, "Is NFC a better option instead of EPC gen-2 in safe medication of inpatients," *Radio Frequency Identification, Springer Berlin Heidelberg*, pp.19-33, 2013.
- [9] M.H. Özcanhan, G. Dalkılıç and S. Utku, "Analysis of two protocols using EPC Gen-2 tags for safe inpatient medication," *IEEE Innovations in Intelligent Systems and Applications (INISTA)*, 2013.
- [10] Information technology -- Radio frequency identification for item management -- Part 6: Parameters for air interface communications at 860 MHz to 960 MHz, http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=46149 (Accessed on 19 December 2014).
- [11] EPC Global Class1 Gen2 RFID Specifications, http://www.gs1.org/gsm/kc/epcglobal/uhf1g2/uhf1g2_1_2_0-standard-20080511.pdf (Accessed on 19 December 2014).
- [12] H.Y. Chien, "SASI: A new ultralightweight RFID authentication protocol providing strong authentication and strong integrity," *Dependable and Secure Computing*, pp. 337-340, 2007.
- [13] M.H. Özcanhan, "Improvement of a Weak RFID Authentication Protocol Making Drug Administration Insecure," *Life Science Journal*, vol. 11(10), pp. 269-276, 2014.
- [14] P. P. Lopez, A. Orfila, J. C. H. Castro and J. C. A. Lubbe, "Flaws on RFID grouping-proofs guidelines for future sound protocols," *J. of Network and Computer Appl.*, vol. 34(3), pp. 833-845, 2011.
- [15] S. Dhal and S. G. Indranil, "A new authentication protocol for RFID communication in multi-tag arrangement," *IEEE Computing for Sustainable Global Development (INDIACom)*, 2014.
- [16] Y. C. Yen, N. W. Lo and T. C. Wu, "Two RFID-Based Solutions for Secure Inpatient Medication Administration," *J. Med. Syst.*, vol. 36, pp. 2769-2778, 2012.
- [17] T. V. Deursen and S. Radomirovic, "Algebraic Attacks on RFID Protocols," *Information Security Theory and Practices (WISTP'09)*, LNCS, vol. 5746, pp. 38-51, 2009.
- [18] M.H. Özcanhan, G. Dalkılıç and S. Utku "Cryptographically Supported NFC Tags in Medication for Better Inpatient Safety," *J. Med. Syst.*, vol. 38(8), pp. 1-15, 2014.

Model Driven Testing of Web Applications Using Domain Specific Language

Viet-Cuong Nguyen

Department of Computer Science and Engineering
Faculty of Electrical Engineering
Czech Technical University in Prague
Prague, Czech Republic

Abstract—As more and more systems move to the cloud, the importance of web applications has increased recently. Web applications need more strict requirements in order to support higher availability. The techniques in quality assurance of these applications hence become essential, the role of testing for web application becomes more significant. Model-driven testing is a promising paradigm for the automation of software testing. In the web domain, the challenge however remains in the creation of models and the complexity of configuring, launching, and testing big number of valid configuration and testing cases. This paper proposes a solution towards this challenge with an approach using Domain Specific Language (DSL) for model driven testing of web application. Our techniques are based on building abstractions of web pages using domain specific language. We introduce WTML - a domain specific language for modeling web pages and provide automatic code generation with a web-testing framework. This methodology and techniques aim at helping software developers as well as testers to become more productive and reduce the time-to-market, while maintaining high standards of web application quality.

Keywords—Domain specific language (DSL); model-driven development; model-driven testing; WTML

I. INTRODUCTION

Advances in web-based technologies today has led to the rapid growth in the number of web applications used in business. As the demand for mobility and internet-of-things requires more complexity in web applications, the existing testing frameworks used to test software system struggle to get up to speed. Methods from model driven can support the rapid evolution of such system by building an abstract model of a web application and use the created models instead of specific code to generate tests. In general, the model of the web application does not need to include all the details of the implementation, but should be precise enough to guarantee that the test cases represent actual use scenarios of the web application [1].

In this paper, we present an attempt to build an approach using Domain Specific Language for model driven testing of web application.

Our techniques are based on building abstractions of web pages and modeling state-machine-based test behavior using domain specific language. This is used to form a more generic testing framework that can apply to many web-based systems to save time and cost.

This paper is structured as follows: In the next section, we review some knowledge of Model-driven Development (MDD), model-based testing and domain specific language as background information. The subsequent section discusses the current challenge to build a testing platform that can automate the process from development to execution. In the next section, we introduce WTML (our designed DSL) for test modeling and test development of web-based applications. In the last section, we present some conclusions on the methodology of using a domain specific language in model-driven testing of web applications.

II. BACKGROUND

Automated model driven testing has received much attention in recent years, both in academia and in industry. This interest has been stimulated by the success of model-driven development in general, by the improved understanding of testing and formal verification as complementary activities, and by the availability of efficient tool support [2]. Model driven engineering approach as a methodology could be described as follow:

A. Model driven engineering

Model-driven engineering (MDE) is a software development methodology, which focuses on creating and exploiting domain models. Models can be perceived as abstract representations of the knowledge and activities that govern a particular application domain. Models are developed through-out various phases of the development life cycle with extensive communication among product managers, designers, developers and users of the application domain. MDE aims to increase productivity by maximizing compatibility between systems, simplifying the process of design and promoting communication between individuals and teams working on the system [3].

The Object Management Group's (OMG) initiatives on MDE contain the Model-driven Architecture (MDA) specification. MDA allows definition of machine-readable applications and data models that enable long-term flexibility with regards to implementation, integration, maintenance, testing and simulation [4] [5]. There are two main modeling classes in MDA:

- Platform Independent Models (PIMs): these are models of the structure or functionality, which are independent of the specific technological platform used to

implement it.

- Platform Specific Models (PSMs): these are models of a software or business system, which are bound to a specific technological platform.

In the MDA, models are first-class artifacts, which are later integrated into the development process through the chain of transformations from PIMs through PSMs to coded application. The mapping and transformation between PIMs and PSMs are based on meta-model concepts. These concepts can be described by technologies such as Unified Modeling Language (UML), Meta Object Facility (MOF) or Common Warehouse Meta-model [3], [6], [12]. These languages are considered as general-purpose modeling languages. Currently, there are many challenges in implementing model driven testing due to the lack of standardization and tools. There are specific desired aspects for each application within its domain and this makes it difficult to design a tool that can be applied to all situations.

B. Model based testing

Model-based testing is application of model-based design for designing and optionally also executing artifacts to perform software testing or system testing. Models can be used to represent the desired behavior of a System Under Test (SUT), or to represent testing strategies and a test environment.

A model describing a SUT is usually an abstract, partial presentation of the SUT's desired behavior. Test cases derived from such a model are functional tests on the same level of abstraction as the model [10].

Currently, testing usually comprises between 30% and 70% of all software development projects. Hence, a good testing methodology and toolset will enable software developers and testers to become more productive and reduce the time-to-market, while maintaining high standards of software quality.

The purpose of the model-driven testing in the web domain is to provide a framework that helps developers perform the following tasks:

- Create models of web applications or pages: This enables developers to create the abstraction of the components. Developers can later use the model created as a skeleton for the test project. In this way, the test plan can be reviewed and simulated to discover problems in the implementation or model before the actual code is ready for test.
- Model behaviors: The behaviors and interactions of the web application are modeled using the modeling language to later support test case generation. These behavior models simulate the features of the web application.
- Generate test cases for the web components. The tools generate tests using data from the component (page) models and the behavior models. It is often a good practice to have the test cases that cover all required test specifications.

Test execution: The generated tests can be later executed either manually or automatically by some triggers. This test execution automatically compares the observed results with the results predicted by the model. Thus, developers can walk through a unit test case to examine each test interaction and identify where the test failed.

C. Domain specific language

In software development and domain engineering, a domain specific language is a programming or specification language dedicated to a particular problem domain, a particular problem representation technique, and/or a particular solution technique. The concept is not new. Special-purpose programming languages and all kinds of modeling or specification languages have always existed, but the term has become more popular due to the rise of domain specific modeling [7].

Adoption of domain specific language can be a solution to several problems encountered in various software development aspects. A DSL can reduce the costs related to maintaining software [8]. In comparison to other techniques, DSL is considered as one of the main solutions to software reuse [9]. On the other hand, using DSL also promotes program readability and makes its understanding easier. This enables users without experience in programming to create the models or programs as long as they possess knowledge of the targeted domain.

Another advantage of a DSL for modeling is the ability to generate more verification on the syntax and semantics than a general modeling language. This can reduce errors (and burden) on the debugging process.

III. CHALLENGE

The process of web application development starts with concepts, mock-ups and requirements. After that, following a lot of iterations, more and more mature prototypes are gradually created towards a working solution. Testing needs to be performed within every iteration in this process. This nature makes testing web applications a routine task from designing the tests to tests execution and report. When maintaining such systems, any change to the system also requires the execution of a complete regression test. Therefore, there is a need to build a testing platform that can automate this testing process from development to execution.

There exist many model-based testing approaches and tools that vary significantly in their specific designs, testing target, tool support, and evaluation strategies. In the web domain, there is a noticeable increase in the number of model-driven testing techniques in recent years. Firstly, the challenge in this area is to have a good design of a modeling language that used to represent the system. Secondly, there is the challenge for effectively defining the process of test case generation and evaluation. There are several aspects of a model-driven testing technique that need to be considered:

- Effective Modeling Language: The modeling language used to model system, can be a generic UML approach or a domain specific language, should bring up good solution on the web domain while being easy to read

and to understand. This language needs to be effective and designed with agility support to ensure that models can adapt to changes seamlessly.

- Automation: This is an important aspect in model driven development, it is the ability to generate final artifacts from high-level specifications. Automation also enables test case generation and execution mechanism to perform easily without manual refinements.
- Good Tool Support: The tool chain and platform support is essential for any approach. This allows the integration with other parts of the development process. This means that the platform should provide tools for editing, debugging, compiling and transformation. The tools should also be able to be integrated together with other languages and platforms without a lot of effort.

Although there exist many techniques that tend to vary significantly in their design, they usually don't provide adequate results in every applicable domain [11]. There are also challenges in other aspects of the modeling process. On one hand, the model has to be written in a notation powerful enough to describe any elements of the web page. At the same time, it has to be abstract enough to ease the process of model creation and promote software reuse.

IV. OUR APPROACH: DSL FOR WEB PAGE MODELING

Our approach is based on the principle of raising the level of abstraction by modeling web pages and describes their behaviors using the theory of State Machines. In order to check the conformance between the application and the model, the automated process for generating test cases from the model is used. Our approach uses DSL to develop the testing model together with the functional web page model development. We aim at introducing a DSL and the tool set that fit for this purpose.

In this approach, designing a new DSL with the support for modeling at a good abstraction level is crucial. This DSL can later be used for automatic generation of the model artifacts and code that implement the services. In theory, a general modeling language could also be used for this purpose but an appropriately designed DSL can perform the same job much more effectively.

There are three essential requirements to the DSL design that we aim to achieve during the creation of a DSL to ensure the quality of the language. Firstly, the language needs to be effective, while being easy to read and to understand. Secondly, as the modeling language can raise the level of abstraction away from programming code by directly using domain concepts, automation needs to be achieved to generate final artifacts from these high-level specifications. This automatic transformation at the same time has to fit the requirements of the specific domain. Finally, the DSL has to be able to provide support via tools and platforms. The DSL needs to be able to integrate with other parts of the development process. This means that the language is used for editing, debugging, compiling and transformation. It should

also be able to be integrated together with other languages and platforms without a lot of effort.

The starting point for a DSL for web page modeling is an abstraction of a web page. This abstraction model comprises the effective elements that are involved in the testing process and, optionally, the behavior of the transitions to be simulated and validated during the test execution. Following diagram depicts the simplified syntax rules of a page model:

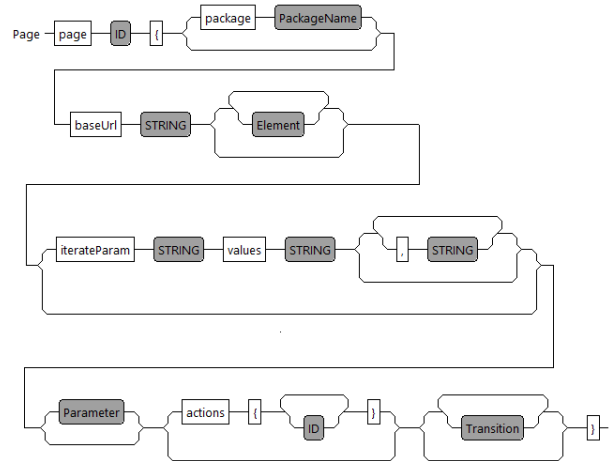


Fig. 1. Simplified syntax of a Page in WTML

The semantics of the language expressions starts with the page definition identified by its name (ID). In order to have package information for code generation, a package name can be optionally declared. Base URL is then assigned to each page. This gives us the possibility for customization of the parameters for the URL. Main information for a page is the elements. A page can have arbitrary number of elements. In order to query elements in a web page, we identified it with the XPath expression. The syntax for an element can be seen as in Fig. 2.

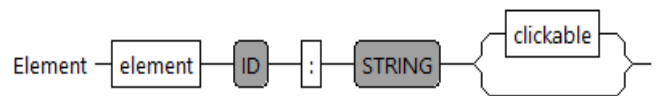


Fig. 2. Syntax of an Element model in a page

Each element starts with the keyword element followed by its name (ID). We then use a string literal to store its XPath expression. An element can optionally be clickable, this can be declared by the keyword *clickable*.

We then define the parameters of the page. A page can have any number of parameters. Each parameter starts with the keyword *param* as in Fig. 3.

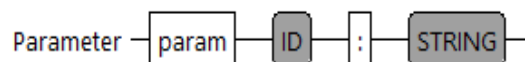


Fig. 3. Syntax of parameters in a page

Another type of parameter can be seen in the next block in a page as in Fig. 1 is the set of parameters to later be used in the code generation process to repeatedly test against. This is

defined by the keyword *iterateParam*. The parameters for iteration are comma-separated.

This information is enough if we just simply want to model a page for testing. The last components are the actions and transitions. These are the optional components to define the actions and transition between pages. This can be used later when we want to use state machine to model the test cycle of the whole web platform.

To demonstrate the simplicity of the model creation process in this approach, we can see how simple it is to write a textual model of a web component from a web application in a case study.

```
page RatingPage{
  baseUrl "http://www.webtest.org/login"
  element content "//*[@id='content']"
  element user "//*[@id='user']"
  element submit "//*[@id='submit']" clickable
  iterateParam itemID "12,13,14,15,16"
  param action "add"
}
```

This eight-line-of-code model at this abstraction level allows us to be very flexible on building the elements and logic needed for the test. At this level code reused is heavily promoted. This can be reused on many pages yet enables us to generate large amount of codes for test automation. Our benchmark pointed out that 90 lines of Java code were generated from this. This means we saved a significant amount of time that was otherwise supposed to be spent on test development. Overall, even if we take into account the time spent on developing and learning a new DSL such as WTML, this could still potentially provide a good productivity gain in test development.

V. AUTOMATION OF TEST GENERATION WITH WTML

According to IEEE standards, a test case is “a documentation specifying inputs, predicted results, and a set of execution conditions for a test item”. As the definition states, there are three required elements in a test case; test inputs, predicted results and conditions of execution. IEEE’s definition also clarifies the difference between test data and test case.

In order to generate test cases, test data must first be generated. Test data can be generated using different sources that describe the system, system’s behavior or how the system will be used, such as source code, system specifications, system requirements, business scenarios and use cases. Our approach utilizes specification-based method for test case generation.

In this approach, we focus on the verification of the web system against the design specification that was available on the test models. This comprises of abstract information on the available operations and its parameters. Information from the specifications allows generation of test cases for boundary-value analysis, equivalence class testing or random testing [13].

In WTML platform, in order to generate the tests, we first need to generate the model implementation of the page to be

tested against. A sample on how the page in Java was generated is as bellow:

```
@Page
@ComponentScan(basePackages=
    {"net.webmodeling.testing"})
public class FirstPage {
    private final static String
        baseUrl = "http://www.testpage.org/";
    private final static String
        iterateParamName = "value";
    @Autowired
    private AutoBrowser browser;

    @Value("#{ 'AAA,BBB,CCC'.split(',') }")
    private List<String> iterateParams;
    private static final By
        rating = By.xpath("//*[@id='viewcomments_click']");

    public String getRating() {
        return browser.getTextValue(rating);
    }
    public void clickOnRating() {
        browser.clickOn(rating);
    }
    ...
}
```

From the web page model syntax as seen on previous section, *iterateParam* is used when we want to iterate over a set of input parameters when testing a page. This becomes handy especially on the development of regression tests.

Another important aspect is *@Page* annotation, we introduced this annotation to inject special configuration to a page. This allows us to use Spring framework for processing pre- and post- Java bean creation. Testing data is injected directly into the page from the test models by using Spring *@Value* annotation. All setters and getters are also automatically generated from elements in the models.

This approach also provides a solution for automating regression testing. These tests are the reuse of the existing test cases from the previous system tests. Regression testing is performed when additions or modifications are made to an existing system. Since this could be run and generated automatically, regression testing could be performed anytime using WTML platform when there is a requirement.

VI. INTEGRATION WITH OTHER PLATFORM

One of the essential features of the modeling tool is the ability to integrate with other platforms. Selenium is a suite of tools to automate web browsers across many environments. WTML can utilize Selenium to provide automatic simulation with browser. WTML raises the level of abstraction by modeling the elements and actions on the web page. This model will then be used as input to generate code for modeling page accordingly. We use Java as the target language. Using Spring framework dependency injection we then can integrate layered architecture in the code generated. Configurations are injected into JUnit tests via Spring annotation.

To support WTML platform, we created our defined annotations in Java, this Page annotation consists of Configuration that can be later injected and directives to load

the application context. We also defined our browser implementation in order to integrate with web driver from Selenium and provide automatic processing. In general this browser is defined in the following way:

```
...
@Component
public class AutoBrowser {
    private static final int TIME_OUT_SEC = 10;
    private static final Logger LOGGER =
        LoggerFactory.getLogger(AutoBrowser.class);
    @Autowired
    private WebDriver webDriver;

    public void clickOn(By location) {
        webDriver.findElement(location).click();
    }

    public WebElement findElement(By location) {
        return webDriver.findElement(location);
    }
    public void goToPage(String url) {
        webDriver.get(url);
    }
    public void goToUrlWithParam(String baseUrl,
        Map<String,String> params) {
        final StringBuilder pageUrl =
            new StringBuilder();
        pageUrl.append(baseUrl + "?");
        for (Map.Entry<String, String>
            entry : params.entrySet()) {
            pageUrl.append(entry.getKey());
            pageUrl.append("=");
            pageUrl.append(entry.getValue());
            pageUrl.append("&");
        }
        goToPage(pageUrl.toString());
    }
    public void goToUrlWithSingleParam
        (String baseUrl, String paramName,
            String paramValue) {
        final StringBuilder pageUrl =
            new StringBuilder();
        pageUrl.append(baseUrl + "?");
        pageUrl.append(paramName);
        pageUrl.append("=");
        pageUrl.append(paramValue);
        goToPage(pageUrl.toString());
    }
}

@PreDestroy
private void destroy() {
    webDriver.quit();
}
public int getNumberOfElements(By location) {
    return webDriver.findElements(location)
        .size();
}
public String getTextValue(By location) {
    return webDriver.findElement(location)
        .getText();
}
public String getAttributeValue(By location,
    String attributeName) {
    return webDriver.findElement(location)
        .getAttribute(attributeName);
}
```

```
    }
    public String getCssValue(By location,
        String propertyName) {
        return webDriver.findElement(location)
            .getCssValue(propertyName);
    }
}
...
```

After the configuration of Selenium web driver is defined and loaded, we inject web driver into our *AutoBrowser*, this way we keep the Selenium code separated from our browser logic. This allows us to only focus on the requirements and logics of code generation and automation test runners. After that we define all necessary methods for our automated browser such as *getNumberOfElements* from a given XPath address inside any page.

With the integration of Selenium, we are able to perform automatic browser actions. This enables us to write automated tests for a web application directly in WTML, which allows for better integration in existing unit test frameworks.

VII. RELATED WORK

In the UML world, there has been effort on proposing techniques to automatically generate and execute test cases starting from a UML model of the Web Application by Filippo Ricca and Paolo Tonella [14]. This approach requires a manual work in several phases. There is manual work on the creation of models for testing and in the test refinement phase. Our approach has an advantage of fully automation in test case generation using the abstract web model and its action.

Alessandra Cavarra, Charles Crichton, Jim Davies, Alan Hartman and Laurent Mounier [15] presented the approach on test case generation utilizing UML. The authors' approach is based on extending UML using UML profiling capabilities. In these approaches, two profiles are created for different purposes. The first one is used to model the system under test by extending class diagrams, object diagrams, and state diagrams to support testing properties. The other profile is used to capture the test directives which are composed of the object diagrams and state diagrams. A transformation is then used to verify and produce scripts that can later be used to generate test cases.

A model-based testing approach is presented by Bouquet et al. in [16]. Their approach is based on a combination of class, object, and state diagrams, which can be found in UML and OCL expressions to automatically generate test cases from these models. Test cases are generated using a test generator that takes these diagrams and constraints as input. The authors discuss the need to alter the semantics of OCL to allow OCL expressions to have a side effect on the system state. In an overview of model driven testing techniques from the work of Mussa M., Ouchani S., Al Sammane W. and Hamou-Lhadj A. [11], the authors pointed out the shortcomings of this approach that it violates OCL semantics, which may hinder the acceptance of the approach by the UML community. One possible solution is to use an action language to express expressions that change the state of the system.

There has been also a direct attempt to use UML activity diagrams to generate test cases for Java programs in the work of Chen, Qiu and Li [17]. The approach is based on the

generation of test cases then compares the running traces with the activity diagram to reduce the test case set. The disadvantage of this approach is the limitation to the UML activity diagram that makes it impossible to obtain concurrency or loops for the tests.

Deutsch, Sui and Vianu in [18] introduced an approach that models data-driven web applications. This approach used Abstract State Machine to model the transitions between pages, determined by the input provided to the application. The structure and contents of web pages, as well as the actions to be taken, are determined dynamically by querying the underlying database as well as the state and inputs. The properties to be verified concern the sequences of events (inputs, states, and actions) resulting from the interaction, and are expressed in linear or branching-time temporal logic. This approach has an advantage of wide-range error detection. However, this leads to complex models that can make the integration with development methodologies not feasible.

Q. Yuan, J. Wu, C. Liu and L. Zhang [19] present an automatic approach to generate test cases of a given business process of a web service. The modeling of business process uses notations from Business Process Execution Language and UML activity diagrams. This approach is an example of applying MDA and the conformed transformation techniques. This approach aimed to build concise test models using given notations and generate test cases from these models. The advantage of this approach is the ability to apply in many types from unit testing to integration testing.

VIII. CONCLUSIONS

With the strict requirements of web-based systems, techniques to assure the quality of these systems play a very important role in the development process. Model-driven testing tools reduce overall testing time by supporting the reuse of many common testing functions. They also enhance test quality and complexity by offering a systematic approach to test suite generation. In this paper, we outlined the theoretical ideas and analysis from lessons learned during the real industry implementation of the framework. The approach introduced in our research provides a methodology for using a domain specific language in model-driven testing of web applications. Adopting WTML in combination with the MDA initiative allows early testing of model-driven systems and eases the sharing of models between the system developers and the system testers.

WTML was designed at the appropriate abstraction level to have better model readability and more support for integration. This is aimed at reducing test maintenance costs, since changes happen at the model level and are captured by the test models. When there are changes, we only have to regenerate the tests from the test models and all test cases are updated to the new specifications. This framework also enhances team communication because the model, test cases provide a clear, unambiguous, and unified view of both the system under test and the test. This technique decouples the testing logic from the actual test implementation. This makes the test architecture more robust and scalable. The shortcomings of this approach include a learning curve needed to adopt a new modeling language and the limitation of test

behaviors based only on the possible elements modeled in a page abstraction.

Domain specific language such as WTML can be applied to automation testing of web-based applications and pages. In practice, this approach has initially gained adoption in testing of web systems in the financial industry where the authors had the chance to work with. Our future work will continue on the improvement of the framework in terms of consistent methodology, wider code generation coverage and more efficient notations and syntax.

ACKNOWLEDGMENT

This work has been supported by the Department of Computer Science and Engineering, Faculty of Electrical Engineering and by the grant of Czech Technical University in Prague number SGS14/078/OHK3/1T/13.

REFERENCES

- [1] F. Bolis, A. Gargantini, M. Guarnieri, E. Magri, L. Musto, "Model-Driven Testing for Web Applications Using Abstract State Machines", in M. Grossniklaus, M. Wimmer, ed., *Current Trends in Web Engineering* vol. 7703, (Springer Berlin Heidelberg, 2012), pp. 71-78.
- [2] J. Peleska, "Industrial-Strength Model-Based Testing - State of the Art and Current Challenges", *Electronic Proceedings in Theoretical Computer Science* 111 (2013), pp. 3-28.
- [3] X. Qafmolla, V. Nguyen, Automation of Web Services Development Using Model-driven Techniques. In Institute of Electronics Engineers, The 2nd International Conference on Computer and Automation Engineering (ICCAE 2010), pp. 190-194, 2010.
- [4] Object Management Group (OMG): Meta Object Facility (MOF) Core. Retrieved March 20, 2012, <http://www.omg.org/spec/MOF/2.4.1/>, 2012.
- [5] Object Management Group (OMG): The Architecture of Choice for a Changing World. Retrieved April 20, 2013, <http://www.omg.org/mda>, 2013.
- [6] V. Nguyen, X. Qafmolla, Agile Development of Platform Independent Model in Model Driven Architecture. In Proceedings of the 2010 Third International Conference on Information and Computing, Vol. 2. IEEE Computer Society, Washington, DC, USA, pp. 344-347, 2010.
- [7] Wikipedia: Domain specific language. Retrieved January 15, 2013, from http://en.wikipedia.org/wiki/Domain_specific_language, 2013.
- [8] A. Deursen, P. Klint, Little languages: Little maintenance. *Journal of Software Maintenance*, pp. 75-93, 1998.
- [9] C. W. Krueger, "Software reuse", *ACM Computing Surveys (CSUR)* 24, 2 (1992), pp. 131-183.
- [10] Wikipedia, "Model Driven Testing", *Model Driven Testing* (2014).
- [11] M. Mussa, S. Ouchani, W. Al Sammane, A. Hamou-Lhadj, "A Survey of Model-Driven Testing Techniques", in *Quality Software, 2009. QSIC '09. 9th International Conference on* (, 2009), pp. 167-172.
- [12] X. Yu, Y. Zhang, T. Zhang, L. Wang, J. Hu, J. Zhao, X. Li, A model-driven development framework for enterprise Web services. *Information Systems Frontiers*, pp. 391-409, 2007.
- [13] M. Bozkurt, M. Harman, Y. Hassoun, "Testing Web Services: A Survey", Department of Computer Science, King's College London (2010).
- [14] F. Ricca, P. Tonella, "Analysis and Testing of Web Applications", in *Proceedings of the 23rd International Conference on Software Engineering* (Washington, DC, USA: IEEE Computer Society, 2001), pp. 25-34.
- [15] A. Cavarra, C. Crichton, J. Davies, A. Hartman, L. Mounier, "Using UML for automatic test generation", in *International symposium on testing and analysis ISSA* (Springer-Verlag, 2002).
- [16] F. Bouquet, C. Grandpierre, B. Legeard, F. Peureux, "A Test Generation Solution to Automate Software Testing", in *Proceedings of the 3rd International Workshop on Automation of Software Test* (New York, NY, USA: ACM, 2008), pp. 45-48.

- [17] C. Mingsong, Q. Xiaokang, L. Xuandong, "Automatic Test Case Generation for UML Activity Diagrams", in *Proceedings of the 2006 International Workshop on Automation of Software Test* (New York, NY, USA: ACM, 2006), pp. 2-8.
- [18] A. Deutsch, L. Sui, V. Vianu, "Specification and verification of data-driven Web applications", *Journal of Computer and System Sciences* 73, 3 (2007), pp. 442 - 474. Special Issue: Database Theory 2004.
- [19] Q. Yuan, J. Wu, C. Liu, L. Zhang, "A model driven approach toward business process test case generation", in *Proc. of the 10th International Symposium on Web Site Evolution (WSE)* (2008), pp. 41-44.

A Design of Pipelined Architecture for on-the-Fly Processing of Big Data Streams

Usamah Algemili

Department of Computer Science
The George Washington University
Washington, DC 20052, USA

Simon Berkovich

Department of Computer Science
The George Washington University
Washington, DC 20052, USA

Abstract—Conventional processing infrastructures have been challenged by huge demand of stream-based applications. The industry responded by introducing traditional stream processing engines along-with emerged technologies. The ongoing paradigm embraces parallel computing as the most-suitable proposition. Pipelining and Parallelism have been intensively studied in recent years, yet parallel programming on multiprocessor architectures stands as one of the biggest challenges to the software industry. Parallel computing relies on parallel programs that may encounter internal memory constrains. In addition, parallel computing needs special skillset of programming as well as software conversions. This paper presents reconfigurable pipelined architecture. The design is especially aimed at Big Data clustering, and it adopts Symmetric multiprocessing (SMP) along with crossbar switch and forced interrupt. The main goal of this promising architecture is to efficiently process big data streams on-the-fly, while it can process sequential programs on parallel-pipelined model. The system overpasses internal memory constrains of multicore architectures by applying forced interrupts and crossbar switching. It reduces complexity, data dependency, high-latency, and cost overhead of parallel computing.

Keywords—Big Data; Clustering; Computer Architecture; Parallel Processing; Pipeline Design; Variable Lengths; Symmetric Multiprocessing; Crossbar Switch; Forced Interrupt

I. INTRODUCTION

Conventional computing has been thoroughly challenged by the emerging situation of Big Data. Big Data is the problem of managing huge amount of unstructured data. The complexity of Big Data calls for new form of software clustering and hardware organization. At the beginning of this centenary, studies reported enormous growth of information that exceeded Moore's Law [1]. Big Data introduces unconventional pressure on time and memory performance. Consequently, new computation models are significantly required to cope up with Big Data situation. Researchers introduced "on-the-fly" clusterization of amorphous data. On-the-fly processing deals with a continuous stream of data, and it must maintain certain throughput of information flow. In this pattern, hardware design should not tolerate any postponement of oncoming stream. Multicore pipelined architecture provides a simple yet effective solution to the on-the-fly computation by transferring the operating states from core to core down the pipeline [2]. This pipelining device requires practically the same sequential programs that are currently used based on single processor system. Pipeline computing offers very

effective solution for big data streams. It increases the throughput considerably when processing intensive streams of data. Pipelined architectures consist of sequence of processing elements where the output of one processor is the input of the next one. "By pipelining, processing may proceed concurrently with input and output, and consequently overall execution time is minimized. Pipelining plus multiprocessing at each stage of a pipeline should lead to the best-possible performance" [3].

This paper investigates the previous work on multicore processing and parallel computing architectures. It discusses stream processing requirements, followed by general outlook over the current limitations of parallel systems. This paper suggests a hardware model that is especially intended to process Big Data clustering on-the-fly, while this model can process sequential programs using parallel-pipelined multicore design. Finally, it proposes the same model based on Symmetric multiprocessing (SMP) and forced interrupts.

II. MULTICORE PROCESSING AND PARALLEL COMPUTING ARCHITECTURES

A. Multi-Core processors

Most modern processors include huge number of transistors on one chip. The architectures of general purpose multicore processors allow multiple related tasks for execution, this would be conducted in different cores such as IBM Cell processor, Intel and AMD multicore processors. Usually, these cores are heterogeneous in time requirement because of advanced scheduling algorithms that intend to exploit these architectures effectively.

On the other hand, these architectures support shared access of global caches or memory, this support faces some limitations in accessing the same block by other cores which decreases their efficiency. Consequently, memory design has significant influence on high clock rates, and indexing references is important to attain high processing performance. Optimizing the instructions and developing the data queues may increase the performance. However, these solutions have obvious limitations in heavy processing queues like graphics manipulation [4].

For instance, SIMD (Single Instruction Multiple Data) technique was one of the earliest programming methods to stress parallelism in microarchitecture design. More instructions were added by Intel in 2004. The introduction of 90 nm process-based Pentium games processor was followed

by (Streaming SIMD Extensions) SSE3 and SSE4. This was to improve thread synchronization and math capabilities.

Nevertheless, computing field is extensively progressive, so conventional processors such as multicore CPUs (Central Processing Units) are replaced by power-aware multi-core CPUs coupled with GPUs (Graphics Processing units). The idea behind this new trend is to take the advantage of chip die area. This integration can increase the efficiency of SIMD, and it may provide superior environment that supports stream processing and vectorization. This approach uses large memory units and large register sets that are distributed among different levels of system hierarchy [5].

B. Parallel Programming Architectures

HMPP (Hybrid Multicore Parallel Programming Environment) based on GPUs (Graphics Processing Units) can provide tremendous computing power. With current NVIDIA and AMD hardware group of graphic products, a peak performance can reach hundreds of gigaflops. GPUs designed originally for graphic cards, and it have emerged as the most powerful chip in high-performance workstations. Unlike CPUs with multicore architecture that uses two or four cores on chip, GPUs consist of manycore architecture that can run thousands of threads by hundreds of cores in parallel.

CUDA and OpenCL is a coevolved hardware/software architectures that enable HPC (High-Performance Computing). Developers were encouraged to utilize GPUs' tremendous power in computation and memory bandwidth, this by familiar programming environment -C programming language-. Apparently, the advantages of GPUs over CPUs are undoubtedly interesting. However, an application must retain certain characteristics in order to insure performance benefits: First, well-designed parallelism for massive amount of data. Second, Intensive Kernels that represent very large fraction of execution time should be computed (Amdahl's law) [6]. Third, an application that requires arithmetic intensity and density. These types of applications usually favor the use of the multi-computing units. Forth, Local memory access that is simple and regular, and it should avoid pointer tracking code. Last but not least, local memory accesses that can exploit the pipeline structure of GPU boards [7].

GPU programing mainly uses data-parallel paradigm, which frequently called stream computing. Stream computing relies on a map operation that consists of using the same computation on all elements of a stream such as arrays. Basically, a stream is one or multidimensional array with homogeneous points. Usually, parallel paradigm is based on operations such as Mapping were the map applies kernel function to all elements of stream, yet Kernel function can access elements from many input streams. Also it uses Reduction were an array of elements processes single value.

Typically, there are two types of memory-access operations that can be executed. First, gather operation which assumes kernel is able to read any element of a stream. It is usually of the form of $(x = s1[s2[i]])$. Second, Scatter operation that assumes kernel is able to write any element of a stream. The form of memory operation is similar to $(s1[s2[i]] =)$.

Previous hardware architecture of GPUs was not able to efficiently implement these operations. Luckily modern architectures have overcome this constraint, however, it should note that if $(s2[i])$ is not permutation of $(s1)$, then the result of this operation is non-deterministic. For example, if $(s2[i] = s2[j] = x)$, we should consider that if $(i \neq j)$ then $(s1[x])$ will be assigned more than once in undefined order.

C. Systolic Arrays

Another form of parallelism is systolic arrays, in which the data flows between processors in synchronization. Systolic arrays are specialized form of parallelism where different data may flow in different direction (down or right). Processors compute and store data independently. H. T. Kung and Charles Leiserson were the first to introduce systolic arrays in 1978 showing multiple processors in arrays connected by short buses [8]. Typically, systolic arrays use joint form of parallel computing and pipelined flow of data.

Systolic arrays miss two key features that we aim to achieve in the context of processing big data clusterization. First, Systolic arrays rely on parallel programing, and we target an architecture that would utilize sequential programs by forced interrupts. Second, each processor stores data independently of each other, which adds unfavorable dependency in case we process variable-lengths of data. Finally, the multidirectional nature of systolic arrays adds global synchronization limits due to signal delays. Running time that allows parallel overhead on several processors may exceed the ideal program running time.

D. Graphic Processing Units (GPUs)

Recently, the use of modern GPU (Graphic Processing Unit) increased considerably. The GPUs have been evolving since the very first computer system placing number of key challenges that are facing programmers to fulfill future application demands and support various platforms.

These challenges such as effective use of GPU's architecture and performance increase have led to outstanding results, yet this computational advancement stimulated software engineers to formulate innovative programming techniques in order to utilize GPUs' capability.

The increasing interest of parallel software requirements heightened the need for deep analyzing and scientific comparison of software methods whether in programming or architecture. The development of GPUs' technology has led to the hope that GPUs will contribute in many applied sciences and will open a new window for continuous growth. Recently, GPUs applications are being adopted by sciences which need processing massive data sets such as physical simulations, image processing, computer vision, data mining and text processing as well as smart phones and portable devices. Consequently, multicore processors and parallel programing has become favorable topic among HPC (High-performance computing) teams [9]. Many researches have been extensively studying general purpose GPUs coupled with multicore processors; they are looking for paralleling the tasks and keeping the sequential execution to its minimum. However, the optimal use of parallelism depends on GPUs architecture.

III. PROBLEM STATEMENT AND CONTRIBUTION

A. The Limitations of GPUs

In the context of big data streams, GPUs may not provide enough main memory for large chunks of data. GPUs mitigate this problematic situation by accessing multiple GPU boards to single node. PCIe (Peripheral Component Interconnect Express) can interface with many GPU boards providing an aggregated storage. For instance, eight GPU boards that contains 6 MB of internal memory can allow up-to 48 MB of shared memory. However, this solution does not work efficiently on algorithms that requires random access to large data due to local physics and multi-access constrains. Moving the data among GPUs by high-latency PCIe bus can create huge computational intensity. The moment when algorithms get larger than internal memory of GPU, the performance of PCIe's net system decreases dramatically. In fact, transferring anything over PCIe can lower the speed twentyfold compared to onboard main memory [10].

B. Parallel Programming Constrains

Real-time processing has led to widespread use of multicore architectures. Experts took the advantage of multiprocessors by embracing parallel programming in order to exploit parallel cores. This approach provided promising opportunities in HPC (High Performance Computing). However, it created big challenge for software industry. Most existing programming languages are designed to perform sequential execution. Parallel Programming added extra skill-set requirements whereas programmers already deal with many hardware and software complications. As a result, multicore processing unit such as GPU (Graphical Processing Units) has evolved to accommodate these challenges providing well-distributed and properly managed parallel cores, yet software engineers have to place additional effort and time in developing GPU applications that adhere to the promising parallel-core architecture [11].

C. Contribution

Clearly, this paradigm of parallel computing relies on parallel programming in order to utilize parallel hardware arrangement. It has memory limitation where variable lengths of Big Data streams may require processing data on-the-fly. Nevertheless, the hardware design may change this fact, and it can overcome these constrains. This paper presents a novel multicore pipelined architecture by forced interrupts. It is parallel model of computing that executes Big Data clusters on-the-fly, while it still can utilize sequential programs. Our promising architecture designed to handle variable lengths of data, and it can achieve very low-latency in time. This discourse proposes reconfigurable FPGA hardware architecture, where pipeline length can comply with Big Data clustering work-load. It provides scalable framework that can add more processing units, and it can change HW configuration according to our processing needs. The proposed hardware organization is especially aimed at Big Data clustering. It consists of three main components. First, pipelined multiprocessing elements that operates on multicore tasks in parallel. Second, sophisticated Main Memory management by crossbar switching. Third, forced-interrupt that allows conventional software programs to utilize our proposed

platform, and it may eliminate the overhead cost of parallel programming effort.

IV. DISCUSSION

A. Multicore Pipelined Architecture for Executing Big Data Streams

The conventional realization of parallel programming introduced three main approaches that have been used in multicore processing. First, Task Parallel that partitions input software into functions and tasks. This approach schedules each function or task onto multicore processor. Second, Data Parallel that partitions input data, then it schedules data segments onto multicore processor system. Third, Pipeline Parallel that uses elements of sequential processing capability. It decomposes a program into states then run each state simultaneously.

While the first approach requires sophisticated software development, the second method is useful for data-independent applications. These applications may require complicated modules for data-scheduling. Pipeline processing is common solution for high-intensive volumes of data; however, it is limited by the maximum length of processing stage and/or the ability for task-decomposition [2].

Multiprocessor Pipelined Architecture proposes hardware design that utilizes the benefits of these three methods in order to process Big Data streams. It uses parallel computation coupled with data-parallel method and pipeline-parallel approach. The main goal of this architecture is to progressively process incoming data flows. We determine the length of the pipeline by the size of data chunk that we receive [11]. Multiprocessor Pipelined Architecture divides the program into equal processing times by forced interrupts as it is described in [Ber90][Ber94][Ber00]. This technology has US PATENT No. 6145071. Given specific time of processing for each processor, this pattern ensures the receiving of incoming data without interruption, while it still using the conventional implementation of sequential software [12][13].

B. The Multiprocessor Pipelined Architecture by Forced Interrupt

Figure 1 shows multiprocessor pipeline architecture that uses forced interrupts in order to automatically slice each program into fixed durations. Each processing cycle starts with (L) loading, (P) processing and (U) unloading. Since each cycle has fixed duration, this design performs efficiently on large volume of data with relatively small algorithms. The initial design is limited by algorithm size that requires additional round of processing.

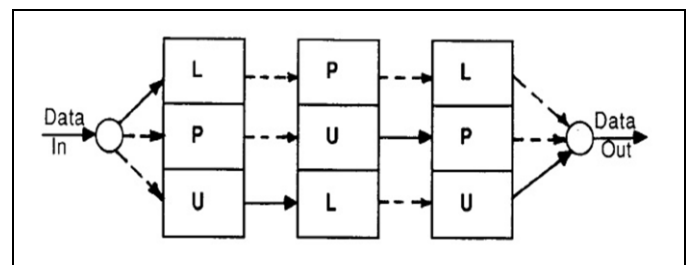


Fig. 1. General diagram of data flow in the pipeline [12]

Figure 2 zooms in the internal structure of each stage. Every microprocessor interacts with three memory blocks M1-M3. MPU sends addresses and control signals, while Sa1-Sa3 direct the data flow from previous stage or to the next cycle [12].

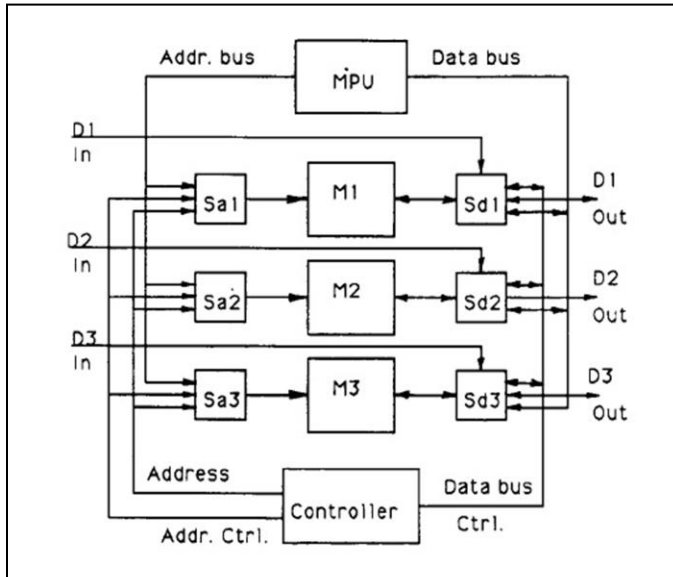


Fig. 2. Internal structure of one stage [12]

The challenges of algorithm size as well as the overhead of memory data transformation are well addressed in recent Multicore Pipeline Organization. The multicore system uses

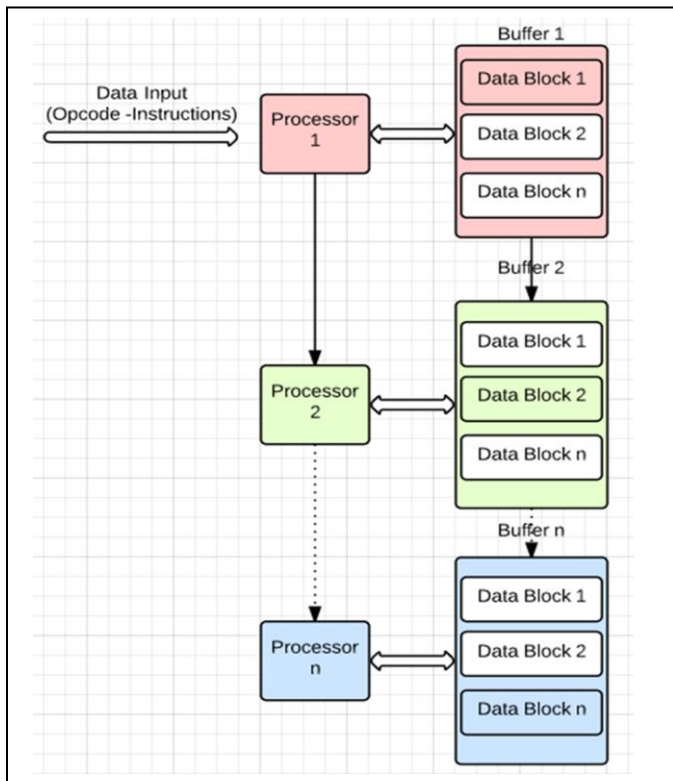


Fig. 3. Ideal situation where the blocks are in equal size

program slicing and forced interrupts. Theoretically, it receives data, then it generates blocks of different sizes based on slice function. Figure 3 shows the ideal situation where the blocks are in equal size, and each processor would execute one block respectively. In this case, each processor executes the block with the same color timing/slicing is not a big issue. In contrast, figure 4 illustrates variable lengths processing. This situation presents data blocks that are not equal in size or processing time. Hence, we need special handling by forced interrupts and program slicing. Each processor may process certain amount of data, then it can be stopped.

The number of processors required for one block execution may vary according to data length and processing time. The multiprocessor pipeline allows an arbitrary algorithm to be performed on-the-fly on a data chunk, given a sufficient number of processors. The major condition for continual mode of stream operations – equal durations of time intervals for computations at each section of the pipeline – is realized by forced interrupts at each processing stage. Time of processing is of no significance to this design due to forced interrupt. Figure 4 shows how different processors can share processing different blocks by forced interrupts.

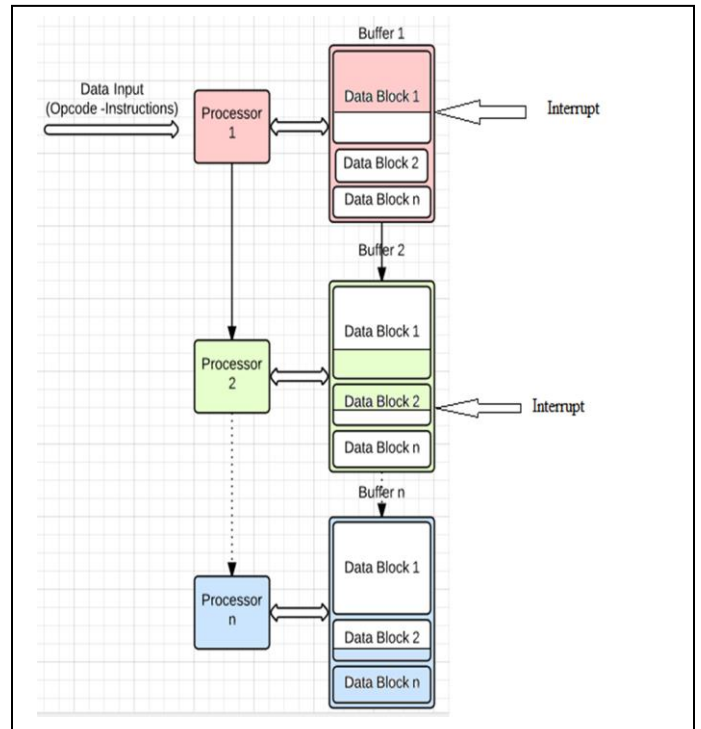


Fig. 4. Data blocks are not equal in size or processing time

In the context of processing Big Data clusters, memory management plays important role, and novel technology of memory and cache management is proposed. This multicore pipelined architecture provides very simple, yet, effective solution by switching the state of processors among data blocks, and it eliminates the overhead of internal-data transformation. As a result, the performance of this pipeline architecture increases significantly.

C. Multicore Memory and Cache Architecture Based on Crossbar Switching

This technology does not relocate memory data down the pipeline; instead, it uses crossbar switch in order to assign memory data blocks to the corresponding processor. Memory data may include program status information that allows the next processor to resume the work starting from last state. This approach has been applied on multi-memory/multi-cache design. Figure 5 illustrates the use of Crossbar Switch internally.

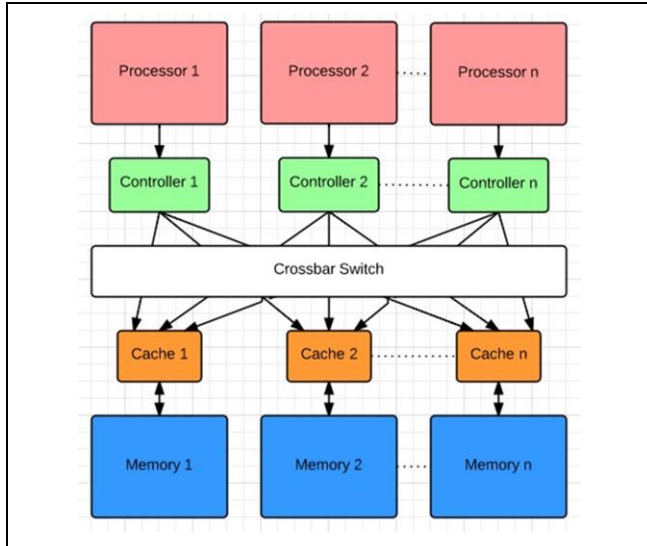


Fig. 5. The use of crossbar switch and shared cache management internally

This design does not assign corresponding memory for each processor; it reduces the cost of data relocation. Cache and memory can be grouped into one set, and Crossbar Switch would assign each processor to one group. The number of groups should be equal to processing elements. This

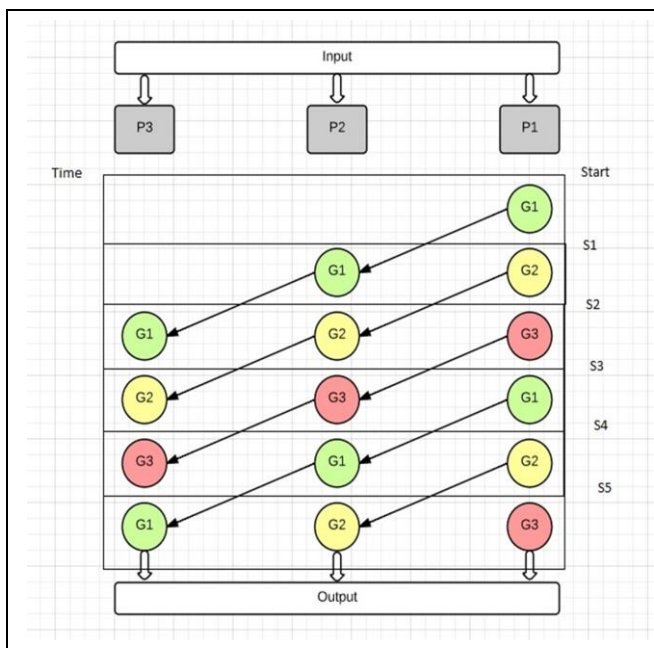


Fig. 6. Processing steps of the pipeline

P: Processor
G: Memory Cache Group
S: Switch

organization eliminates the cost overhead behind memory relocation as well as additional round setup as it described by forced interrupts technique. This organization replaces loading/unloading operations by switching mechanism. Theoretically, we can add as much processing elements as we need to reasonably process any given input of Big Data clusters.

Figure 6 shows the processing steps of Multicore Memory and Cache Architecture Based on Crossbar Switching. Once the data loaded in G1, the corresponding core P1 starts processing G1 data for fixed time, then crossbar switch assign the remaining data to P2. At this stage, P1 starts processing the new incoming assigned by crossbar switch [11].

D. New Multicore Architecture Based on Symmetric multiprocessing (SMP)

This technology follows Symmetric multiprocessing (SMP) architecture whereas many processors can share single main memory. This approach solves the issue of algorithm size, one main memory can receive input data, and then it can assign each memory chunk to processor that would process the instruction set in fixed time. After-which it shifts to the next block of data respectively; that allow the next processor proceed operations where the previous processor stopped.

Computer cluster systems have proved the efficiency of this technique on intensive amounts of data. However, this organization works at processing level. In stream processing, data storage can add high-cost operation within processing path, hence, the system must minimize unnecessary storage operations to archive low latency. Figure 7 shows an architecture that decreases time-intensive operations by processing messages on-the-fly.

In real time situations, we try to avoid dependencies, the program must process messages in given time by timeouts during-which this architecture can proceed with partial data.

The system promotes multi-threading by allowing data partition among processing blocks without having the developer writes low-level code. That would prevent blocking external data thereby minimizing latency.

The objective of this system is to be able to efficiently process external data that can arrive in either variable-lengths, high-volumes or both. In order to achieve better performance, the system should optimize execution path, and it must minimize boundary-crossing overhead. The desired length of the pipeline must be tested with performance in mind to insure sufficient processing path while decreasing the additional cost of multi-processing passage.

Figure 7 illustrates the desired communication between each processor and main memory. This architecture provides more flexibility to add more processing unites into the pipeline sharing the same data source.

Raman and Clarkson [14] carried out interesting project that proof the efficiency of this specific type of architectures.

Their project recognizes parallelism with non-identical processing unites. These unites can work simultaneously with one shared memory.

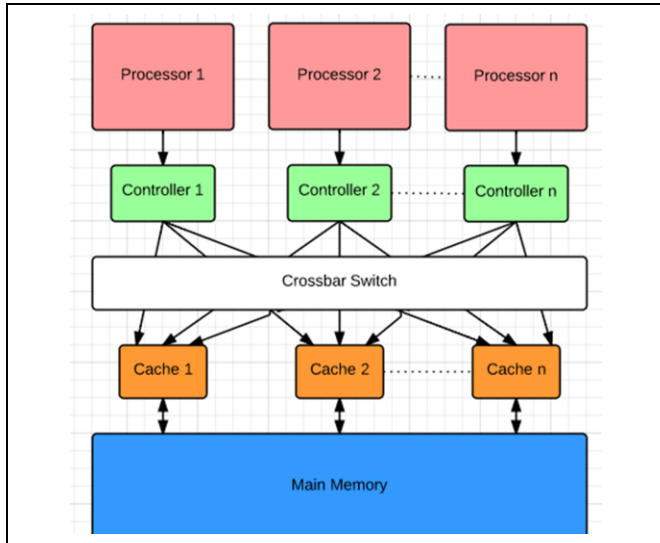


Fig. 7. The desired communication between each processor and main memory

Conventionally, the design of multiprocessor pipeline moves data chunks when processor loading-state changes. In contrast, this new architecture offers two advantages over conventional models. First, the data blocks does not relocate when switch operation occurs. Second, it allows other processors to load data as long it is in ideal state in order to utilize the pipeline.

E. Time Table and Analysis

Table 1 demonstrates one-cycle processing throughout pipeline of three core processors. It explains how variable-lengths of data blocks can be executed in different processors

TABLE I. ONE-CYCLE PROCESSING THROUGHOUT PIPELINE OF THREE CORE PROCESSORS. L: LOADING. P: PROCESSING. B: DATA BLOCK. S: SWITCHING

Time	P1	P2	P3
2	L-B1		
3	P-B1		
4	P-B1		
5	S-B1		
6	L-B2	P-B1	
7	L-B2	P-B1	
8	P-B2	P-B1	
9	P-B2	P-B1	
10	S-B2	S-B1	
11	L-B3	P-B2	P-B1
12	L-B3	P-B2	P-B1
13	P-B3	P-B2	P-B1
14	P-B3	P-B2	P-B1
15	S-B3	S-B2	S-B1

using switching operations and forced interrupts. Switching Operations (S) and forced interrupts happen at the same time. Table 1 illustrates each data block (B) by one color throughout the pipeline. This design assure availability of processor 1 by a given input and hardware specification, and it hinders any complications behind scheduling and load balancing.

Based on multi-cycle processing, we calculate the time required to perform system operation (T). We assume that a pipeline of (n) cores would execute an operation in (C) cycles. We discretize system operations in one cycle by (D).

A typical processor may execute in time of

$$Tc = (C) \times (D) \tag{1}$$

Given a number of cores (n) to execute input-data on average, the number of cycles performed on each given core presented as

$$X = C \div n \tag{2}$$

Therefore, the time given for typical processor can be also described as

$$Tc = (x) \times (n) \times (D) \tag{3}$$

The length of the required pipeline (L) can be found by

$$L = x + (n - 1) \tag{4}$$

factored by number of cycles that are required for equal duration process (m).

$$Lp = x + (n - 1) \times (m) \tag{5}$$

Let us assume that the internal operations of the pipeline take $O = Lp/m$ that is equal to

$$x + (n - 1) \times (m) \div m \tag{6}$$

$$O = \lceil x \div m \rceil + (n - 1) \tag{7}$$

In multicore system, switching operations can also increase system latency, and each hardware may have different switching capability. We present this overhead for each switch operation by (S). The total switch overhead expressed by

$$Ts = (O - 1) \times (S) \tag{8}$$

Hence, the total number of cores required to process a given data block including switching overhead expressed by

$$L = Lp + Ts \quad (9)$$

and the time required to perform all operations described as

$$T = (L) \times (Ts) \quad (10)$$

This discretization allows us to estimate performance speedup of multicore parallel pipelined architecture compared to typical single core system where Improvement = (Tc)/T.

$$\text{Improvement} = \frac{(X \times n)}{([X + (n - 1) \times m] + [(|x/m| + n - 2)] \times S)} \quad (11)$$

F. Reconfigurable FPGA Design

1) Overview

The design consists of three processing units all processors share the same AXI stream bus, each processor samples the incoming data when it receives an interrupt signal from the control unit, the output of these processors is send to a multiplexer which selects which output stream AXI bus to be used, the select for the multiplexer is also received from the control unit.

A MicroBlaze processor was chosen to collect the data from the processors; it receives no control signal from the control unit and it treats the data coming from the AXI stream bus as a data from single source.

The control unit is responsible for controlling the multiplexers at the processors input and output also it send interrupt signals for the processors to start sampling data.

2) Processors

Each processor contains the following ports:

- a) AXI stream input
- b) AXI stream output.
- c) Interrupt input.
- d) Busy output.
- e) Offload input.

The design consists of a finite state machine with 2 states and a 256 32-bit Ram. It stays on the first state waiting for the interrupt from the control unit, upon receiving an interrupt, the finite state machine goes to second state where it samples and stores the incoming data in the Ram. After storing 256 word, it goes to final state it enables the processor AXI stream output and reads from the Ram until it reaches final address then it goes back to first state.

3) AXIS DEMUX

The demultiplexer routes the system input stream AXI bus to any of the three processors according to the select signal it receives from the control unit. It has four main ports:

- a) One AXIS stream input.
- b) Three AXIS stream outputs.

This module can be removed from the design of a single AXIS data source. It is used to feed the three processors (AXIS counter for example).

4) AXIS DEMUX

The AXIS multiplexer is responsible for selecting which processor output to be fed into the MicroBlaze, it receives its select input from the control unit, all parameters are adjustable However, the control sends the select signal for clock cycles which is the time required by the processor to offload the data in its Ram.

5) Control unit

The control unit is responsible for synchronizing the complete design, and it shares the same data valid signal with the processors from system AXIS input bus in order to track the number of data words in the design. It consists of a finite state machine with 3 states. In the first state, it gives control signals to enable the AXIS Input and output of processor 1, and it counts the incoming data words until it reaches 256 (all parameters are configurable). After that, it go to the second state where it selects the AXI input and output of the second processor and after 256 data word it go to final state where it selects the third processor and then it return back to first state to repeat the complete process.

a) Interrupt:

The control unit can send the interrupt signal for the processor according to processing requirements, in the initial design the control unit sends the interrupt signal when the data reaches 256 word. This can be adjusted to interrupt according to any data word length or to data rate if a timer is used.

G. Design performance

This is a system designed using AXIS stream bus. Therefore all data transfer are constrained by the AXIS protocol timing performance as a result a word can be send each clock cycle in case of holding the valid signal high and sending the data each clock cycle.

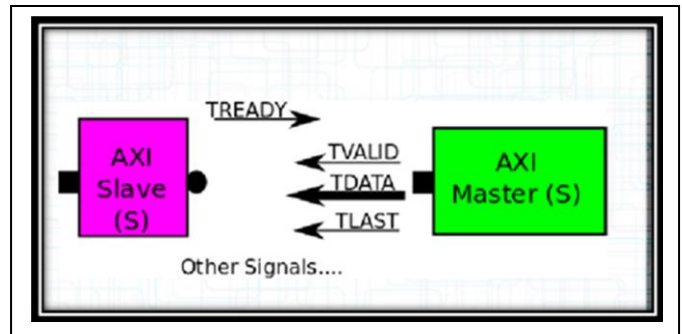


Fig. 8. AXIS protocol stream bus

In this design we assumed that TVALID is set high and a new data word is available each clock therefore the design processes a data word each clock cycle.

The complete design is synthesized at 100 Mhz which is the available clock source in the FPGA Zedboard [15]. For future work, we recommend to upgrade the project to include AXI timer and embedded design to access the timer; this would

enable reading the number of clock cycles for any future processing operations which may affect the processing time.

V. CONCLUSION AND FUTURE WORK

Parallel programming on multiprocessing systems is a challenging software domain. We proposed program slicing by a novel method of dynamic resource management that allows organizing on-the-fly processing of arbitrary complexity without parallel programming. The new architecture is very suitable for handling Big Data systems. The experimental results and performance comparison with existing multicore architectures demonstrate the effectiveness, flexibility, and diversity of the new architecture, in particular, for large data parallel processing.

The considered pipelining processing is of especial significance for applications of the presented technique of Golay Code clustering [16] as it involves very diverse and rather sophisticated computations for realization of multiple data cross-sections with sophisticated "Meta Knowledge" templates. Performance analysis introduces promising opportunities in real-time processing from pre-processing steps of clustering algorithms until final visualization.

REFERENCES

- [1] Brown, John Seely, and Paul Duguid. *The social life of information*. Harvard Business Press, 2002.
- [2] Liao, Duoduo, and Simon Y. Berkovich. "A new multi-core pipelined architecture for executing sequential programs for parallel geospatial computing." In *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*, p. 23. ACM, 2010.
- [3] Dewdney, A.K. *The (New) Turing Omnibus*. Henry Holt and Company, New York, 1993.
- [4] Raman, S. & Clarkson, T. (1990) Parallel image processing system – a modular architecture using dedicated image processing modules and a graphics processor. *IEEE, Conference on Computer and Communication Systems*, September 1990, Hong Kong, pp. 319–323.
- [5] Che, S, Boyer, M and others. (2009), *Rodinia: A Benchmark Suite for Heterogeneous Computing*, Department of Computer Science, University of Virginia.
- [6] Amdahl, Gene M. "Validity of the single processor approach to achieving large scale computing capabilities." *Proceedings of the April 18-20, 1967, spring joint computer conference*. ACM, 1967.
- [7] Bodin, F., & Bihan, S. (2009). Heterogeneous multicore parallel programming for graphics processing units. *Scientific Programming*, 17(4), 325-336. doi:10.3233/SPR-2009-0292 From : <http://ehis.ebscohost.com.proxygw.wrlc.org/eds/pdfviewer/pdfviewer?sid=0dc925aa-da7b-4e0e-852d-2df3c6def810%40sessionmgr112&vid=3&hid=2>
- [8] Quinton, Patrice, and Yves Robert. *Systolic algorithms & architectures*. Prentice Hall, 1991.
- [9] Brown, J.D.; , "High Performance Processor Development for Consumer Electronics Game Processor Perspective," *VLSI Circuits, 2007 IEEE Symposium on* , vol., no., pp.112-115, 14-16 June 2007 doi: 10.1109/VLSIC.2007.4342680 URL: <http://ieeexplore.ieee.org.proxygw.wrlc.org/stamp/stamp.jsp?tp=&number=4342680&isnumber=4342661>
- [10] Novakovic, Nebojsa A. "CPUs Will Fight Back as GPU Computing Hits the Limits." - *The Inquirer*. *The Inquirer*, 1 Aug. 2012. Web. 02 Dec. 2014. <<http://www.theinquirer.net/inquirer/feature/2195344/cpus-will-fight-back-as-gpu-computing-hits-the-limits/page/2>>.
- [11] Liao, Duoduo. *Real-time solid voxelization using multi-core pipelining*. Diss. The George Washington University, 2009.
- [12] [Ber00] S. Berkovich, Z. Kitov, A. Meltzer: On-the-fly processing of continuous data streams with a pipeline of microprocessors. In *Proceedings of the International conference on Databases, Parallel Architectures, and Their Applications (PARBASE-90)*, IEEE Computer Society, Miami Beach, Florida, March 1990, pp. 85-97.
- [13] [Ber00] S. Berkovich, E. Berkovich, and M. Loew, 2000. "Multi-Layer Multi-Processor Information Conveyor with Periodic Transferring of Processor's States for On-The-Fly Transformation of Continuous Information Flows and Operating Method Therefor", US PATENT No. 6145071, owned by George Washington University. Date issued - November 7, 2000.
- [14] Raman, S. & Clarkson, T. (1990) Parallel image processing system – a modular architecture using dedicated image processing modules and a graphics processor. *IEEE, Conference on Computer and Communication Systems*, September 1990, Hong Kong, pp. 319–323.
- [15] Zynq, Xilinx. "7000," "Zynq-7000 all programmable soc overview, advance product specification-ds190(v1. 2) available on: http://www.xilinx.com/support/documentation/data_sheets/-ds190-Zynq-7000-Overview.pdf," August (2012).
- [16] F. Alsaby and S. Berkovich. Realization of clustering with Golay code transformations. *Global Science and Technology Forum, J. on Computing (JoC) Vol 4 No 1*, 2014.
- [17] Alhudaif, Adi, Tong Yan, and Simon Berkovich. "On the organization of cluster voting with massive distributed streams." *Computing for Geospatial Research and Application (COM. Geo)*, 2014 Fifth International Conference on. IEEE, 2014.
- [18] Spafford, Kyle L., Jeremy S. Meredith, Seyong Lee, Dong Li, Philip C. Roth, and Jeffrey S. Vetter. "The tradeoffs of fused memory hierarchies in heterogeneous computing architectures." In *Proceedings of the 9th conference on Computing Frontiers*, pp. 103-112. ACM, 2012.
- [19] Vivado, H. L. S. "Vivado high level synthesis." (2012).

Review of Cross-Platforms for Mobile Learning Application Development

Nabil Litayem^{1,2}, Bhawna Dhupia¹, Sadia Rubab¹

¹Computer Science and Information, Salman Bin Abdulaziz University
Wadi College of Arts and Science, Kingdom of Saudi Arabia

²LSA Laboratory, INSAT-EPT, University of Carthage, TUNISIA

Abstract—Mobile learning management systems are very important for training purpose. But considering the present scenario, the learners are equipped with a number of mobile devices that run by different operating systems with diverse device features. Therefore, developing a mobile application for different platforms is a cumbersome task if appropriate tools and techniques are not practiced. There are three categories of mobile application namely native, web based and hybrid. For mobile learning system, hybrid is an appropriate choice. In order to avoid re-implementation of the same hybrid application for different platform separately, several tools are proposed for example: PhoneGap, Adobe Air, Sencha Touch and QT, each with their own strength. With proper use of the strength of individual framework or the combination of frameworks, more compatible and more stable, cross-platform mobile learning application specifically for quizzes and assignments can be developed.

Keywords—cross-platform; mobile devices; framework; mobile learning; hybrid mobile app

I. INTRODUCTION

In the present era of proliferating computer networks, and electronic devices (mobiles, tablets, PCs) every individual and organization is trying to get access to information and use these devices for their advancement and improved performance. Electronic information is accessible to a huge population in the world. Mobiles have become so much common that we have started to use it in every sphere of life. Whether it is entertainment or education, we can see a very intense use of mobiles and we can consider them as a new personal computer. It doesn't mean that desktop computers are now useless, but the mobile devices market is growing fast. They are cheap, convenient because of their portability, and due to geo location often more useful than PC.

Desktop applications are now commonly communication based and application developers develop a single application in cross-platform that can easily run on different desktop platforms e.g. Mac or Windows. There is an increasing demand of mobile applications to use mobile devices, but an application development for mobile devices is not so simple and it's a big challenge. These mobiles have different operating system and unlike PC OS, mobile OS decides the type of programming language required for applications running on it. So we need to design the mobile application according to its platform. There are various categories of app development for mobiles broadly categorized into three, namely, native apps, mobile web apps and hybrid apps and

according to the app categories and platforms, we have different choices in programming languages and framework for mobile application development. For running an application on multiple platforms, a cross-platform framework is required. Cross-platform frameworks for PC app are not adaptable to mobile app, so vendors provide cross-platform framework for mobile apps (some of them are discussed in section V). With this advancement of technology maybe third party cross-platform emerges that supports both PC app and mobile app irrespective of platform.

In our studies, we explore cross-platform strategies that would be helpful for providing solution to barriers in an app developed for mobile learning systems in a heterogeneous device environment. The mobile learning is defined as “Any sort of learning that happens when the learner is not at a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies” according to [1]. In a university environment incorporating mobile learning for quizzes and assignments requires learning system tools [1] specifically course search, file exchange, report generation, content creation, administration, off-line work, calendar and indeed many others.

With the diversity in devices (like Android, iPhone, Blackberry, Nokia Symbian, Laptops, Windows Mobile Phone), in order to provide above mentioned tools of learning system to its end user, an application suitable for the heterogeneous platform environment is required. We have discussed the different platforms in section II, categories of mobile apps in section III, app development languages and frameworks suitable for heterogeneous platform mobile apps are discussed in section IV and section V respectively. In section VI we discuss cross-platform strategies suitable for online or off-line quizzes and web based assignments and in section VII we conclude the studies.

II. MOBILE PLATFORMS

There are different operating systems that are used in the mobile market: Symbian, Blackberry, iOS, Android, Windows phone and Plam's webOS. Application development in a few of them discussed in detail in [14].

A. iOS

To develop for iPhone, iPod, Intel based Mac computer running OS X or later is required. iPhone SDK provided by Apple includes Xcode IDE, iPhone simulator and a suite of

additional tools for app development. App Store provides user facility to search and download app developed by iOS SDK.

B. Android

It is released by apache license and built on linux. Android app can be built using windows, Mac, or linux and java is primary language of Android. Java classes recompiled in Dalvik byte code and run on Dalvik virtual machine. Android does not support J2ME, and its most commonly used editor is Eclipse. Developers may create native libraries in C or C++. The Google play store is the official site and portal for Android app.

C. Blackberry

Blackberry smartphone platform support web development with HTML, CSS and javascript, java application development using MIDP, CLDC and RIM proprietary APIs. 32 bit windows OS support development of Blackberry app in Java.

D. Windows Mobile

It provides more desktop like user experience. In addition to C++ and C#, Silverlight and XNA are used for application and game development. For windows mobile development visual studio. Net framework, windows mobile SDK and developer toolkit are required.

Symbian

Symbian OS is used by Nokia, Sony Ericsson, Siemens and Samsung [1]. Language supported by Symbian OS is python makes use of Symbian C++ APIs. Python can also run on Mac, Windows and Linux [13].

III. CATEGORIES OF MOBILE APPLICATIONS

Irrespective of the type of mobile platform there are three main categories of mobile apps as in which normally we designed the applications. Depending upon the application requirement, we decide the category suitable to our requirements. First of all, let us discuss about all of them briefly.

A. Native apps

These applications are specific to the mobile operating system environment, take full advantage of its particular features and can be developed with the cross platform approach with a single code base for all devices [3]. Native apps get a performance gain by using the generated native code [2]. As one project maintained for each OS this leads to an increment in the development team, costs, and time and challenge for developers are the new ones that continually appear in market [3]. Native apps works without the connectivity of the internet. If you are working in an environment where there is no connectivity, native app is the way to go. Since native apps work with the device's built-in features, they are easier to work with and also perform faster on the device. Application build in this category are platform dependent. So whatever language you will use, you will have the full access to IDEs which provide better tools to develop

and debug the project faster. Examples of native apps are Angry Birds, Shazam¹.

With all these benefits of native apps, they have a few disadvantages also. Maintenance of the Native app is complicated task both for the users and developers. Developers have to program it according to different platforms and users have to update it regularly. The development cost of this app is more if you are making application for different platforms. Sometimes it becomes very difficult for the developer to give maintain and offer support as users of different mobile may be using different versions of the apps.

B. Web Apps

Web apps work with the devices that have browser therefore they can work on desktop computers as well on mobiles. In responsive web application, design decided by the server and applied at client level renders according to device features [3]. Whereas mobile web app that provides better usability as compare to responsive web, its content, provided merely for mobile devices consequently there is a need to maintain different sites for each device [3]. Users need not to go to mobile app store to update or download the application. Whenever the users will log in, they will get the updated version. Developers also need not to bother about the mobile platform. There will be the single universal version which can be used by any mobile platform. Hence, the maintenance cost of the web app is low. Example of Safari and Chrome web apps given in mobiloud webiste.

On the other hand, web app has some shortcomings also. Internet connection is a must in web apps. Web apps are not compatible with smartphone features like camera, GPs, phone dialing, etc. Web apps are not even listed in play store. Users have to search it on the web to use. Performance of web apps is slower as compared to native apps. They are also more difficult to build a regular user-base, unless they save it as a bookmark. Users won't have the app's icon on their devices as it is a web link which can be open as required. As a developer or publisher you can't send them notifications to bring them back to your content.

C. Hybrid apps

Hybrid apps combine technologies from native and mobile Web apps to gain the benefits of each. They behave like a native app because they are installed from a web store and have access to device specific features as in native app but developed using web app tools [3]. The tools for hybrid app can modify pre-packaged HTML pages, can change user interface according to device platform and allow both offline and online usage [2]. Hybrid mobile apps can be released on multiple platforms when using certain web technologies like HTML5, CSS3 and JavaScript. It will save the overhead time and cost used to prepare softwares for each platform. The Netflix app is one example of a hybrid app which runs the same code base on all platforms. Facebook, TuneIn Radio, LinkedIn are some of the examples of hybrid apps.

There are a few disadvantages of using hybrid app development. Hybrid apps are not executed natively; the

¹ <http://www.mobiloud.com/blog/2012/06/native-web-or-hybrid-apps/>

HTML5 and JavaScript portion of the app is rendered and executed by the platform's Web engine, which adds another layer between the user and the app. This can make the execution of the application slower. This is a new technique as compared to other two, so there are less tools available for the development of hybrid apps. Sometime, performance issues for certain types of apps are there in hybrid apps like on complex native functionality or heavy transitions, such as 3D games.

IV. APPLICATION DEVELOPMENT LANGUAGES

Choosing the right development tool is very crucial and important decision. A new technology's successful adoption often depends on its development tools. Good tools help new developers more easily get started and make experienced developers more productive. For instance, the success of Microsoft's programming environment is closely associated with the success of its Visual Studio tools. This topic summarizes the major app development languages available for different platform that we discussed in section II with their relevance to cross platform mobile apps.

A. Objective-C

Objective-C with Xcode IDE, primary programming language offered by Apple only for iOS [4] provides object oriented capabilities, requires dynamic runtime and runs merely on MAC operating system². It is used to build native apps that run directly on iOS [1]. Xcode suits include interface builder and instrument. In the MVC design pattern for iPhone application development Objective-C class required for view controller and web view can also be included in native iPhone application [14].

B. Java

Java³ is a very popular programming language when it comes to mobile application programming. Many developers are using this language as it easy to use and many online tutorials are available to get the help while development process. One of the popular development tools of Java is Android SDK. It includes many standard Java libraries like data structure libraries, math libraries, graphics libraries, networking libraries as well as special Android libraries that will help programmers to develop best Android applications. Moreover, Java can be used to program native apps, mobile web apps and hybrid apps. It is an object oriented programming environment that is used for platform independent application. The application running on battery driven devices built by J2ME java family, it provides networking support and has API JSR082 for Bluetooth technology [13]. Java with Eclipse IDE provided app running on Android and support web view [14]. For blackberry app both J2ME and Eclipse IDE are suitable. Java CLDC Emulation required for windows phone and J2ME for Symbian [17].

C. JavaScript

Considered as a major app development language in [14] along with HTML and CSS is important for Mobile Web Application and almost all framework support JavaScript. JavaScript supported by all platforms, Android and iPhone have full featured browser while the browser on Blackberry support JavaScript 1.0, 1.1, 1.2, and 1.3. Therefore HTML and CSS code based on platform specific browser features of one device cannot be used on another device with different browser features.

D. HTML 5

It is considered by many developers and companies as a standard for web development and even hybrid can be used with HTML 5 framework [3]. Due to HTML 5 difference in native and hybrid has reduced significantly, moreover an app in HTML and JavaScript deployed on local system can provide a similar structure as native apps [5]. It provides information regarding error handling and also support offline storage, network connectivity, multimedia, sockets and threads, drawing animation and advanced form controls[12]. A very prominent feature of HTML 5 is client side storage and up to 5 MB data can be stored by the user of the web application [12]. iPhone and Android support web application developed using HTML 5.

E. C#

It is used for Microsoft mobile and a web or desktop application implemented in Java can be translated to C# for windows mobile platform [5]. It targets the .Net CLI and take advantage of .Net framework. In C# developer can take advantage of web libraries, database connectivity, and socket programming [13].

F. PHP

In mobile web apps server side applications can be built using PHP, Node.js and ASP.Net [6]. PHP IDE provides a framework to create robust mobile PHP applications⁴. With PHP [15] we can access the user agent string for detecting device, attain information about browser, check device capabilities and image rendering with PHP WURFL API. Open source PHP is required for Android.

G. C\C++

For Microsoft windows phone C++ use within XAML app and in games. Windows phone runtime native API are built in C++ and can be projected in C# or VB.Net [1]. Android native libraries written in C\C++ [16]. Symbian C++ [13] provides full access to device features and improved speed, that is an edge for symbian C++ over python and java and eclipse based carbide C++ is preferred IDE for Symbian and Nokia. Open C [13] provides cross-platform development, a set of middleware libraries for the smartphone platform and TCP/IP socket programming.

H. Python

²developer.apple.com/library/mac/documentation/Cocoa/Conceptual/ProgrammingWithObjectiveC/

³ www.javaworld.com/article/2074670/mobile-java/

⁴ http://www.zend.com/en/products/studio

Python with a rich and standard library of modules is defined in [13]. It provides a scripting solution using Symbian C++API. Used for Symbian particularly Nokia's platform. It can run on Windows, MAC OS, Linux and Symbian OS. It supports a rich set of smartphone features such as sending and receiving SMS / MMS, camera, bluetooth, network access, sound recording and playing, text to speech and 2D / 3D graphics. It supports different dialogs e.g. pop-up notes, query, pop-up menu, select list. By combining with the web services through JSON, REST new type of applications and services can be created. Bluetooth protocol use for Bluetooth connectivity between devices and gsm_ module for location awareness.

The native development languages of different smartphone operating system are mentioned in [14]. Table 1 summarizes mobile platform support for the set of app development languages with IDE according to what it is stated in section II and IV.

To conclude this discussion we have come up with that there are various languages used for cross-platform application development. According to report of developer-economics-q3-2014 research only 15% of mobile developers are targeting browser while 42 % are using HTML, CSS, and Java Script. Whereas Java, C / C++, Objective C and C# are also very popular among developers. According to statistics mentioned in that research we provide the percentage of primary languages share in mobile app development in fig 1.

TABLE I. MOBILE PLATFORM SUPPORT FOR DEVELOPMENT LANGUAGES

Mobile Platform	Development Languages							
	Java	Java Script	PHP	HTML 5	Phyton	C\C++	C#	Obj-C
iOS	Duke Scripts ⁵	Full browser support	Server side app	Full Webkit based browser	-	Xcode IDE	-	Native Xcode IDE
Android	Native Eclipse IDE	Full browser support	Open source PHP	Full Webkit based browser	-	Native APIs	-	-
Blackberry	Sun JDK J2ME platform Eclipse IDE	Partial support	Server side app	Browser dependent	-	Eclipse IDE	-	-
Windows Phone	Java CLDC EMU	Depend on browser and webkit	Server side app	Depend on hardware UI IE Browser	-	Native APIs	Native .Net framework	-
Symbian	J2ME	Depend on APIs	Server side app	Full Webkit based browser	Scripting with Symbian C++ API	Eclipse based Carbide C++, Open C	-	-

⁵ www.javaworld.com/category/java-ios-developer/

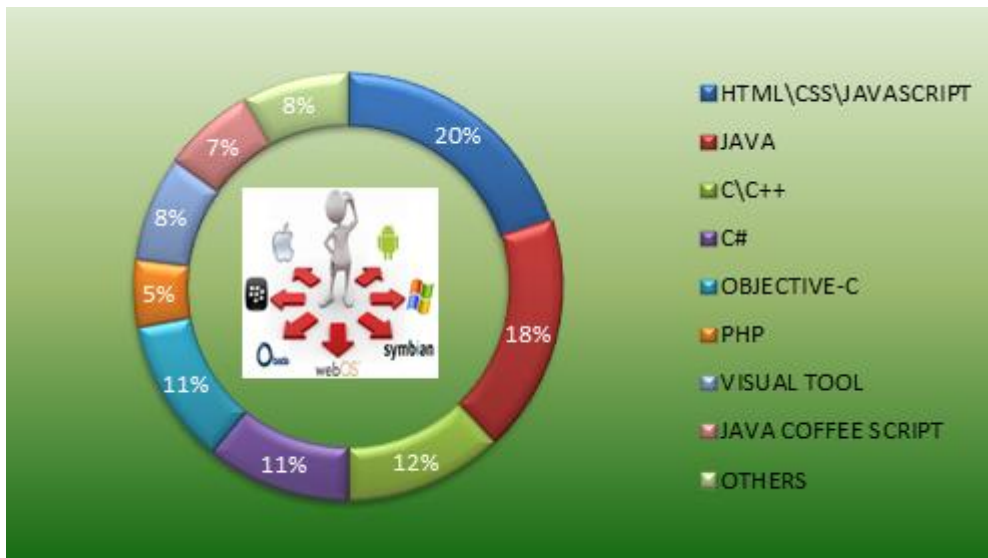


Fig. 1. Primary Languages share in Mobile App Development

V. FRAMEWORKS FOR CROSS PLATFORM MOBILE APPLICATIONS

In the previous section we have discussed different programming languages used for development of mobile applications. Percentage of usage of these languages depends upon the mobile app category and device platform supported by the target language. Although an application can be built

without a framework, but with the framework we get a cross platform app with an easy development and deployment process [3]. The given table 2 summarizes some of the frameworks used for mobile apps. Information of XMLVM, PhoneGap, DragonRad and Rhodes based on [4], Appcelator Titanium, QT, Adobe AIR, mobile voice agent, Sencha Touch, aCME, Processign on [2][7][8][9][10][11][16] respectively and the rest on [14].

TABLE II. FRAMEWORK OF MOBILE APP DEVELOPMENT ENVIRONMENT

Platform	Category of Mobile App	Type Environment of	Programming Language	Cross Platform Deployment
PhoneGap	Hybrid	Web view framework	HTML, CSS, JavaScript	iPhone, Android, Windows Phone, Blackberry, Symbian
JQTouch	Web, Native	Desktop, Browser environment	JQuery Plug-in, JavaScript, CSS	Desktop PC, Windows Mobile
Appcelator Titanium	Native	Rich APIs and low level TCP sockets	JavaScript	Android, iPhone, Blackberry
iWebKit	Web based	Desktop, Browser environment	HTML5, CSS3	Desktop PC, iOS mobiles
Adobe Air	Hybrid	Rich Internet application platform	Action Script, HTML, CSS, JavaScript, MXML	iOS(iPhone, iPad, iPod Touch), Android, Blackberry
Sencha Touch	Hybrid, Web based	Access to subset of phone native API	HTML, CSS, JavaScript	iOS, Android, Blackberry, Windows Phone
iUI	Web based	Browser Environment for iOS	JavaScript, CSS	iOS based devices
QT	Hybrid	Deployable UI and applications	C++, QML	Symbian, Maemo
Mobile Voice agent	Hybrid	Client server app with proprietary speech engine	HTML, JavaScript, speex audio format	iPhone, Android
xUI	Web based	Browser Environment	JavaScript	iPhone, Blackberry
Rhodes	Native Web	MVC support Real business logic	Ruby, HTML, JavaScript, CSS	iPhone, Android, Blackberry, Windows Phone, Symbian

		environment		
Dashcode	Web based	Desktop, Browser environment	HTML, CSS, JavaScript	iOS based devices
XMLVM	Web based	Cross compilation with API mapping	Java Byte Code	Android, iOS
CiUI	Web based	Browser Environment	CSS	iPhone
DragonRAD	Web, Native	Database driven drag and drop environment	Visual Drag & Drop Tiles	Android, Blackberry, Windows Mobile
aCME	Web based	Database driven Rich Internet App	Java	Desktop PC, Smartphone, tablets
QuickConnect Family	Native Web	Database access	HTML, CSS, JavaScript	iPhone, Android, Blackberry, Web OS
Safire	Web based	Browser Environment	HTML, JavaScript, CSS	iOS based device
iPhones-universal	Web based	Browser Environment	HTML, CSS	iPhone
Bedrock	Native	J2ME to native C++	Java, C++	Android, iPhone, Windows Mobile
Corono	Native	Scripting environment	Lua scripting lang	Android, iPhone, iPad
MoSync SDK	Native	IDE with Single Code base	C \ C++, HTML \ JavaScript	Symbian, Windows Mobile, Android
Adobe Flash Lite	Native	Video and dynamic web content	ActionScript	Symbian, Android, Blackberry, Windows Mobile
WebApp.Net	Web based	Browser Environment	JavaScript	iOS, Android, WebOS
Unity	Native	Game development	JavaScript, C#, Python	Android, iPad, PS3
Processing	Native	Electronics Arts, Visual Design	Java, JavaScript, HTML	Android, iOS
Jo	Web based, Native	Browser Environment	JavaScript, HTML5, CSS3	iOS, Android, Windows 8, BlackBerry 10
RhoSync	Native	Middle tier Data store	Ruby	Connection with Rhodes based smart app

With this outline about different frameworks mentioned in table 2, we can make a comparison of their usage across operating systems (mobile platforms) and can provide a layout of the development process according to app category. Since with cross-platform mobile frameworks it is inevitable that developers select software that provide a common development approach across different platforms. In this topic we discuss briefly on some of the popular cross-platform frameworks selected from table 2, focusing on app development process.

A. PhoneGap

PhoneGap is a free and open source framework that allows you to create mobile apps using standardized web APIs for the platforms. To develop for iPhone, we need a Mac OS X computer. PhoneGapLib is a static library that enables users to include PhoneGap in their iPhone application projects. We can also create new PhoneGap-based iPhone application projects through an Xcode project template. Xcode is Apple's development environment for Mac OS X and iPhone that includes the iPhone SDK capabilities of the framework. In case of Android, a developer needs to install the Android SDK and Eclipse plus the Android Development Tools

development plug-in for the Eclipse. ADT extends the capabilities of Eclipse to let you build Android projects and APKs in order to distribute applications. Development for Blackberry device needs Eclipse 3.4 or 3.4.1. Along with this developer has to install BlackBerry JDE Plug-in for Eclipse, and the Eclipse Software Update for the BlackBerry JDE v4.6.1Component Pack.

B. Sencha Touch

Sencha touch is the best environment for cross-platform development based on HTML5 and CSS3. It enables developers to build powerful applications that work on iOS, Android, BlackBerry, Windows Phone, and more. The Sencha Touch API is pure JavaScript. Developers need to be fairly experienced at JavaScript to take advantage of the Sencha Touch framework. Sencha Touch apps can not only be accessed via browsers, but can also be deployed as hybrid apps using native wrappers. Sencha Touch is not dependent on jQuery, so is compatible with both the iPhone and Android. It uses XML and HTML to create interface design and procedural code for creating a UI object. The latest version of it supports Apache Cordova APIs for camera, capture, connection, events, geolocation, media, notification, splash

screen and storage. These are few features available to native apps that are essential to app developers.

C. *WebKit*

The *WebKit* is a framework focused on being fast, lightweight, and specifically for developing web applications and websites for Apple's devices. It can easily be integrated into iPhone application developed in Objective-C and applications developed using Rhodes and PhoneGap frameworks. Developers familiar with HTML and CSS framework can easily use *WebKit*. The *WebKit* framework includes a comprehensive set of style sheets, icons, JavaScript, and a test index page that serves as a basic template for any views you may need to add to your application.

D. *Titanium*

Titanium is mainly used for native application development for mobile environment. It consists of an SDK that provides the necessary tools, compilers, and APIs for building for the target platform, and a visual environment for managing development. It utilizes web technologies that are both trendy and powerful, including AJAX, HTML5, CSS3, and jQuery. *Titanium* is available for Mac, Linux and Windows. Developing for the Android requires the Android SDK and can be done using Mac, Windows, or Linux. The *Titanium* framework comes with a platform-independent API that can make applications feature-rich because it can access advanced features such as touchscreens, cameras, GPS, navigation, contacts, storage, and much more. *Titanium* also supports augmented reality features like Screenshot, Shake and Record Video.

E. *Rhodes*

It supports cross-platform web application development in HTML, CSS, JavaScript and Ruby. Its tool can be used across Mac, Windows, and Linux. The user interface of app is created using HTML and CSS. It requires Apple SDK for iPhone or iPad with Mac OS for development of app. For android Mac, Windows or Linux can be used. The Android native development environment is required, but no need of Eclipse IDE. Blackberry is java based with windows to run its tools and no need for eclipse. Rhodes support windows mobile 6 but not 7. MS Visual studio not used with Rhodes. Device capabilities supported by Rhodes in different platform are geolocation, contacts, camera, date\time picker, audio\video capture, Bluetooth, SMS, Landscape orientation, and native maps.

F. *RhoSync*

Mobile user can access information even in offline mode on device due to synchronization servers. It is a sync server framework that provides web services to Rhodes based app running on smartphones. It is data stores that stores information as an object, attributes values and works as a middle tier between web services and mobile app. In *RhoSync* ruby support query based information retrieval, data submission, creation, deletion, updating and user authentication.

G. *jQueryTouch*

It supports HTML pages that look like a native iPhone app animated transition, swipe detection and themes for HTML based web app are aided by *jQueryTouch*. It can influence cross-platform such as PhoneGap and Rhodes. It is a source code lib that includes JavaScript and CSS. Creating a new app is *jQueryTouch* is being simple, but the modification is difficult.

H. *Adobe Air*

It uses the same technology to build web application for different platforms but this framework is not suitable for an app that requires high computation resources. Irrespective of operating and browser get access to the services of the same site, app portability, and rich user interaction. JSON used for data transfer between web server and mobile app. It provides a rich internet application with desktop software and network capabilities and complete control over app by the user. It is a runtime program with no specific language and heavyweight solution. It supports flash, flex, JavaScript, AJAX and HTML.

I. *Processing*

It is popular for artist, designers and from the edge of productivity, supporting different platforms. It is Java based. Ketai provides Android hardware features in processing. It provides a simple, straight forward, wide range of libraries and tools for highly interactive app. It is available for Windows, Mac, Linux, and support Open GL. It does not support some of the advance features of Eclipse IDE.

VI. DISCUSSION

Well-known operating system and rich set of software development tools are used to drive mobile devices such as cell phones, smartphones and tablet. Each operating system and applications running on that OS works in a different way, e.g. Android applications developed by using Java and run in its own process, while iOS application built by using Objective-C run directly on iOS. Since there is proliferating diversity in devices used by the students, therefore development language and framework should meet the requirements of the heterogeneous device environment. In this section we discuss possible solution for implementation of on-line \ off-line quizzes and web based assignments in mobile learning system to enhance traditional learning practices under a given scenario.

There is a need of framework that is platform independent and creates an app that not depends on particular device requirements. With the discussion in section III and conferring to requirements of learning tools, we consider hybrid app category is suitable for on-line \ off-line quizzes and for web based assignments in the mobile learning system specifically for the one we proposed in [18]. Because they provide web based interface with offline working capacities and access to some of the device features. Hybrid app support highest number of cross-platform [9]. PhoneGap [6] for mobile learning system development support forum, assignment, chat, resources, file upload and download.

Likewise an open platform LMS [1] with middleware based architecture and rich client application using PhoneGap, HTML5 and JavaScript for user interface also support many different student role functionalities on heterogeneous client devices. Mobile voice agent [9] integrated IBM Worklight and PhoneGap with speech platform for audio query and audio summary of query results. Application development with Adobe Air [8] is browser based rich internet application that can be accessed at any location on any device. Sencha Touch use to develop presentation layer and reasonable graphical user interface for mobile app [10], is hardware independent and can run on different devices.

From the literature review given in the last paragraph and online information⁶ and many others, we conclude our discussion on the effectiveness of hybrid app cross platforms for online \offline quizzes and web based assignments. The choice of platform depends upon requirements, for example, if the requirement is a simple hybrid mobile app with web interface Phone Gap is a better choice but for rich application with animation, videos, etc. Adobe AIR gives better performance and application support. PhoneGap is suitable for a complete app rather than just user interface of a browser based app as in Sencha touch, jQTouch and iWebKit. Rhodes suitable for cross-platform, but provide a single code base that has compiled in target platform language and load app in the simulator. Most of the primary app development languages are not supported by Rhodes. Processing is also suitable for rich sensor based app; it is lightweight and easily installable but it is defined for Android in detail. How it will actually work in a cross-platform environment that is not explicitly defined.

Quiz and Assignment Tool

Since quizzes and assignments include many features such as creation of quiz \ assignment content, sending assignment to students on due date, assignment submission within the due date, conducting quiz online or offline in class, result generation, display of result and answer key etc. as mentioned in fig 2. Therefore a client server based mobile app for these tasks could be differentiated as web based assignment tool and offline \ online mode quiz tool supporting different end users on their heterogeneous devices. One of the most important features that are data transfer between devices in offline mode with Bluetooth technology is not discussed in mobile learning systems [1][6][7]. During the development, in order to fulfill requirements of our system we will explore cross-platform frameworks specifically PhoneGap, Adobe Air, Sencha Touch, Processing and aCME (for web based assignment) further as most of them are Java based, Java is compatible development language to almost all mobile platforms as shown in table 1 and has API for Bluetooth technology. Other than that, except for Processing all other frameworks support many varied mobile platforms. Either we will use them individually or use a combination of frameworks for implementation of the application in a heterogeneous device environment.

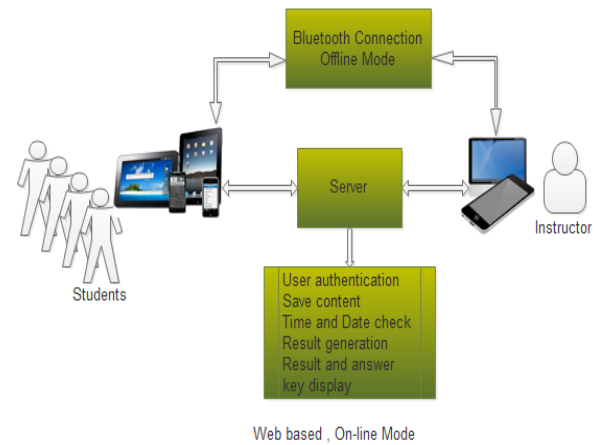


Fig. 2. Basic Tasks in Quiz and Assignment Tool

VII. CONCLUSION

In this paper, we discussed cross-platform mobile learning development with special focus on quizzes and assignments. Through the comparison and contrast we have proposed that adopting a hybrid app technique is most suitable for such mobile app developers. From this study, we have learned that there are several frameworks that are feasible for hybrid cross-platform mobile app with each of them having some pros and cons. In order to avoid re-implementation of the same applications for different platform separately, it is vital to compare and make use of the strength of individual framework or the combination of frameworks, so that a better, more compatible and more stable cross-platform application can be developed. Studies also show that there are many existing learning systems that are not being explored fully and will probably have a bright future if some more functionalities are considered and proper implementation is carried out. Therefore, in our future work we will still continue the studying of the latest cross-platform framework, device features accessible by these frameworks, and implement a learning system incorporating new features. Hence, this study is important for a cross-platform mobile learning application development.

REFERENCES

- [1] D. V. Sanchez, E. H. Rubio, E. F. Ruiz, and A. M. Viveros, "Student Role Functionalities Towards Learning Management Systems As Open Platforms Through Mobile Devices", IEEE 2014
- [2] I. Dalmaso, S. K. Datta, C. Bonnet, and N. Nikaein, "Survey, Comparison and Evaluation of Cross Platform Mobile Application Development Tools".
- [3] N. Serrano, J. Hernantes, and G. Gallardo, "Mobile Web Apps", IEEE Software 2013
- [4] N. M. Hui, L. B.. Chienget, W. Y. Ting, H. H. Mohamed, and M. Rafie, "Cross-Platform Mobile Applications for Android and iOS", IFIP WMNC, IEEE 2013
- [5] P. Gokhale, and S. Singh, "Multi-platform Strategies, Approaches and Challenges for developing Mobile applications", CSCITA 2014
- [6] D. H. Setiabudi, L. J. Tjahyana, and Winsen, "Mobile Learning Application Based On Hybrid Mobile Application Technology Running On Android Smartphone and Blackberry".

⁶ <http://www.developereconomics.com/pros-cons-top-5-cross-platform-tools/>

- [7] B. A. Babatunde, L. Chia-Feng, and Y. Shyan-Ming, "A Cross-Platform Mobile Learning System Using QT SDK Framework", Fifth International Conference on Genetic and Evolutionary Computing 2011
- [8] J. F. Lu, and Y. Zhang, "Mobile Application Development Based on Adobe AIR", IEEE 2013
- [9] D. Jaramillo, V. Ugave, R. Smart, and S. Pasricha, "Secure Cross-Platform Hybrid Mobile Enterprise Voice Agent", IEEE 2014
- [10] Z. Ji, X. Zhang, I. Ganchev, and M. O'Droma, "Development of a Sencha-Touch mTest Mobile App for a mLearning System", IEEE 13th International Conference on Advanced Learning Technologies 2013
- [11] D. Malandrino, I. Manno, et al "A Tailorable Infrastructure to enhance Mobile Seamless Learning", IEEE Transactions on Learning Technologies, 2014
- [12] A. Kosmaczewski, "Mobile javascript application development", O'Reilly Media, Inc., 2012
- [13] H. P. Frank, and F. Reichert, "Mobile Phone Programming and its application to wireless networking", Springer 2007
- [14] S. Allen, V. Graupera, and L. Lundrigan, "Pro Smartphone Cross-Platform Development", Springer 2010
- [15] P. MacIntyre, B. Danchilla, and M. Gogala, "Mobile PHP", Pro PHP Programming, pp 31-35, 2011
- [16] D. Sauter, "Rapid Android Development Build Rich, Sensor-Based Applications with Processing", The Pragmatic Bookshelf, 2012
- [17] S. Wachenfeld, M. Madeja, and X. Jiang, "Developing Mobile Multimedia Applications on Symbian OS Devices", Mobile Multimedia Processing Lecture Notes in Computer Science Volume 5960, pp 238-263, 2010
- [18] N. Litayem, B. Dhupia, and S. Rubab, "Automatic Attendance and Mobile Learning System in Sensor Enabled Heterogeneous and Dynamic University Environment", International Journal of Emerging Technology and Advanced Engineering Vol 4 Issue 12 Dec 2014

Fault-Tolerant Attitude Control System for a Spacecraft with Control Moment Gyros Using Multi- Objective Optimization

Ai Noumi

School of Science for Open and Environmental Systems
Keio University
Yokohama, Japan

Misuzu Haruki

Aerospace Research and Development Directorate
Japan Aerospace Exploration Agency
Tsukuba, Japan

Takuya Kanzawa

Aerospace Research and Development Directorate
Japan Aerospace Exploration Agency
Tsukuba, Japan

Masaki Takahashi

Department of System Engineering
Keio University
Yokohama, Japan

Abstract—Recent years have seen a growing requirement for accurate and agile attitude control of spacecraft. To both quickly and accurately control the attitude of a spacecraft, Control Moment Gyros (CMGs) which can generate much higher torque than conventional spacecraft actuators are used as actuators of the spacecraft. The drive on the motors is needed for rapid maneuverability, negatively affecting their life. Thus, in designing spacecraft the conflicting requirements are rapid maneuverability and reduced the drive on motors. Furthermore, the attitude control system needs to be fault-tolerant. The dominant requirement is different for each spacecraft mission, and therefore the relationship between the requirements should be shown. In this study, a design method is proposed for the attitude control system, using multi objective optimization of the skew angle and parameters of the control system. Pareto solutions that can show the relationship between the requirements are obtained by optimizing the parameters. Through numerical analysis, the effect with fault-tolerance and parameter differences for the dominant requirement are confirmed and the method to guide for determining parameters of the attitude control system is established.

Keywords—Control Moment Gyros; Spacecraft; Attitude Control; Multi-objective Optimization

I. INTRODUCTION

These days spacecraft require rapid rotational maneuverability because of the diversity and complexity of missions. Rapid rotational agility as well as a precision steady attitudinal state are required for the attitude control of spacecraft[1]. Rapid rotational agility as well as a precision steady attitudinal state are required for an attitude control of spacecraft. To meet this demand, Control Moment Gyros (CMGs) are ideal as an attitude control actuator of an agile spacecraft. Compared with previously used actuators, for example Reaction Wheels (RWs), CMGs can effectively generate higher torque. Many methods have been proposed to solve CMG's specific singularity problem[2].

The pyramid-type four-CMG system, as shown in Fig. 1, is commonly used with a skew angle set to 54.74 degree. In actual operation, it is necessary to combine several CMGs for redundancy. Skew angle is usually selected as 54.74 degree because the maximum angular momentum for each axis in Fig. 2 is the same. However, it is not necessary for the three axes to have the same angular momentum in the case of a spacecraft such as earth observation satellites whose mission angle is fixed. In fact, the skew angle is set to 30 degree for Pleiades-HR1 because a roll slew maneuver is assumed to be the main mission[3].

Therefore the author previously proposed optimizing the skew angle and parameters of the control system in Fig. 3 to achieve the shortest settling time, assuming a specific mission[4]. However, the load on the motor and bearings cannot be ignored for a long term mission because the drive on the gimbal motor is needed to shorten the settling time merely to achieve rapid maneuverability. At the same time reduced drive on CMG is required for spacecraft.

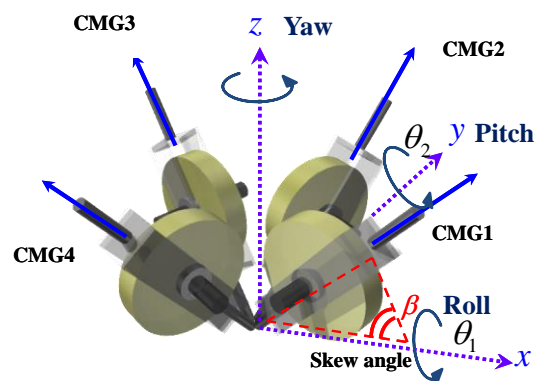


Fig. 1. Skew array CMG system

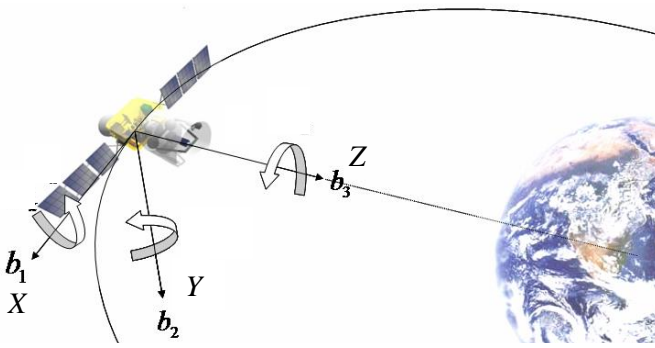


Fig. 2. Coordinate of spacecraft

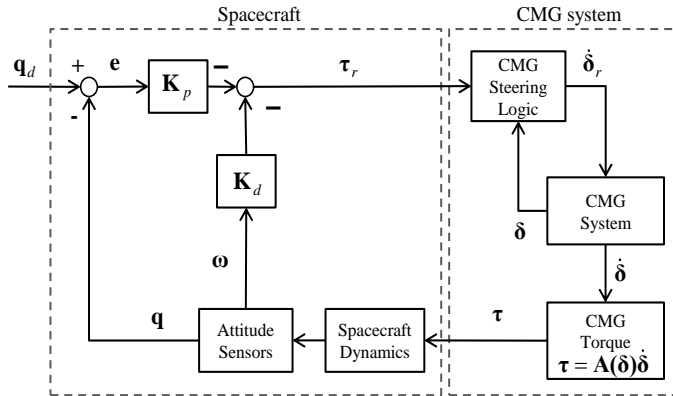


Fig. 3. Attitude control system of agile spacecraft with CMG system

The failure of the ISS CMGs in 2002 was mainly caused by excessive accumulation of the load as a result of metal fatigue of the gimbal axis, especially in the bearing[5],[6]. Therefore, reduced the drive on the gimbal can be effective in extending the operating life of CMGs, as well as reducing power consumption. To achieve rapid maneuverability is important for spacecraft with CMGs but it is not necessary to achieve the shortest settling time because the mission should be achievable within a given preset time. Therefore it is appropriate to design the attitude control system taking into account the conflicting requirements, which are to achieve rapid maneuverability and reduce the drive on the gimbal, for a long operating life.

In this study, multi-objective optimization of skew angle and parameters of control system considering conflicting requirements is proposed. The Pareto solutions considering conflicting objectives are obtained using optimization of parameters. The relationship between requirements and parameters is shown by calculating Pareto solutions from the optimization of parameters. As a specific design method, the settling time and motion of gimbal axes are evaluation criteria when considering rapid maneuverability and reduced the drive on CMG. The combination of parameters that minimizes both evaluation values is obtained by multi objective genetic algorithm (MOGA)[7].

In numerical simulation, the three types of optimization were conducted in addition to the proposed method, as comparative methods. From simulation results, the changes in the parameters with or without consideration of the drive on

the gimbal, the effectiveness of optimizing skew angle, effect with fault-tolerance and the parameter differences for the dominant requirement are confirmed. Therefore, the method to guide for determining parameters of the attitude control system is established.

II. CMG SYSTEM

A. Schema of CMG

CMG is an actuator that can generate torque using the gyro effect with a swinging wheel that rotates at a constant rate in a gimbal axis, perpendicular to the axis of wheel rotation. In actual operation, it is necessary to combine several CMGs for redundancy. In this study, an agile spacecraft is considered to have a CMG system that has a pyramid arrangement of four single-gimbal CMGs as shown in Fig. 1.

B. Modeling of CMGs

A block diagram representation of the CMG-based attitude control system of the agile spacecraft is illustrated in Fig. 3. When a target angle is required, a torque command vector is calculated using both the current Euler angle and angular velocity vector of the spacecraft ω , which are detected by the spacecraft's own sensors of angular position and velocity. The CMG gimbal angular velocity vector command, which is needed to achieve the torque command, is calculated using the equation of inverse kinematics, named steering logic[8]. Torque is generated from the gimbal angular velocity which, in turn, is generated by activating CMGs to follow the gimbal angular velocity vector command.

The attitude quaternion error vector $\mathbf{e} = [e_1 \ e_2 \ e_3 \ e_4]^T$ is computed using the quaternion of the reference angle of spacecraft $\mathbf{q}_d = [q_{1d} \ q_{2d} \ q_{3d} \ q_{4d}]^T$ and the quaternion of the current angle of the spacecraft $\mathbf{q} = [q_1 \ q_2 \ q_3 \ q_4]^T$, as follows:

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} q_{4d} & q_{3d} & -q_{2d} & -q_{1d} \\ -q_{3d} & q_{4d} & q_{1d} & -q_{2d} \\ q_{2d} & -q_{1d} & q_{4d} & -q_{3d} \\ q_{1d} & q_{2d} & q_{3d} & q_{4d} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{bmatrix} \quad (1)$$

The linear state feedback controller is defined as:

$$\begin{aligned} \boldsymbol{\tau}_r &= -\mathbf{K}_p \mathbf{e} - \mathbf{K}_d \boldsymbol{\omega} \\ &= - \begin{bmatrix} 0 & K_p & 0 & 0 \\ 0 & 0 & K_p & 0 \\ 0 & 0 & 0 & K_p \end{bmatrix} \mathbf{e} - \begin{bmatrix} K_d & 0 & 0 \\ 0 & K_d & 0 \\ 0 & 0 & K_d \end{bmatrix} \boldsymbol{\omega} \end{aligned} \quad (2)$$

where K_p is the proportional controller gain, and K_d is the derivative controller gain, of the spacecraft and $\boldsymbol{\tau}_r$ is the torque command vector. The angular momentum for four skew type CMGs shown in Fig. 1 is a function depending on the CMG gimbal angle vector $\boldsymbol{\delta} = [\delta_1 \ \delta_2 \ \delta_3 \ \delta_4]^T$ as follows:

$$\mathbf{h} = h_{CMG} \sum_{i=1}^4 \mathbf{H}_i(\delta_i)$$

$$= h_{CMG} \begin{bmatrix} -c\beta \sin \delta_1 - \cos \delta_2 + c\beta \sin \delta_3 + \cos \delta_4 \\ \cos \delta_1 - c\beta \sin \delta_2 - \cos \delta_3 + c\beta \sin \delta_4 \\ s\beta \sin \delta_1 + s\beta \sin \delta_2 + s\beta \sin \delta_3 + s\beta \sin \delta_4 \end{bmatrix} \quad (3)$$

where h_{CMG} is the angular momentum of the CMG wheel, \mathbf{H}_i ($i=1,2,3,4$) is the angular momentum vector of the i th CMG, β is the skew angle of the four CMGs, $c\beta = \cos \beta$ and $s\beta = \sin \beta$. A time derivative of the CMG angular momentum vector is given by:

$$\dot{\mathbf{h}} = h_{CMG} \begin{bmatrix} -c\beta \cos \delta_1 & \sin \delta_2 & c\beta \cos \delta_3 & -\sin \delta_4 \\ -\sin \delta_1 & -c\beta \cos \delta_2 & \sin \delta_3 & c\beta \cos \delta_4 \\ s\beta \cos \delta_1 & s\beta \cos \delta_2 & s\beta \cos \delta_3 & s\beta \cos \delta_4 \end{bmatrix} \dot{\boldsymbol{\delta}}$$

$$= \mathbf{A}(\boldsymbol{\delta}) \dot{\boldsymbol{\delta}}$$

$$\equiv \boldsymbol{\tau} \quad (4)$$

where $\dot{\boldsymbol{\delta}} = [\dot{\delta}_1 \ \dot{\delta}_2 \ \dot{\delta}_3 \ \dot{\delta}_4]^T$ is the CMG gimbal angular velocity vector, \mathbf{A} is a 3×4 Jacobian matrix, and $\boldsymbol{\tau}$ is the torque vector. From Eq. (4), the gimbal angular velocity vector command is calculated and determines the gimbal angular velocity to generate the torque command. This is called Steering Logic[8] and is an inverse kinematics equation for calculating the gimbal angular velocity vector command. Most simple CMG steering logic uses a pseudo inverse matrix of \mathbf{A} :

$$\dot{\boldsymbol{\delta}}_r = \mathbf{A}^T (\mathbf{A}\mathbf{A}^T)^{-1} \boldsymbol{\tau}_r \quad (5)$$

where $\dot{\boldsymbol{\delta}}_r$ is the gimbal angular velocity command vector, $\mathbf{A}^+ = \mathbf{A}^T (\mathbf{A}\mathbf{A}^T)^{-1}$, which is often referred to as the pseudo inverse steering logic. Most CMG steering laws determine the gimbal rate commands with some variant of the pseudo inverse. If the rank of $(\mathbf{A}) < 3$ for certain sets of gimbal angles, or, equivalently, the rank of $(\mathbf{A}\mathbf{A}^T) < 3$, the pseudo inverse does not exist and the pseudo inverse steering logic encounters singular states. This singular situation occurs when all the individual CMG torque output vectors are perpendicular to the commanded torque direction. Equally, the singular situation occurs when all the individual CMG momentum vectors have external projections onto the commanded torque vector direction. This is needed to avoid singular states when the chance of failure increases because the value of the gimbal angular velocity vector command is extremely high and would be a strain on the gimbal axis of the CMG at the singularity.

In this study, Generalized Singularity Robust (GSR)-Inverse logic[8],[9], proposed by Bong Wie, is used for the steering logic. This is a method for avoiding a singularity by generating a torque with given gimbal angular velocity command, even in a singular situation, as follows. The GSR-Inverse steering logic can be represented as:

$$\dot{\boldsymbol{\delta}}_r = \mathbf{A}^\# \boldsymbol{\tau}_r \quad (6)$$

$$\mathbf{A}^\# = \mathbf{A}^T [\mathbf{A}\mathbf{P}\mathbf{A}^T + \lambda \mathbf{I}_4]^{-1} \mathbf{A}^T \mathbf{P}$$

$$= \mathbf{A}^T [\mathbf{A}\mathbf{A}^T + \lambda \mathbf{P}^{-1}]^{-1}$$

$$= \mathbf{A}^T [\mathbf{A}\mathbf{A}^T + \lambda \mathbf{E}]^{-1} \quad (7)$$

$$\mathbf{P}^{-1} \equiv \mathbf{E} = \begin{bmatrix} 1 & \varepsilon_3 & \varepsilon_2 \\ \varepsilon_3 & 1 & \varepsilon_1 \\ \varepsilon_2 & \varepsilon_1 & 1 \end{bmatrix} > \mathbf{0} \quad (8)$$

where $\varepsilon_i = \varepsilon_0 \sin(\omega t + \phi_i)$ ($i=1,2,3$), $\lambda = \lambda_0 \times e^{-\mu m^2}$, $\omega = \pi/2$, $\phi_i = (1-i) \times (\pi/2)$ ($i=1,2,3$). t is the current time, and $m = \sqrt{\det(\mathbf{A}\mathbf{A}^T)}$ is the singularity measure. ε_0 , λ_0 and μ are constant parameters to be properly selected. The GSR-Inverse logic is used for singularity-avoidance steering logic in this study.

The equation to calculate the gimbal angular acceleration of i th CMG $\ddot{\delta}_i$ is obtained from:

$$\ddot{\delta}_i = K_g (\dot{\delta}_{ri} - \dot{\delta}_i) / J \quad (i=1, \dots, 4) \quad (9)$$

where K_g is the feedback gain of the CMG control system, $\dot{\delta}_i, \dot{\delta}_{ri}$ are the gimbal angular velocity and command of i th CMG and J is the inertia moment matrix of the CMG wheel.

III. DESIGN OF A FALUT TOLERANT ATTITUDE CONTROL SYSTEM CONSIDERING CONFLICTING REQUIREMENTS

The purpose of this chapter is to describe a specific design method for a fault-tolerant attitude control system, considering conflicting requirements. Initially, the parameters to be optimized are shown. Second, the conflicting requirements, such as rapid maneuverability and reduced drive on gimbals, are defined as the evaluation functions. Finally, the optimization of the parameters is defined using the Multi Objective Genetic Algorithm (MOGA) taking fault-tolerance into consideration.

A. Design Parameters

1) Skew angle

The maximum angular momentum of each axis can be changed by changing skew angle β in Fig. 1. For a typical pyramid configuration of four single-gimbal CMGs with a skew angle of β , the angular momentum for the three axes $\mathbf{h} = [h_x \ h_y \ h_z]^T$ can be obtained analytically as[3],[9]:

$$h_x = \frac{c\beta(-s\beta u_z + c\beta u_x)}{n_1} + \frac{u_x}{n_2} + \frac{c\beta(s\beta u_z + c\beta u_x)}{n_3} + \frac{u_x}{n_4} \quad (10)$$

$$h_y = \frac{u_y}{n_1} - \frac{c\beta(s\beta u_z - c\beta u_y)}{n_2} + \frac{u_y}{n_3} + \frac{c\beta(s\beta u_z + c\beta u_y)}{n_4} \quad (11)$$

$$h_z = \frac{s\beta(-c\beta u_x + s\beta u_z)}{n_1} + \frac{s\beta(s\beta u_z - c\beta u_y)}{n_2} + \frac{s\beta(s\beta u_z + c\beta u_x)}{n_3} + \frac{s\beta(s\beta u_z + c\beta u_y)}{n_4} \quad (12)$$

$$n_1 = \pm\sqrt{1-(s\beta u_x + c\beta u_z)^2} \quad (13)$$

$$n_2 = \pm\sqrt{1-(s\beta u_y + c\beta u_z)^2} \quad (14)$$

$$n_3 = \pm\sqrt{1-(-s\beta u_x + c\beta u_z)^2} \quad (15)$$

$$n_4 = \pm\sqrt{1-(s\beta u_y + c\beta u_z)^2} \quad (16)$$

where $u_x = \sin \theta_2$, $u_y = -\sin \theta_1 \cos \theta_2$ and $u_z = \cos \theta_1 \cos \theta_2$. θ_1 and θ_2 are the rotation angles of two successive rotations about the x and y axes. From these equations, it is apparent that the maximum angular momentum is directly related to the skew angle β .

Fig. 4 shows the relationship between the skew angle and the maximum angular momentum $\mathbf{H}=[H_x \ H_y \ H_z]^T$ for each axis in normal time when the skew angle was changed from 0 degree to 90 degree every 10 degree and 54.74 degree, which is commonly used. From Fig. 4, it can be seen that the maximum angular momentums of the roll and pitch axes decrease, whereas the maximum angular momentum of the yaw axis increases with the increasing skew angle. Furthermore, it is also apparent that the maximum angular momentums of the three axes are almost the same when the skew angle is 54.74 degree. However, the skew angle needs to be designed taking into consideration the requirements for this study, which are rapid maneuverability and reduced drive on the gimbals. Moreover, the conventional design only considers normal situations, whereas the control system should be designed to consider fault-tolerance when failure of CMGs has been reported during the operation.

A method for dynamically changing the skew angle when the spacecraft is in use has also been proposed[10],[11] because the maximum angular momentum of each axis can be changed by changing the skew angle, which is a valid method when using CMGs. However, the design of a unique skew angle before the launch, is proposed in this study, because potential failure of the added moving element must be taken into consideration when dealing with the failure of a CMG. In this study, it is assumed that one CMG can be shut down entirely in use.

It is apparent that CMG shutdown can be classified into two patterns, the failure of CMG 1 or CMG 3, or failure of CMG 2 or CMG 4. Figs. 5 and 6 show the relationships

between the skew angle and the maximum angular momentum in each situation. These figures show that the maximum angular momentums for every axis decrease than when the CMGs are functioning normally. Moreover, it is the same as the normal situation in that the maximum angular momentums of the roll and pitch axes decrease, while the maximum angular momentum of the yaw axis increases with an increasing skew angle. Comparing the failure of CMG 1 or CMG 3 with the failure of CMG 2 or CMG 4, it is apparent that the values of H_x and H_y switch place, although the value of H_z remains the same, which is verified by a deformation Eqs (10) to (16). For that reason, three situations, normal, failure of CMG 1 and failure of CMG 2, are dealt with in this study.

2) Parameters of the control system

In this study, both the skew angle and parameters of control system are tuned simultaneously because there is the possibility that the parameters of system, which can achieve the preset goal within target time, can be changed depending on the skew angle.

The parameters to be designed are the gains of the spacecraft attitude control system K_p and K_d , the gain of the CMG control system K_g , the parameters of the GSR-Inverse logic, λ_0 , ε_0 and μ .

B. Method for design of a fault-tolerant attitude control system considering the conflicting requirements

In this section, the method to determine skew angle and parameters of the control system using MOGA is outlined. As shown in the previous section, the unique skew angle and the six parameters of the control system which are appropriate to each situation are optimized.

First, several initial chromosomes are generated for 19 parameters, as shown in Fig. 7. They contain the gains of the spacecraft attitude control system K_p and K_d , the gain of the CMG control system K_g , the parameters of the GSR-Inverse logic λ_0 , ε_0 , and μ in each of the three situations and the unique skew angle.

$$10 \leq \beta \leq 55$$

$$30 \leq K_p \leq 200$$

$$50 \leq K_d \leq 250$$

$$0.27 \leq K_g \leq 0.285$$

$$-4 \leq \log_{10} \lambda_0 \leq 4$$

$$-4 \leq \log_{10} \varepsilon_0 \leq 4$$

$$-4 \leq \log_{10} \mu \leq 4$$

The range of the skew angle is set from 10 degree to 55 degree which provides the large maximum angular momentum for the roll and pitch axes from the results in Fig. 4.

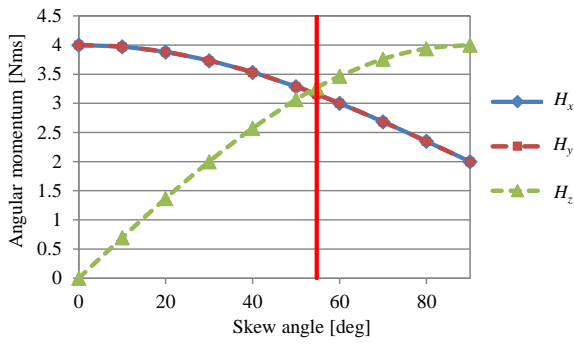


Fig. 4. Relationship between the skew angle and the maximum angular momentum

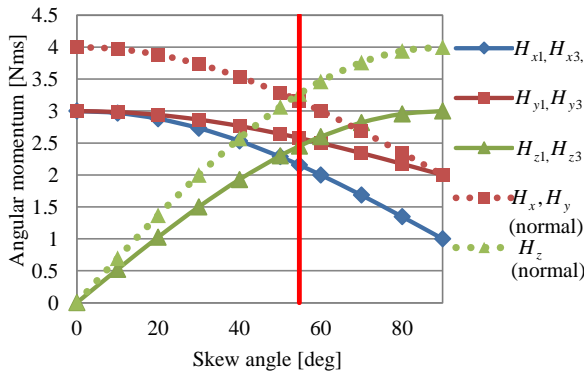


Fig. 5. Relationship between the skew angle and the maximum angular momentum (Failure of CMG 1 or CMG 3)

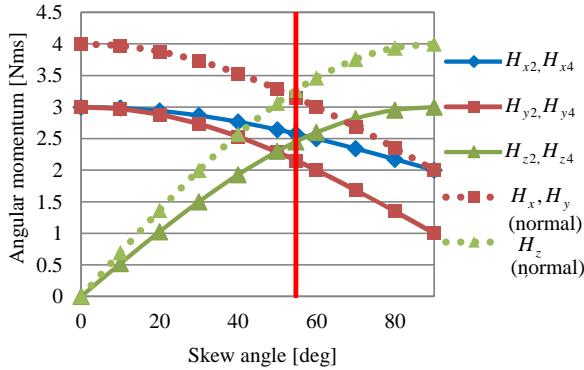


Fig. 6. Relationship between the skew angle and the maximum angular momentum (Failure of CMG 2 or CMG 4)

The settling time t_j and drive on the gimbal (gimbal angle, velocity and acceleration) in each of the three situations are evaluated for each chromosome. Where j means the situation, $j=1$ is the normal situation, $j=2$ is the situation of a failure of CMG 1 and $j=3$ is the situation of a failure of CMG 2. The settling time and drive on the gimbal for each chromosome are defined as f_1 and f_2 , and these are the summation of the evaluation value for each situation. Here the weight of the evaluation value for each situation a_j is designed according to the preferred situation. In this paper, $a_j=1$ to deal every situation equivalently.

$$f_1 = \sum_{j=1}^3 f_{1j} / \sum_{j=1}^3 a_j \quad (17)$$

$$f_{1j} = a_j (t_j / t_{\max}) \quad (18)$$

$$f_2 = \sum_{j=1}^3 f_{2j} / \sum_{j=1}^3 a_j \quad (19)$$

$$f_{2j} = a_j (f_{\delta} + f_{\dot{\delta}} + f_{\ddot{\delta}}) \quad (20)$$

$$f_{\delta} = \frac{1}{4t_j} \sum_{i=1}^4 \sum_{k=0}^{t_j/dt} \left(\frac{\delta_i(kdt)}{\delta_m} \right)^2 \quad (21)$$

$$f_{\dot{\delta}} = \frac{1}{4t_j} \sum_{i=1}^4 \sum_{k=0}^{t_j/dt} \left(\frac{\dot{\delta}_i(kdt)}{\dot{\delta}_m} \right)^2 \quad (22)$$

$$f_{\ddot{\delta}} = \frac{1}{4t_j} \sum_{i=1}^4 \sum_{k=0}^{t_j/dt} \left(\frac{\ddot{\delta}_i(kdt)}{\ddot{\delta}_m} \right)^2 \quad (23)$$

where δ_i , $\dot{\delta}_i$ and $\ddot{\delta}_i$ are the i th gimbal angle, velocity and acceleration. $\delta_{\max} = 6$, $\dot{\delta}_{\max} = 1$ and $\ddot{\delta}_{\max} = 3$ because the values of the terms should be the same. In Eqs (21) to (23), the

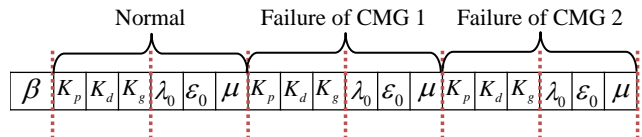


Fig. 7. Gene of an individual for genetic algorithm

summation of the data for each gimbal ($i=1, \dots, 4$) is divided by 4 t , which means the evaluation value is the average of the drive on the gimbal of a CMG for 1 s.

Pareto solutions are obtained by optimizing the combination of parameters which minimizes both evaluation values with the evaluation function f_1 for the settling time and the evaluation function f_2 for the drive on the gimbal using MOGA[7].

In this study, the mission assumed is a rest-to-rest maneuver, which means that the spacecraft body must rest at the beginning and end of a maneuver to observe a ground target, and is analyzed assuming a rigid body spacecraft. The spacecraft model is assumed to be a middle-sized earth-observing satellite whose directional axis is the yaw axis. The parameters used in the numerical simulations are given in TABLE II. The mission assumed for the analysis is to perform a 60 degree roll (cross-track) slew maneuver based on an actual earth-observing satellite, the ALOS-2[12]. The settling time is defined as when the Euler angle has settled to the target maneuver angle within ± 0.01 degree, and is the same for the pitch and yaw axes. The parameters of the CMG are the same as the CMG, 15-4S, in Pleiades-HR1 manufactured by Astrium[13].

IV. NUMERICAL ANALYSIS

Error! Reference source not found. shows the Pareto solutions obtained using MOGA, referred to in the last section.

The three types of optimization were conducted in addition to the proposed method, as comparative methods. 0 shows the comparison of each method. **Error! Reference source not found.** shows the comparative results from the three methods.

A. The changes in the parameters with or without consideration of the drive on the gimbal

In this section, the change in the optimized parameters when the evaluation function of the drive on the gimbal is or is not considered is compared for Methods 1 and 3. TABLE IV. shows the average and standard deviation of parameters for normal situation in each method.

From **Error! Reference source not found.**, the solution which satisfies smaller drive on gimbal than Method 1 when settling time is same as Method 1 is obtained in Method 3. From TABLE IV. smaller skew angle than for Method 3 is obtained in Method 1 to make settling time shorter because the angular momentum is larger for roll axis. On the other hand, the solution which satisfies smaller drive on the gimbal with same settling time than Method 1 is obtained even though skew angle in Method 3 is larger than in Method 1. It is because that appropriate parameters of the control system which satisfies both requirements are obtained using the evaluation function for the drive on the gimbal in addition to for the settling time.

B. Difference with and without optimizing skew angle

In this section, comparing Methods 2 and 3, the changes in the values of the evaluation functions are discussed when the skew angle is optimized and when it is set at 54.74degree.

In Method 3, skew angle is selected at 31.98 ± 7.64 degree which is smaller than 54.74 degree. It is noted that the settling time could have been shorter by optimizing the skew angle compared with when it was set at 54.74 degree. This is because a larger torque can be generated when skew angle is smaller as shown in the previous section. From this result, making the skew angle smaller can make the settling time shorter, but the drive on the gimbal could be greater when the settling time is longer for a skew angle of 54.74 degree. It is confirmed that skew angle can be designed according to requirements.

C. Effect with or without consideration of fault tolerance

In this section, comparing Methods 2 and 3, the effect with or without consideration of fault-tolerance is discussed.

From TABLE IV. skew angle is smaller when considering fault-tolerance. It can be assumed that smaller skew angle is obtained with fault-tolerance because the maximum angular momentum is smaller in failure situations than in a normal situation. Moreover, standard deviations of parameters in Method 4 are smaller than in Method 3. Fig. 9 shows the Pareto solutions for normal situation in Methods 3 and 4. From Fig. 9, the range of solution in Method 4 is smaller than in Method 3. It is assumed that range of solution for normal situation in Method 4 is smaller because of the consideration of fault-tolerance. Therefore, wider design of the attitude control

system in normal situation is available without consideration of fault-tolerance.

D. Parameter differences for the dominant requirement

In this section, parameter differences in Method 4 for the dominant requirement are discussed. Fig. 10 shows Pareto solution obtained by Method 4. The parameters in the normal situation for the three solutions in Fig. 10 which correspond to the dominant requirement as a discriminative solution of Pareto solutions in Method 4 are discussed:

- 1) Solution in which rapid maneuverability is dominant,
- 2) Solution in which both requirements are equivalently dominant,
- 3) Solution in which reduced the drive on the gimbal is dominant

TABLE V. shows the values of parameters in each solution. Figs. 11 to 13 show the time histories of the singularity parameter, gimbal velocity and torque as three solutions It is assumed that the solution satisfying the requirement of rapid maneuverability because the gimbal is driven rapidly in a singularity situation as a result of the combination of parameters of GSR Inverse logic $\lambda_0, \epsilon_0, \mu$, is obtained in I. Therefore torque errors for pitch and yaw axes are larger than other solutions to avoid singularity as quick as possible. In addition, it is assumed that the solution satisfying the requirement of reduced drive on the gimbal because the gimbal is driven slowly even in a singularity situation as a result of the combination of parameters of GSR Inverse logic $\lambda_0, \epsilon_0, \mu$, is obtained in III.

From these results, it is verified that the fault-tolerant attitude control system which satisfies the dominant requirement by changing characteristic of the control system by parameters of the control system.

TABLE I. THE VALUES OF THE SYMBOLS USED IN THE EVALUATION FUNCTIONS

Symbols	Values
t_{max}	40
δ_m	6
$\dot{\delta}_m$	1
$\ddot{\delta}_m$	3

TABLE II. PARAMETERS AND VALUES FOR THE NUMERICAL SIMULATIONS

Parameters	Symbols	Values
Inertia moment of spacecraft	I_s	diag(5000, 5000, 3000) kgm ²
Inertia moment of CMG wheel	J	0.19 kgm ²
Angular momentum of CMG	h_{CMG}	75 Nms
Max. gimbal rate	$\dot{\delta}_{max}$	1.0 rad/s
Max. gimbal acceleration	$\ddot{\delta}_{max}$	3.0 rad/s ²
Control cycle	dt	0.01 s

TABLE III. COMPARISON OF EACH METHOD

Method	Evaluation function		Fault-tolerance	Optimizing skew angle
	Settling time	Drive on gimbal		
1	○	×	×	○
2	○	○	×	54.74 degree
3	○	○	×	○
4 (Proposed)	○	○	○	○

TABLE IV. AVERAGE AND STANDARD DEVIATION OF PARAMETERS FOR NORMAL SITUATION IN EACH METHOD

Parameter	Average ± Standard deviation			
	Method 1	Method 2	Method 3	Method 4
Skew angle [deg]	18.1	54.74	31.98±7.64	28.77±0.58
K_p	171.15	101.97±28.38	70.33±15.53	92.37±3.77
K_d	299.1	282.68±15.59	211.27±28.38	247.10±4.24
K_g	0.284	0.281±0.004	0.277±0.0022	0.280±0.003
$\log_{10} \lambda_0$	-1.84	-0.21±0.57	-0.83±0.25	-0.92±0.11
$\log_{10} \varepsilon_0$	0.19	-0.99±0.25	-1.22±0.34	-1.12±0.15
$\log_{10} \mu$	-1.42	-1.66±0.72	-0.11±0.91	0.97±0.32

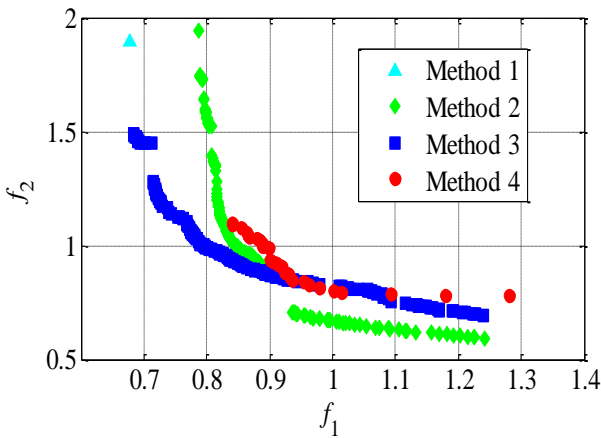


Fig. 8. Pareto solutions

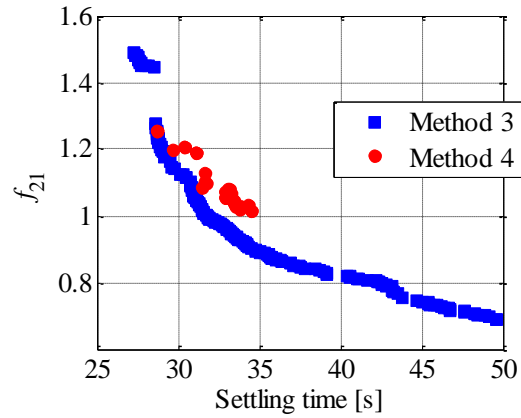


Fig. 9. Pareto solutions (Normal situation in Methods 3 and 4)

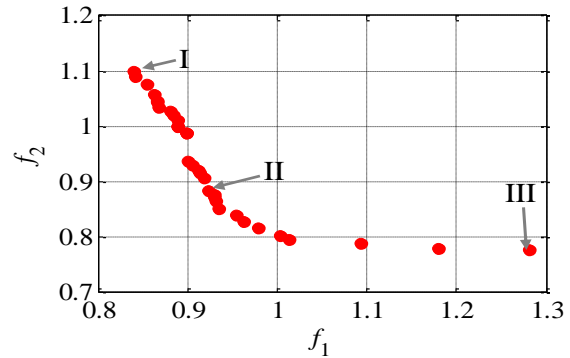


Fig. 10. Pareto solutions (Method 4)

V. CONCLUSIONS

This study proposed multi-objective optimization of the skew angle and parameters of the control system to design of a fault-tolerant attitude control system that would take into account conflicting requirements for a spacecraft with a skew array of Control Moment Gyros. The relationship between the requirements and the relationship between the requirements and parameters can be shown by calculating the Pareto solutions which is a class of solutions that comprehensively consider conflicting requirements.

From simulation results, the changes in the parameters with or without consideration of the drive on the gimbal, the effectiveness of optimizing skew angle, effect with fault-tolerance and the parameter differences for the dominant requirement were confirmed. Therefore, the method to guide for determining parameters of the attitude control system was established. To optimize the parameters considering constraint of spacecraft and to verify the effectiveness of the proposed method using an actual operation are the future works.

TABLE V. THE VALUES OF PARAMETERS IN EACH SOLUTION (METHOD 4)

	Settling time [s]	f_2	Skew angle [deg]	K_p	K_d	K_g	$\log_{10} \lambda_0$	$\log_{10} \varepsilon_0$	$\log_{10} \mu$
I	28.71	1.26	28.06	89.93	247.07	0.280	-0.86	-0.95	0.83
II	32.96	1.05	28.40	90.63	245.81	0.277	-0.97	-1.15	0.87
III	34.32	1.03	29.31	92.50	252.87	0.280	-1.09	-1.10	0.76

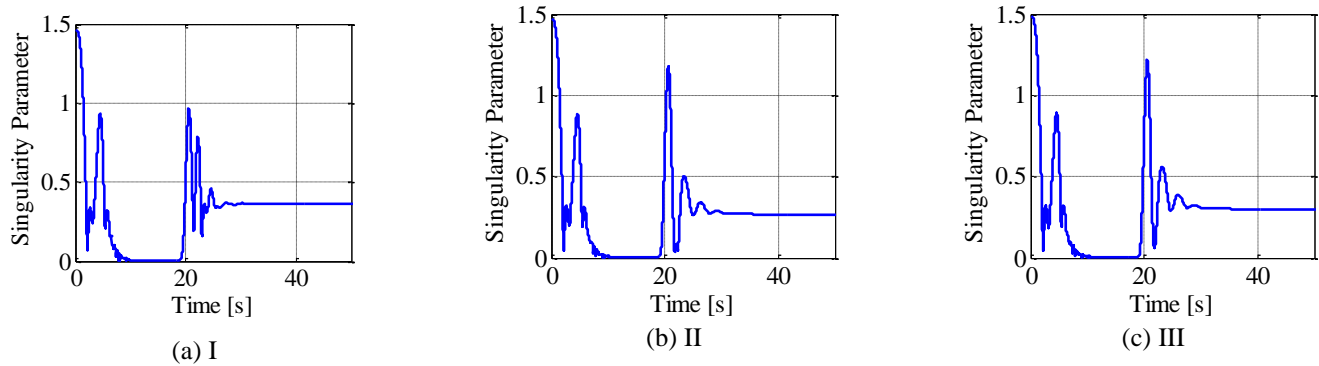


Fig. 11. Time histories of the singularity parameters

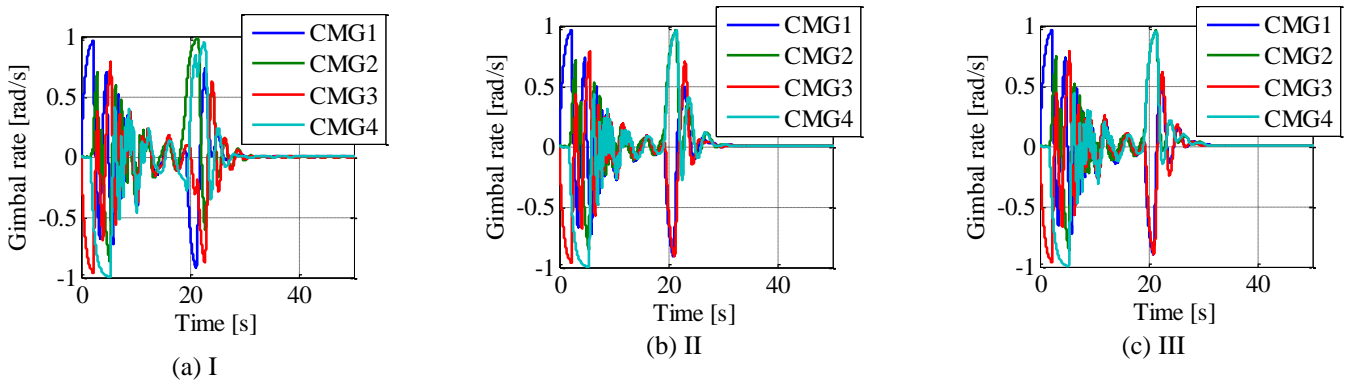


Fig. 12. Time histories of gimbal angle velocity

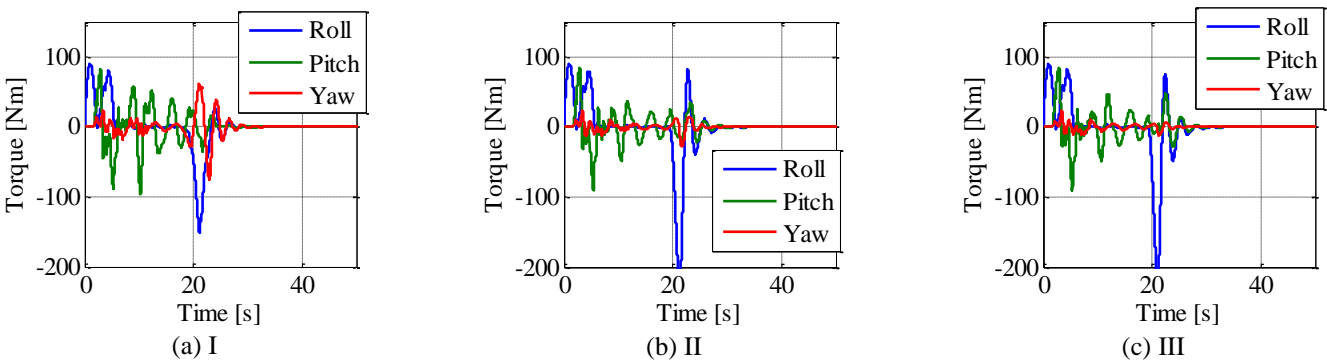


Fig. 13. Time histories of torque

REFERENCES

- [1] The Mechanical Social Systems Foundation, "Research on the realization of Earth observation satellite monitoring with advanced maneuver capability", 2009, pp. 57 – 61.
- [2] Kurokawa, H., "A geometric study of single gimbal control moment gyros - singularity problems and steering law -", Report of Mechanical Engineering Series, No.175, 1998.
- [3] Wie, B., Space Vehicle Dynamics and Control Second Edition, AIAA Education Series, American Institute of Aeronautics and Astronautics, Inc., 2008.
- [4] Noumi, A., Takahashi, M., "Fault-tolerant attitude control systems for a satellite equipped with control moment gyros", AIAA Guidance, Navigation, and Control Conference, AIAA, Boston, 2013, AIAA2013 – 5119.
- [5] Burt, R. R., Loffi, R. W., "Failure analysis of international space station control moment gyro", Proceedings of the 10th European Space Mechanisms and Tribology Symposium, Spain, 2003, pp. 13 - 25.
- [6] Gurrissi, C., Seidel, R., Dickerson, S., Didziulis, S. and Frantz, P., Ferguson, K., "Space station control moment gyroscope lessons

- learned”, Proceedings of the 40th Aerospace Mechanisms Symposium, 2010, pp. 161 - 176.
- [7] Deb, K., Pratap, A., Agarwal, S. and Metarivan, T., “A fast and elitist multiobjective genetic algorithm: NSGA- II”, IEEE Transaction on Evolutionary Computation , Vol.24, No.5, 2001, pp. 865- 872.
- [8] Wie, B., Bailey, D. and Heiberg, C., “Singularity robust steering logic for redundant single-gimbal control moment gyros”, Journal of Guidance, Control, Dynamics, Vol.24, No.5, 2001, pp. 865 - 872.
- [9] Wie, B., “Singularity escape/avoidance steering logic for redundant single-gimbal control moment gyros”, Journal of Guidance, Control, Dynamics, Vol. 28, No.5, 2005, pp. 948 - 956.
- [10] Kojima, H., Matsuda, N., and Takada, K., “Adaptive skewing pyramid-type CMGs for fast attitude maneuver”, Transactions of Japan Society for Aeronautical and Space Sciences, Space Technology, Japan, Vol.7, 2009, pp. 19 - 24.
- [11] Kojima, H., “Singularity analysis and steering control laws for adaptive-skew pyramid-type control moment gyros”, Acta Astronautica, Vol.85, 2013, pp. 120 - 137.
- [12] Kanzawa, T., Iwata, T., Arikawa, Y., Natori, T., “Attitude and orbit control system of the advanced land observing satellite-2 (ALOS-2)”, Proceeding of 55th Space Science and Technology Conference, The Japan Society for Aeronautical and Space Sciences, Ehime, 2011, JSASS-2011-4242.
- [13] Defendini, A., Faucheux, P., Guay, P., Morand, J., and Heimel, H. “A compact CMG products for agile satellites”, Proceedings of the 10th European Space Mechanisms and Tribology Symposium, Spain, 2003, pp. 27 - 31.

Golay Code Transformations for Ensemble Clustering in Application to Medical Diagnostics

Faisal Alsaby

Computer Science Department
The George Washington University
Washington, DC

Kholood Alnowaiser

Computer Science Department
The George Washington University
Washington, DC

Simon Berkovich

Computer Science Department
The George Washington University
Washington, DC

Abstract—Clinical Big Data streams have accumulated large-scale multidimensional data about patients' medical conditions and drugs along with their known side effects. The volume and the complexity of this Big Data streams hinder the current computational procedures. Effective tools are required to cluster and systematically analyze this amorphous data to perform data mining methods including discovering knowledge, identifying underlying relationships and predicting patterns. This paper presents a novel computation model for clustering tremendous amount of Big Data streams. The presented approach is utilizing the error-correction Golay Code. This clustering methodology is unique. It outperforms all other conventional techniques because it has linear time complexity and does not impose predefined cluster labels that partition data. Extracting meaningful knowledge from these clusters is an essential task; therefore, a novel mechanism that facilitates the process of predicting patterns and likelihood diseases based on a semi-supervised technique is presented.

Keywords—*medical Big Data; clustering; machine learning; pattern recognition; prediction tool; Big Data classification; Golay Code*

I. INTRODUCTION

Medical research is one of the most significant fields of science for people since no one is completely protected from physical ailments and biological degradation. It is not a surprise that health care is expensive. In 2010, the United States alone spent \$2.6 trillion in health care expenditures, nearly 17.9 percent of the United States gross domestic product (GDP). The expenses are projected to consume 19.9 percent of GDP by 2022 [1]. According to estimates, 3 million baby boomers will hit retirement age every year for the next 20 years, challenging an already stressed health care system [2]. Chronic diseases form an even bigger challenge, considering that more than 75 percent of health care expenditures are spent on people with chronic conditions [3]. Even though this number is high, it can be dramatically decreased by the power of prevention. Although we are able to generate and store enormous amounts of patients' medical data, physicians nowadays lack techniques that deal with Big Data challenge. More specifically, physicians are not capable of effectively quantify and analyze the relationship between medical data and causes of diseases, and predict the likelihood of diseases based on discovered patterns. However, risk is estimated by considering the patient's family history and the results of

necessary laboratory exams. This is highly dependent on the physician's limited experience. Therefore, this model of health care must be replaced with a new one that helps not only to early predict diseases but to prevent them even before patients show any symptoms. This paper presents a mechanism to encode the medical records patterns and generate the codewords that will be clustered by utilizing the perfect Golay code. This novel approach is suitable for processing continuous data streams [4]. With this clustering methodology, sensible information from underlying clusters can be extracted.

A cluster is defined as a data container with homogeneous data points inside of it. On the other hand, the data points from different clusters are non-homogenous. Technically, clusters isolate data points with boundaries such that the data points within the same cluster share common patterns or characteristics [5]. The Golay code clustering technique requires using vectors to represent any type of data, such as person's information, RNA sequencing, DNA sequencing, diseases, drugs and their side effects and so on. Each vector consists of 23-bit, where each bit represents the presence or the absence of a feature. For example, if a patient is tested positive to symptom x , it is represented in the vector as 1. Otherwise, the symptom is represented as 0. However, in some cases the proposed methodology provides the option of using Gray code property where 2 bits might be used to represent a single feature such as blood pressure level. Using Gray code property, this can be represented as 10, 00, or 01, to express high, normal, or low blood pressure level respectively. For the realization of this clustering method, a particular ontology approach must be considered which is called "Meta Knowledge of 23-bit Templates" [6]. These templates are an essential aid that assists in providing highly efficient clustering algorithm. The 23 questions needed to form vectors are included in 23-bit Meta knowledge template, which must not be arbitrary developed. In some circumstances the number of questions might be less than 23. Golay code clustering algorithm is distinctive, because of its linear time complexity and by allowing Fuzzy clustering. Therefore, it outperforms all other conventional clustering methods such as K-means. As a result, this method is an effective tool for handling the convoluted problems arising with the "Big Data" computational model in the medical field. Many medical applications might be considered in this regard. For instance, comparisons of protein and DNA sequences.

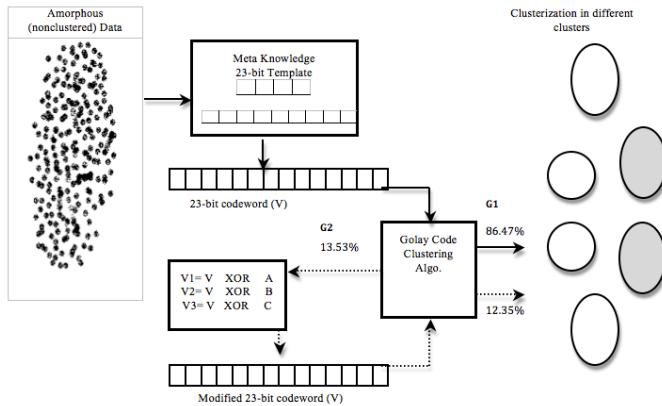


Fig. 1. Overview of Golay Code Clustering Method

This method can also be used to search sequences, find patterns, evaluate similarity and periodic structures based on local sequence similarity [7]. This paper is organized as follows: in section 2, we present some theoretical analysis to demonstrate the applicability of the proposed algorithm. In section 3, we discuss the proposed clustering algorithm. Section 4 presents the pattern recognition method. In section 5, experimental results on synthetic data are presented. Finally, section 6 contains our concluding remarks.

II. GOLAY CODE

The proposed clustering system is based on a reverse of the traditional error-correction scheme using the perfect Golay code (23, 12, 7) as described in [4]. Utilizing this perfect code, the whole set of 23-bit vectors is partitioned into 2^{12} spheres with radius 3. Thus, a transformation that maps the 23-bit string into the 12-centers of these spheres is able to tolerate certain dissimilarity in some bit positions of the 23-bit strings. Luckily, the Golay code is a perfect code that can tolerate up to three error bits [8]. Hence, this property allows adequate codewords to be associated with a single data word i.e. $\binom{23}{3} = 1771$ different codewords. The binary Golay code has a very large data word (2^{12} data words) and a larger codeword space ($2^{23} = 8,388,608$ codewords). This large space makes Golay Code appropriate for clustering. One interesting property of the Golay code scheme appears when decoding different codewords from the same sphere m . The different codewords will all be restored into the same data word. Hence, two random spheres $n1, n2$ will have one or more data words (indices) in common if and only if they have common hosting spheres. Therefore, the six data words that are associated with any n can be used to create clustering keys for the codeword n (Yu, 2011). For example, suppose we have two 23-bit vectors represented by two integers: 1036 ($2^{10} + 2^3 + 2^2$) and 1039 ($2^{10} + 2^3 + 2^2 + 2^0$). The two vectors differ in the two last bit positions. Their six hash indices turn out to be (0, **1054**, 1164, **1293**, 1644, 3084) and (527, **1054**, 1063, 1099, **1293**, 3215) respectively. The hamming distance between the code words 1036 and 1039 is 2, thus, they generate more than one identical index. This property guarantees that the two codewords are placed into a common cluster. As shown in the example, there are two common indices that are generated by both vectors,

1054 and 1293. Intrinsically, concatenating these two data words would provide us with a clustering address where both 1036 and 1039 would be placed in it. Such an approach leads to access the same cluster that contains both of them when searching for either pattern or their neighbors. Based on that, in order to utilize this clustering scheme, n must be restored back to six different data words. But the only way that n can be decoded into six different data words is when the center of n is 3 Hamming distance away from the hosting sphere [9]. In practice, using only one Golay code scheme results in clustering 86.5% of the total vectors (we call them $G1$) while the remaining 13.5% does not fit to this scheme (we call them $G2$). In other words, 86.5% of the vectors are able to generate the six data words (indices), which are required for the clustering process, while the remaining 13.5% of the codewords are not able to produce the necessary indices. One possible attempt of clustering $G2$ codewords is shown in Fig.1. We investigate the ability of each vector $V \in G2$ to generate the six indices. A tolerance of 1-bit mismatch can be implemented by probing each hash index corresponding to all 1-bit modification of a given codeword. Therefore, we create three 23-bit codewords A, B , and C where their values are the numbers 1,2 and 4 respectively. After that, by performing bitwise XOR operation between the original codeword and each one of the new codewords A, B and C , new vectors $V1, V2$ and $V3$ are created. As a result of applying Golay code hash transformation to these vectors, two situations are presented. In the first case, 12.35% of the modified $G2$ codewords are able to generate the six indices, thus the clustering method proceeds as normal. The remaining 1.17% can only generate one index; hence, in some circumstances, these codewords might be neglected [10]. Another way of clustering $G2$ codewords is based on using double Golay codes, which can be generated by the polynomials 2787 and 3189. Based on a previous work [4], this approach, however, is able to cluster 98.2% of the 8,388,608 codewords.

III. CLUSTERING COMPONENTS

A. Meta Knowledge Template

To facilitate the using of our clustering algorithm, a template of yes/no questions for each data item is necessary. A group of “23-bit Metadata Template” that is suitable for the medical case is designed. Questions should be based on acute physiological measurements. Each of these questions investigates the presence or absence of a property, a symptom, or a feature as shown in Fig.2. Moreover, this 23-bit Metadata Template can be utilized in a way such that complex values like DNA and RNA sequencing can be represented in the 23-bit codeword. One possible approach of designing such a template is to investigate DNA sequencings and find patterns and examine the correlation between diseases, mutations, Single Nucleotide Polymorphism (SNP), as well as the surrounding environment. Answering the questions results in a unique 23-bit vector V as in Fig. 1. V is then computed by the Golay code clustering algorithm where the output of this process is six different indices. A pairwiseing process for these six indices is applied to compose 15 cluster addresses. Subsequently, V is stored in each of the corresponding 15 clusters. This technique guarantees storing data items in one cluster if the difference between each two of them does not

exceed a certain number of bit-position mismatches. In other words, this clustering technique assures that the distance between any two vectors included in one cluster does not exceed a certain Hamming distance, as with Fig.2. It is important to recall that when mapping each codeword, we employ the binary Golay code, which guarantees that close decimal numbers have low Hamming distance in their binary representation. When applying our proposed clustering algorithm, all 23-bit codewords are classified into a number of clusters. The maximum Hamming distance within each cluster is either 7 or 8. The total number of bit positions that have common bit values within each cluster is either 15 or 16. This is specifically significant since it represents the total number of common attributes between codewords within a certain cluster. Put in mind that as bit positions may have different physical meanings, a low Hamming distance alone does not mean that two codewords are similar.

23-bit Metadata Template	
Q1	Is the patient a female?
Q2	Is the patient obese?
Q3	Does the patient have a family history of breast cancer?
Q4	Does the patient have BRCA1 or BRCA2 gene mutations?
.....
Q23	Did the patient receive any radiation treatments?

Fig. 2. 23-bit Metadata Template

As discussed above, codewords are created through answering the questions in the 23-bit Meta Knowledge Template. Each codeword consists of a 23-bit; each bit represents the presence or absence of a feature. Consequently, when two codewords have similar answers for the same questions within the template, these two codewords have similar features. Hamming distance is used to measure the similarity between codewords. Hamming distance between two codewords is the number of bits we must change to convert one codeword into the other. For example: the Hamming distance between the vectors **01101010** and **11011011** is 4. This methodology is considered one of the most simple, efficient, and accurate distance measures [11].

B. Composing Clustering Addresses

The overall clustering algorithm structure is shown in Fig.1. To illustrate the proposed methodology that uses the Golay code hash transformation, let V be a 23-bit codeword that is created by answering the question within the 23-bit Meta Knowledge Template. By using only one Golay code scheme and utilizing the Gray code property; the six 12-bit data words are generated for V . Clustering keys will be influenced by these six 12-bit data words. We start by choosing two arbitrary 12-bit data words of the 6 generated indices. Then, we order the selected two data words (such as, $w1 < w2$).

After that, we remove the least significant bit (LSB) of the smallest pair $w1$ and concatenate the result with the second data word to form a 24-bit A . After that, we shift A one bit to the right to get another 24-bit B . We then perform bitwise XOR operation between A and B to get a 24-bit C . The last 23-bit of C is the clustering key. The following algorithm shows how clustering keys are generated:

At least two common indices are generated by two 23-bit vectors at Hamming distance 2, as with the example aforementioned where the codewords were 1036 and 1039. Thus, when we pairwise (concatenate) these two common indices to generate the 23-bit clustering key, it is possible to place these vectors into the same cluster.

C. The Structure of Clusters

Clusters are essential components in our classification and prediction methodology due to its ability to discover the connected components of patients [12]. Because fuzziness is one of the most salient features of the “Big Data” concept, underlying relationships can be detected by using Golay code clustering technique. Furthermore, clusters assist in reducing the influence of patients who have little or no similarity i.e. common symptoms. When applying the Golay Code clustering algorithm to the possible 23-bit vectors (8,388,608 vectors), a total of 1,267,712 non-empty clusters were created. Each one of the generated clusters contains (139) or (70) codewords. For simplicity, we call them larger cluster (LC) and smaller cluster (SC) respectively. The maximum Hamming distance within each cluster is either 7 or 8. More importantly, the minimum total number of bit positions that have common bit values within each cluster is either 15 or 16. This is specifically a significant feature since it represents the total number of common attributes between codewords within a certain cluster.

Algorithm 1: Composing Clustering Addresses

1. generate the 6 data words
2. loop $i=1$ to 15
3. pick 2 random data words: $w1, w2$
4. order them such as $w1 < w2$
5. right shift the smallest data word such as $w1 >> 1$
6. $A = w1 \ w2$
7. $B = A >> 1$
8. $C_i = B \text{ XOR } A$
9. $\text{clustering_keys}[i] = C_i$ (only last 23 bit of C is used)
10. end loop
11. Return clustering_keys

For example, in Fig.2, the first two codewords have 19 features in common. More importantly, within each one of the SLs, 98.55% of the codewords have at least 17 common features, while the remaining codewords have either 16 or 15. On the other hand, 86.25% of the codewords in LCs have at least 17 common features, while 13.75% share either 16 or 15.

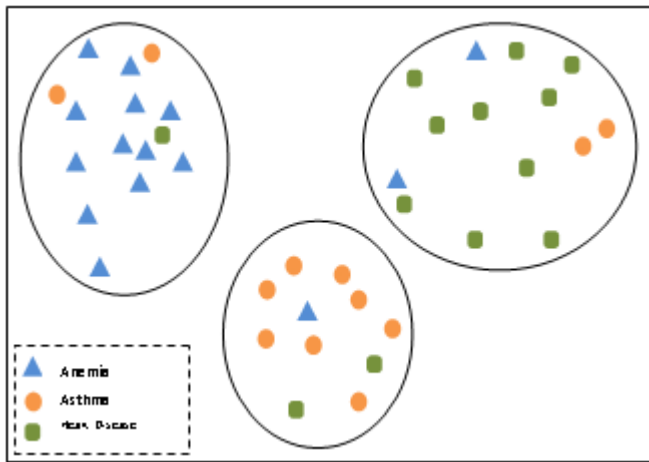


Fig. 3. Labeled Clusters based on the Majority Vote

IV. DATA ITEMS AND CLUSTERS LABELING METHOD

A. Training Method

Unlabeled data forms a major challenge that machine learning and data mining systems are facing [13][14][15]. Far better results can be obtained by adopting a machine learning approach in which a large set of N vectors $\{x_1, \dots, x_N\}$ called a training set is used to tune the parameters of an adaptive model[book]. Our pattern recognition procedure starts by training the system with a fully labeled training dataset (we call them centers). Specifically, the dataset is a collection of vectors that represent the identity of corresponding medical conditions or diseases, for instance, heart disease, Asthma, Breast cancer etc. These vectors will be employed to label objects that already were clustered. We sequence through clusters and find the nearest center to each clustered codeword in terms of Hamming distance. The label of the codeword is basically the exact label of the nearest center. When all codewords labeling process is fulfilled, labeling clusters becomes trivial. For example, assume that V_1 represents Asthma, V_2 represents Anemia, and V_3 depicts Heart diseases. Subsequently, we find the minimum Hamming distance between each vector in the system and V_1, V_2, V_3 . If the Hamming distance does not exceed a certain number of distortions, the vector's label is the same as the label of the nearest center. After labeling codewords, we rank objects among every cluster according to their frequency, regardless of whether they occur in other clusters within the system [14][15]. The label of the cluster depends solely on the majority weight within this cluster, i.e. prevalent element. Some clusters have different types where one type dominates that cluster or weighs more. Thus, the weight W_n of each object within a cluster is simply its frequency in that cluster.

$$w_n = \text{Frequency}[n]$$

W_n is the weight of the object n .

As a result, the vote of the majority within a cluster influences the label of the cluster. Noise is a factor that might reduce the accuracy of labeling process. Therefore, a threshold is recommended to insure high accuracy and efficiency. Cluster is granted the right to vote when it contains at least 10

codewords. Table (1) presents an example of the labeling process. Eventually, when the majority of clusters are labeled, the process of assigning a label to a new vector becomes a trivial. The label of a new vector is determined during the clusterization method. In particular, after attaching the new vector to the appropriate 15 clusters, its label will be assigned instantly. The assignment works by receiving a vote from each one of the 15 clusters i.e. the vote of a cluster is basically its label. Hence, the label of the new vector is the majority vote among its 15 clusters [16]. For example, if 10 out of 15 clusters are labeled with Asthma and 5 are labeled with Heart disease, the new vector is labeled with Asthma. Prior work indicates that the accuracy of the assignment is 92.7% [4].

TABLE I. IDENTIFYING THE PATTERN OF A NEW DATA ITEM

CLUSTER #	Object Frequency		Cluster size	Label
	Asthma	Anemia		
1	20	3	23	Asthma
2	12	2	14	Asthma
3	104	10	114	Asthma
4	14	1	15	Asthma
5	1	2	3< threshold	Ineligible for voting
6	65	6	71	Asthma
7	2	1	3< threshold	Ineligible for voting
8	15	7	22	Asthma
---	----	----	--	----
14	78	3	81	Asthma
15	30	2	32	Asthma
The new pattern is:				Asthma

V. CLASSIFICATION AND PREDICTION METHODOLOGY

This proposed approach is suitable for Big Data problems, because it requires less complex mathematical calculations. Not like other conventional methods that depend on performing complex probabilistic operations, which are time consuming and requiring large-scale computational capabilities. The approach is an efficient technique in a sense that smarter decisions can be made much faster for quick responses. To simply describe the prediction methods, assume that a codeword C is generated based on diagnosing a patient P and answering the 23-bit questions of the Meta knowledge template. Thus, C represents the symptoms S that P has or has not. Our prediction approach works as follows: C goes in a process of generating and composing the clustering keys which was described above. Then, a pointer to C is placed in each one of the 15 clusters. Two different ways of prediction and classification are presented. First prediction approach works by identifying the type of the disease that P might develop based on the majority vote among the 15 clusters. For instance, if 10 out of the 15 clusters were labeled with "Asthma", then P is most likely to develop asthma based on the current symptoms.

The second approach works by discovering relationships between symptoms based on other patients' metadata analysis. This relationship yields a prediction on the type of the S that P might develop in the future. For example, let A be the group of neighbor vectors. Vectors in A are placed with C in the same

cluster(s) and have no more than a certain Hamming distance, let's say 1. Then, we follow Fuzzy search method to retrieve codewords in A . After that, we sequence in A to find all the bit positions that mismatch with C , and place these mismatches in a group named L . As we described earlier, each bit represents the presence or absence of a property, which is in our example a symptom. Subsequently, we rank these symptoms in L based on their frequency. Therefore, our system can predict the S and their likelihood for a specific P based on the frequency of S . As a result, a symptom S with high frequency has high chance of occurrence in P and vice versa.

VI. CONCLUSION REMARKS

Formulating meaningful groups of scattered data is beginning to gain popularity in many fields, including the medical field. In fact, it is one of the most demanding fields due to the enormous amounts of data generated on a daily basis. In this paper, we presented an efficient medical Big Data processing model based on Golay Code clustering algorithm. Our Big Data methodology works by clustering diverse information items in a data stream mode. The result is a group of clusters where the data items in each cluster are homogeneous. In contrast, the data points from different clusters are non-homogenous. This technique improves our ability to extract knowledge and insights from large and complex collections of medical data. Granting all the clustering methods that have been published before, the proposed method surpasses others as it improves the time complexity to $O(n)$. We recommend the presented algorithm to be used as a tool in the medical field due to its competence in classification and prediction of risks, symptoms, and diseases.

REFERENCES

- [1] Centers for Medicare and Medicaid Services, Office of the Actuary (2012), National Health Expenditure Projections, 2012–2022. Washington, DC: CMS.
- [2] Barr, P. (2014, January). *The Baby Boomer Challenge. Hospitals & Health Networks*. Retrieved from http://www.hhnmag.com/display/HHN-news-article.dhtml?dcrPath=/templatedata/HF_Common/NewsArticle/data/HHN/Magazine/2014/Jan/cover-story-baby-boomers
- [3] Chronic Conditions: Making the Case for Ongoing Care. 2002. Johns Hopkins University. Baltimore.
- [4] D. Greene, A. Tsymbal, N. Bolshakova and P. Cunningham, "Ensemble Clustering in Medical Diagnostics," Proc. 17th IEEE Symp. Computer-Based Medical Systems (CBMS '04), pp. 576-581, 2004.
- [5] F. Alsaby and S. Berkovich. Realization of clustering with Golay code transformations. Global Science and Technology Forum, J. on Computing, 2014.
- [6] D. Liao, and S. Berkovich. "On clusterization of Big Data streams," Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications, article no.26. ACM press, New York 2012
- [7] H. Yu. Golay Code Clustering Using Double Golay Encoding Technique, Doctoral Dissertation, GWU, October 2011.
- [8] D. Davis, N. Chawla, N. Blumm., N. Christakis, and N. Barabasi. Predicting individual disease risk based on medical history, Proceedings of the 17th ACM conference on Information and knowledge management, October 26-30, 2008, Napa Valley, California, USA [doi>10.1145/1458082.1458185]
- [9] H. Yu, T. Jing, and S. Berkovich. Golay Code Clustering for Mobility Behavior Similarity Classification in Pocket Switched Networks, J. of Communication and Computer, USA, 2012.
- [10] W. Pearson, and D. Lipman, Improved tools for biological sequence comparison. Proc Natl Acad Sci USA, 1988. 85: p. 2444 - 2448.
- [11] S. Berkovich and E. El-Qawasmeh. "Reversing the Error-Correction Scheme for a Fault-Tolerant Indexing", The Computer Journal, Vol. 43, No. 1, pp. 54 – 64
- [12] M. Yammahi, K. kowsari, C. Shen, and Simon Berkovich. An efficient technique for searching very large files with fuzzy criteria using the Pigeonhole Principle.
- [13] A. Gruber1, S. Bernhart, I. Hofacker, and S. Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures, BMC Bioinformatics, Volume 9, 2008
- [14] A. Blum, and S. Chawla. Learning from Labeled and Unlabeled Data using Graph Mincuts, Proceedings of the Eighteenth International Conference on Machine Learning, p.19-26, June 28-July 01, 2001
- [15] M. Charkhabi, T. Dhot, and S.Mojarad. Cluster Ensembles, Majority Vote, Voter Eligibility and Privileged Voters International Journal of Machine Learning and Computing, Vol. 4, No. 3, June 2014
- [16] T. Yokoi, T. T. Yoshikawa and T. Furuhashi. Incremental learning to reduce the burden of machine learning for P300 speller, Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on , vol., no., pp.167,170, 20-24 Nov. 2012
- [17] G. Kowalski, Document and Term Clustering, In Information retrieval architecture and algorithms, p. 173, New York: Springer,2011
- [18] E. Berkovich. Method of and system for searching a data dictionary with fault tolerant indexing. US patent No.: US 7,168,025 B1 (2007)
- [19] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)
- [20] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- [21] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [22] K. Elissa, "Title of paper if known," unpublished.
- [23] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
- [24] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
- [25] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

A Monitoring Model for Hierarchical Architecture of Distributed Systems

Phuc Tran Nguyen Hong
Danang University of Education
The University of Danang
Danang, Vietnam

Son Le Van
Danang University of Education
The University of Danang
Danang, Vietnam

Abstract—Distributed systems are complex systems and there are a lot of the potential risks in the system, so system administrators need to have some effective support tools for network management. The architecture information of distributed systems is an essential part of distributed system monitoring solutions, because it provides general information about monitored objects in the system for administrators, as well as supports administrator in quickly detecting change of topology, error status or potential risks that arise during operation of distributed systems. The modeling approaches have an important role in developing monitoring solutions, in which they are background to develop algorithms for monitoring problems in distributed systems. This paper proposes an approach in order to model for hierarchical architecture of objects in distributed systems, in which consists of architecture of monitored objects, networks, domains and global distributed systems. Based on this model, a basic monitoring solution for hierarchical architecture of distributed systems is developed and this solution is able to provide administrators more important architecture information such as the topology of hardware components, processes, status of monitored objects, etc.

Keywords—architecture; distributed systems; model; monitored objects; monitoring

I. INTRODUCTION

Distributed systems (DS) are complex systems, which have always challenged for system administrator a lot [5,9]. A hardware malfunction, a faulty process or an abnormal event occurs on the system may affect other events taking place at different locations in the running environment of system. These symptoms can cause a bad effect on performance and stability of the system, they can also cause of errors of related processes and incorrect results of distributed applications. In order to ensure the effective operation of DS, global system information in general and information of each object in particular is critical issues. Many technical solutions have been researched and developed to support administrators in monitoring the system. Through the survey and review some typical monitoring works such as [10,11,13,14,15,16,17] in paper [4,8], in which we presented in detail the technical details, some advantages as well as disadvantages of these solutions. The survey result on implementation solutions and function of monitoring systems is presented in Table I and II.

TABLE I. THE IMPLEMENTATION SOLUTION

Monitoring System	Implementation Solution		
	Software	Hardware	Hybrid
JADE [13]	•		
MonALISA [11]	•		
MOTEL [17]	•		
ZM4/SIMPLE [14]			•
NON-INVASIVE MONITOR [8]		•	

We are aware that there are many implementation solutions to deploy monitoring system. However, with the advantages such as flexibility and mobility, the ease of maintenance, etc the software solution has been widely deployed in many TCP/IP monitoring products.

TABLE II. THE FUNCTION OF MONITORING SYSTEMS

Monitoring System	Monitoring Function			
	Computing	Performance	Object	Operation
ZM4/SIMPLE [14]	•			
JADE [13]	•			
MonALISA [11]		•		
SNMP [8,16]		•		
MOTEL [17]			•	
CorbaTrace [10]			•	
Tools (OS,...) [8]				•

From Table II, we see that the monitoring systems for DS can be divided into two groups: specific monitoring (SM) and general operations (GM) for monitored object in DS.

- SM consists of monitoring systems that monitor specific issues of monitored objects in DS such as MonALISA, MOTEL, SNMP, etc. SM can be seen as a special monitoring layer to monitor details such as traffic, performance, computing, etc. Most of these solutions in SM are only focused on solving the requirements for specific monitoring issues between

objects and have not yet been really interested in the global architecture of monitored objects in DS. For example, ZM4/SIMPLE is deployed to do performance evaluation for and parallel and distributed programs; MonALISA is deployed to monitor and help manage and optimize the operational performance of Grids; etc.

- GM consists of monitoring systems that monitor general operations of the monitored objects in DS such as built-in tools of devices or utilities in OS (Operating System). GM can be seen as a common monitoring layer in which provide abilities to monitor architectures and operations of monitored objects (MO) such as configuration, status, communication, connections, etc. For example, taskmgr and netstat commands are in Windows OS; prstat command is Solaris OS, etc.

Therefore, we can divide monitoring for DS into two basic stages:

- The first stage is general monitoring with monitoring solutions in GM, the global architecture information of monitored DS in general and the information about general operations of monitored objects in particular are essential in this stage, because they can support administrator for quickly detecting common errors or error domains that arise during operation of the system [4].
- The second stage is extended survey with monitoring solutions in SM in order to go into more detail in special monitoring information.

Thus, the monitoring solutions in GM are considered as a high level monitoring facilities to monitored DS before using other monitoring solutions in SM to deeper analysis. However, GM are now mainly based on tools (OS, utilities) that developed by device vendors side or operating systems side. These built-in tools have some disadvantages such as discrete monitoring information, independent of each device, etc [4,8], hence the global of DS cannot be solved with these built-in tools. The global architecture should be continued to research and develop more effective, the goal of the paper focus on solving this problem base on modeling for architecture of MO and building hierarchical monitoring entities respectively.

When monitored systems have basic changes about architectures, behaviors and operation environments, the technical solutions must be modified and updated appropriately for new changes and management requirements. As system specification methodology is generally and flexibly, the modeling approach is considered more appropriate for systems that have a lot of changes and the approach is widely used in discrete event systems, computer protocols [1,3,7]. In the DS, the modeling approach also achieved some certain results [2,6]. The modeling approaches play an important role, in which it is used as a basis layer for algorithm and solution development in monitoring, diagnosing and controlling issues independently. Therefore, the modeling for MO in DS is really necessary, the objective of the paper is based on the research results on DS and set theory [1,4,6], we focus on building a formal model for the hierarchical architecture of MO in DS, in which consists of architecture of monitored objects, networks, domains and

global distributed systems. We also present a basic monitoring model for the hierarchical architecture of DS, in which can show DS topology visually as well as the local operations and the communication operations of MOs in the DS.

The paper is organized as follows: In section II, we present architectural model for a MO in DS and the composition operation that allows us to combine many MOs into a composition model, we describe hierarchical architecture of monitored objects in DS. Section III focuses on the modeling solution that is able to monitor the architecture of DS. Finally, section IV concludes with the current work and future perspectives.

II. THE ARCHITECTURE MODEL FOR DISTRIBUTED SYSTEMS

DS consists of many heterogeneous devices such as stations, servers, routers, etc. These devices are considered physical objects in DS and communicate to each other in the system; each device consists of many hardware components such as CPU, HDD, etc. and software components such as processes. These components are associated with information about the corresponding states and behaviors, general operations of MO is described by Fig. 1, they can be divided into two basic parts such as internal part – local operations and external part – communication operations [4].

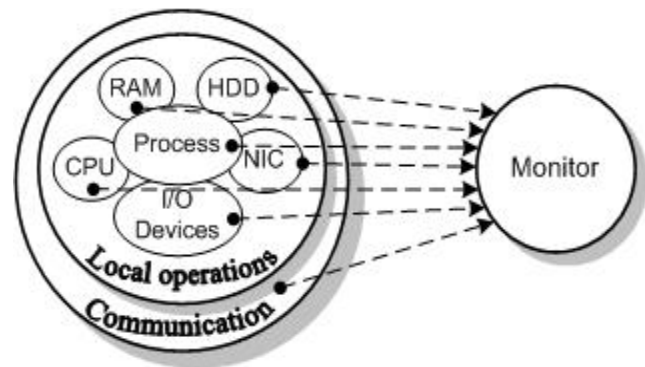


Fig. 1. General operations of the monitored object

- The local operations: these operations include processing, computing, resource requirements for process computations. The operations are locally performed within that object and use system resource such as CPU, RAM, etc. in running time.
- The communication operations: these operations are functions that interact with other objects in the system such as inter-process communication, controlling to interact with management system, etc. These operations are used to communicate with other objects on the system.

All of local and communication operations are based on system resource of MO such as CPU, RAM, I/O, etc. and information of these operations is dynamic in their running process, while system resource of MO is static information. Therefore, architecture of MO will consist of static information of MO and dynamic information of local and communication operations.

MOs are considered as nodes that are connected according to specific architecture and can perform interactive communication to each other. Hence, Architecture model describes the structure of nodes along with the related information of each node, the link between nodes, message propagation via its port, etc. Based on this information, we can determine the physical structure and the state of the nodes in the system.

From result of research on DS and monitoring systems, we can see that DS consists of many heterogeneous objects and topologies that communicate to each other. With point of view the domain-based management for large scale systems, the multi-level domain has been used to manage for DS [18], in which consists of local object level, network and domain level. The hierarchical architecture of monitored objects in DS can be presented as Fig. 2:

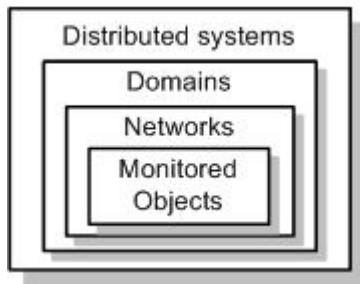


Fig. 2. The hierarchical architecture of objects in DS

Therefore, the architecture model of DS and monitoring model are presented in hierarchical architecture as Fig. 2 in order to deploy a suitable monitoring solution.

A. Architecture Model of MO

Let AM be an architecture model of monitored node, the AM is a 9-tuple and expressed as follows:

$$AM = (NODES, DOMAINS, NETS, LINKS, PORTS, port, communication, status, event) \quad (1)$$

$NODES = \{\text{set of static and dynamic information of nodes}\} = \{NODES_S\} \cup \{NODES_O\} \cup \{NODES_A\}$

where: $NODES_S$ consists of system resource information of MO and this is static information such as Cpu, RAM, etc; $NODES_O$ consists of information about local and communication operations such as processes; $NODES_A$ consists of error or abnormal information of hardware and software components such as I/O errors, overload; $NODES_O$ and $NODES_A$ are dynamic information.

$DOMAINS = \{\text{set of domain information such as name, ...}\}$

$NETS = \{\text{set of network information such as IP, network, ...}\}$

$LINKS = \{\text{set of link information between nodes}\}$

$PORTS = \{\text{set of port: internal and external port}\}$

$port$ is a function that identify communication ports in $NODES$: local ports (internal) and external ports (send/receive to nodes not in $NODES$), $port(NODES) \in PORTS$

$communication$ is a function that identify communication connections between nodes, $\{(NODES, PORTS) \rightarrow (NODES, PORTS \times d)\}$, delay $d = [t_{min}; t_{max}]$

$status$ is a function that identify node states in which consist of normal or abnormal status, $status(NODES) \in \{S_NOR\}$ or $\{S_ABNOR\}$, where: S_NOR is set of normal status such as up, communicating,...; S_ABNOR is set of abnormal status such as down, overload,...

$event$ is a function that identify node events such as request, messages,... These events consist of internal (internal_events) and external events (external_events)

In order to visually present architecture model, we denote AM for architecture model, $n \in NODES$, $d \in DOMAIN$, $net \in NETS$, $L \in LINKS$, $\{p_1, p_2\} \in PORTS$. So architecture model AM can be visually described as Fig. 3

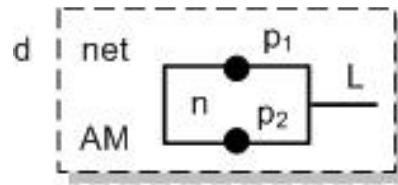


Fig. 3. The architecture model of a node

With this architecture model, we can determine the related information of node such as resource information, operations, status, etc base on elements of AM.

For example, give an architecture model AM of running node as Fig. 3 without communication operations, and then AM can be expressed as follows:

$$AM = (\{n\}, \{d\}, \{net\}, \{L\}, \{p_1, p_2\}, port, communication, status, event)$$

Where $\{n\} = \{\text{system information such as device name, CPU, ...}\} \cup \{\text{running processes, I/O operations, ...}\} \cup \{\text{error status, ...}\}$;

$port = \{\text{internal ports: } p_1, p_2\}$;

$communication = \{\text{no communication with others}\}$;
 $status = \{\text{up}\}$;

$event = \{\text{local operation events}\}$

Therefore, architecture model of monitored object will give us more important information about that object such as local operations (internal operations) as well as communication operations (external operations). Based on this architecture information, we can determine operations, errors or abnormal states that occur in running time of the node.

B. Composition Model

DS is complex system in which consists of many heterogeneous devices (nodes) and is organized according to hierarchical architecture as Fig. 2. So architecture model of DS will be set of architecture model AM of nodes in system. In order to ensure more efficient to build architecture model of DS, we use composition operation as described here.

Let AM_1, AM_2 be architecture model of node 1 and node 2 in system, let \parallel be composition operator (concurrent) for AM_1 and AM_2 . Composition operation is shown in Fig. 4.

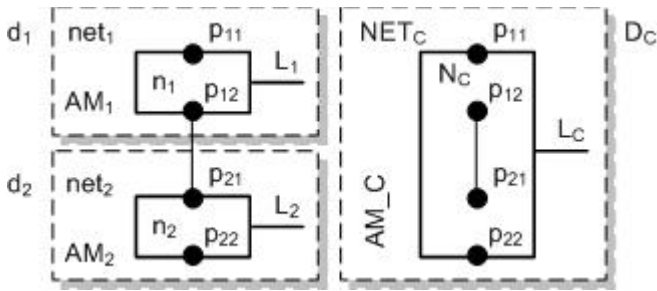


Fig. 4. Composition architecture of two nodes

The architectural model after composition AM_1 and AM_2 is AM_C , and AM_C is also a 9-tuple as expression (2).

$$AM_C = AM_1 \parallel AM_2 = (NODES_C, DOMAIN_C, NETS_C, LINKS_C, PORTS_C, port, communication, status, event) \quad (2)$$

Where:

$$NODES_C = NODES_1 \cup NODES_2 = \{N_C\} = \{n_1, n_2\};$$

$$DOMAIN_C = DOMAIN_1 \cup DOMAIN_2 = \{D_C\} = \{d_1, d_2\};$$

$$NETS_C = NETS_1 \cup NETS_2 = \{NET_C\} = \{net_1, net_2\};$$

$$LINKS_C = LINKS_1 \cup LINKS_2 = \{L_C\} = \{L_1, L_2\};$$

$$PORTS_C = PORTS_1 \cup PORTS_2 = \{p_{11}, p_{12}, p_{21}, p_{22}\};$$

$$port = port(NODES_C) = PORTS_C.internal \cup PORTS_C.external$$

$$\text{with } PORTS_C.internal = \{p_{12}, p_{21}\};$$

$$PORTS_C.external = \{p_{11}, p_{22}\}$$

$$communication = communication(NODES_C, PORTS_C)$$

$$= \{(n_1, p_{12}) \leftrightarrow (n_2, p_{21}), (n_1, p_{11}) \leftrightarrow (n_i, p_i), (n_2, p_{22}) \leftrightarrow (n_i, p_i)\}, i \notin \{1, 2\}$$

$$status = status(NODES_C) \rightarrow \{S_{NOR}\} \text{ or } \{S_{ABNOR}\}$$

where:

$$status(NODES_C) \in \{S_{NOR}\} \text{ when } status(n_1) \in \{S_{NOR}\} \text{ and } status(n_2) \in \{S_{NOR}\};$$

$$status(NODES_C) \in \{S_{ABNOR}\} \text{ when } status(n_1) \in \{S_{ABNOR}\} \text{ or } status(n_2) \in \{S_{ABNOR}\}$$

$$event = event(NODES_C) = internal_events(NODES_C) \cup external_events(NODES_C)$$

$$internal_events(NODES_C) = internal_events(n_1) \cup internal_events(n_2) \cup \{\sigma_{12}\};$$

$$external_events(NODES_C) = external_events(n_1) \cup external_events(n_2) - \{\sigma_{12}\}$$

with $internal_events(n_1)$: local events in node 1;

$internal_events(n_2)$: local events in node 2;

σ_{12} : communication events between node 1 and 2

Therefore, composition model AM_C describes operation information of two nodes in which consist operations of each node and communication between node 1 and node 2.

Similar to architecture information of MO, we can easily determine operations, errors or abnormal states of node 1 and node 2 that occur in running time based on elements in the model AM_C .

C. Modelling for Architecture of DS

As we presented in section II, topology of DS can be seen as hierarchical structure consists of many levels such as local object, network and domain level, in which global DS consists of n ($n > 0$) domains and can communicate with each other via telecommunication networks, each domain consists of m ($m > 0$) heterogeneous networks interconnect to each other, and each the network consists of k ($k > 0$) physical devices. All off them can collaborate, exchange and share information to each other. Therefore, the modeling for architecture of DS will be done with four levels: MO model, network model, domain and global DS model. The architecture model for DS can be expressed as follows:

- The architecture model of MO (AM_{MO}): AM_{MO} describe architecture information of MO and is expressed as follows:

$$AM_{MO} = (NODES_{MO}, DOMAIN_{MO}, NETS_{MO}, LINKS_{MO}, PORTS_{MO}, port, communication, status, event) \quad (3)$$

- The architecture model of a network (AM_{MS}): Give a network consists of k monitored objects $\{MO_1, MO_2, \dots, MO_k\}$ and set of $\{AM_{MO_1}, AM_{MO_2}, \dots, AM_{MO_k}\}$ is architecture model of these objects. Hence, AM_{MS} is a composition model of architecture model AM_{MO} s respectively:

$$AM_{MS} = AM_{MO_1} \parallel \dots \parallel AM_{MO_k} \quad (4)$$

From composition result of expression (2), AM_{MS} is expressed as follows:

$$AM_{MS} = (NODES_{MS}, DOMAIN_{MS}, NETS_{MS}, LINKS_{MS}, PORTS_{MS}, port, communication, status, event) \quad (5)$$

- The architecture model of domain (AM_{MD}): Similar to the AM_{MS} , give a domain consists of m networks corresponding to $\{AM_{MS_1}, \dots, AM_{MS_m}\}$, AM_{MD} is a composition model of AM_{MS} s respectively:

$$AM_{MD} = AM_{MS_1} \parallel \dots \parallel AM_{MS_m} \quad (6)$$

AM_{MD} is expressed as follows:

$$AM_{MD} = (NODES_{MD}, DOMAIN_{MD}, NETS_{MD}, LINKS_{MD}, PORTS_{MD}, port, communication, status, event) \quad (7)$$

- The architecture model of DS (AM_{DS}): As DS is a set of n domains $\{AM_{MD_1}, \dots, AM_{MD_n}\}$, so AM_{DS} is a composition model of AM_{MD} s respectively:

$$AM_DS = AM_MD_1 // \dots // AM_MD_n \quad (8)$$

AM_DS is expressed as follows:

$$AM_DS = (NODES_{DS}, DOMAIN_{DS}, NETS_{DS}, LINKS_{DS}, PORTS_{DS}, port, communication, status, event) \quad (9)$$

From expression (3)÷(9), we see that AM_MO, AM_MS, AM_MD and AM_DS are built from composing architecture model of basic objects. Thus, information of model AM_MO, AM_MS, AM_MD and AM_DS will describe all of system information, operations, links and state information (normal, abnormal, error) of elements in them. For example, related information of any network will describe in expression (5), so $NODES_{MS}$ will describe information of all MO in a network because $NODES_{MS} = NODES_1 \cup NODES_2 \cup \dots$ in which consists of system information, operations and error or abnormal information of all MO. Communication ports $PORTS_{MS}$ will display all of ports of objects in the network, because $PORTS_{MS} = PORTS_1 \cup PORTS_2 \cup \dots$. Therefore, in order to determine error or abnormal states of network according to AM_MS, we only observe $NODES_{AMS}$, because $NODES_{AMS} = NODES_{A1} \cup NODES_{A2} \cup \dots$.

III. THE MONITORING SOLUTION FOR HIERARCHICAL ARCHITECTURE OF DISTRIBUTED SYSTEMS

A. The Technical Base and Basic Monitoring Solution

The objective of the monitoring system is observation, collection and inspection information about the operations of the hardware and software components, communication events of MO. This information supports actively in system management.

The general monitoring architecture can be divided into 3 parts as Fig. 5.

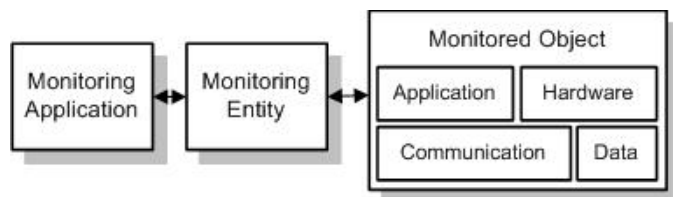


Fig. 5. General monitoring architecture

Monitored Object (MO) consists of independent objects such as switches, routers, workstations, servers, ... these objects have their own hardware and software resource. In order to describe architecture information of MO at time t, we use function $monitoring_info(MO, t)$.

Monitoring Application (MA) is designed to support for the management Objects (administrators or other management agents). MA entity interacts with monitoring entity to support the generation of monitoring requirements and present the results of monitoring are measured from monitoring entity.

ME (Monitoring Entity) is designed to instrument the monitored objects, the instrumentation information of the system will be processed to generate the corresponding

monitoring reports and send to MA. In order to describe result of monitoring entity ME at time t, we use function $result_ME(ME, t)$.

Thus, monitoring result of ME at time t for MO can be expressed as follows:

$$result_ME(ME, t) = monitoring_info(MO, t) \quad (10)$$

The monitoring system for DS consists of more MEs and MAs, they are not fixed and independently operate on each domain of DS, and monitoring information is exchanged between the MEs and MAs by message passing.

With the hierarchical architecture model of DS is presented as the previous session, hierarchical architecture of DS consists of four levels such as MO, network, domain and global DS. In order to collect the architecture information of DS, monitoring entities are designed in accordance with the hierarchical architecture of DS and we use four monitoring entities to monitor hierarchical architecture of DS:

- The monitoring entity ME_MO for object: ME_MO observes and collects the architecture information of MO. Because architecture model of MO is expressed as AM_MO in (3), the monitoring result of ME_MO at time t can be expressed as follows:

$$result_ME(ME_MO, t) = monitoring_info(AM_MO, t) \quad (11)$$

- The monitoring entity ME_MS for network: ME_MS observes and collects the architecture information of a network. Because architecture model of a network is expressed as AM_MS in (4), the monitoring result of ME_MS at time t can be expressed as follows:

$$result_ME(ME_MS, t) = monitoring_info(AM_MS, t) \quad (12)$$

- The monitoring entity ME_MD for domain: ME_MD observes and collects the architecture information of a domain. Because architecture model of a domain is expressed as AM_MD in (6), the monitoring result of ME_MD at time t can be expressed as follows:

$$result_ME(ME_MD, t) = monitoring_info(AM_MD, t) \quad (13)$$

- The monitoring entity ME_DS for distributed systems: ME_DS observes and collects the architecture information of DS. Because architecture model of DS is expressed as AM_DS in (8), the monitoring result of ME_DS at time t can be expressed as follows:

$$result_ME(ME_DS, t) = monitoring_info(AM_DS, t) \quad (14)$$

From expression (11)÷(14), the monitoring system for hierarchical architecture of DS will be set of monitoring entities $\{ME_MO, ME_MS, ME_MD, ME_DS\}$ that are designed as Fig. 6.



Fig. 6. Architecture of monitoring entities

The monitoring entities ME_{MO} will be installed on all of MO in DS, they observe and collect the architecture of MOs, and supply monitoring reports to network monitoring entity ME_{MS} . ME_{MS} runs composition operation in order to synthesize monitored information from ME_{MO} s in the same network and supply network monitoring reports to domain monitoring entity ME_{MD} . The operation of ME_{MD} and ME_{MS} has also run into similar processes as above. The monitoring implementation of ME_{MO} is designed as Fig. 7.

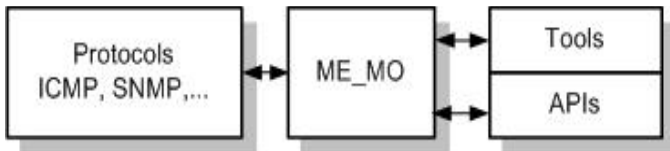


Fig. 7. The monitoring implementation of ME_{MO}

In order to observe and collect the architecture of MO in DS, we use protocols, APIs and built-in tools of operating system. The popular protocols are used in management network to monitor status or traffic of MO such as ICMP (Internet Control Message Protocol) [12,19], SNMP (Simple Network Management Protocol) [4,19]. The APIs and tools are used to observe and collect system information, operations as well as communication ports of components in MO such as the Window API, Linux API, libraries,

The modeling for monitoring solution bases on four levels such as MO, network, domain and global DS which are suitable with point of view the domain-based management, this hierarchical monitoring architecture have advantages to develop some distributed algorithms in levels of DS management in which the level MO focus on observing and collecting the architecture information of MO, level ME_{MS} , ME_{MD} and ME_{DS} are responsible for synthesizing and processing the monitoring information.

Therefore, the collection and composition process for building the architecture of DS is implemented as following sequence:

$$MO \rightarrow network \rightarrow domain \rightarrow global DS$$

The collection and composition process of hierarchical monitoring architecture are described detail in Fig. 8.

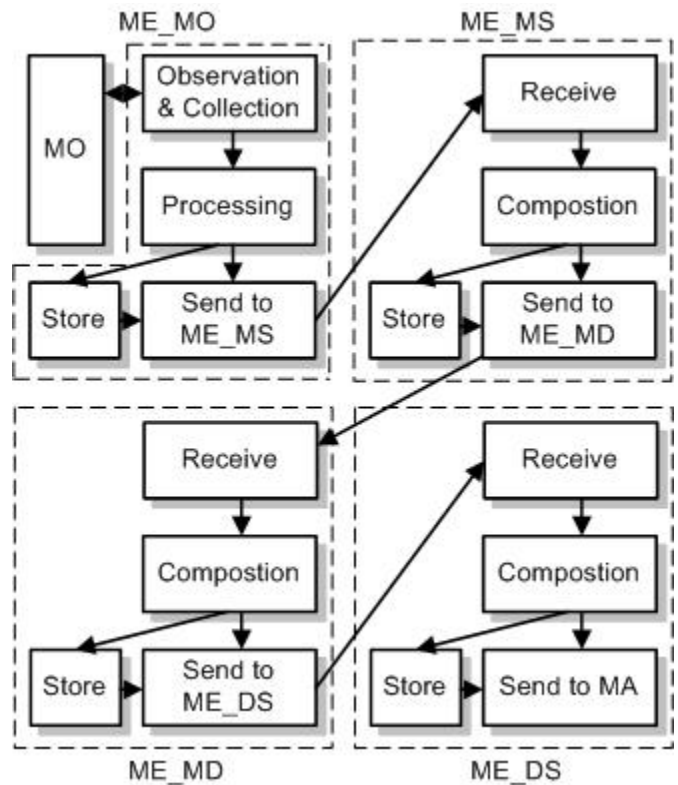


Fig. 8. Basic monitoring process for architecture of DS

At each level of monitoring entities (MO, network, domain, global DS) will collect full monitoring information of their monitored objects. First level, ME_{MO} collects and processes all of monitoring information of components such as Processes, CPUs, etc. Second level, ME_{MS} composes all of monitoring information of MOs in the same network and creates the monitoring report for architecture of this network. Third level, ME_{MD} composes all of monitoring information of networks in the same domain and creates the monitoring report for architecture of this domain. Fourth level, ME_{DS} composes all of monitoring information of domains in DS and builds the monitoring report for architecture of DS.

In order to analyze the architecture information of DS, the sequence of steps is implemented as follows:

$$global DS \rightarrow domain \rightarrow network \rightarrow MO.$$

For example, suppose that distributed system CDS consists of two domains $\{d_1, d_2\}$, each of domains contains one network: net_1 in domain d_1 , net_2 in domain d_2 , network net_1 consists two nodes $\{n_1, n_2\} \in NODES$, and network net_2 consists three nodes $\{n_3, n_4, n_5\} \in NODES$.

After the step ME_DS composes all of monitoring information for architecture of DS, we have all of architecture information of CDS that is expressed by the architectural model AM_DS in (9). Therefore, the architecture of CDS is analyzed as follows:

$$\begin{aligned} DOMAIN_{CDS} &= \{d_1, d_2\}; \\ domain(d_1) &= \{net_1\}; domain(d_2) = \{net_2\}; \\ NETS_{CDS} &= \{net_1, net_2\}; \\ net(net_1) &= \{n_1, n_2\}; \\ net(net_2) &= \{n_3, n_4, n_5\}; \end{aligned}$$

From above architecture information, the hierarchical architecture of CDS is presented as Fig. 9.

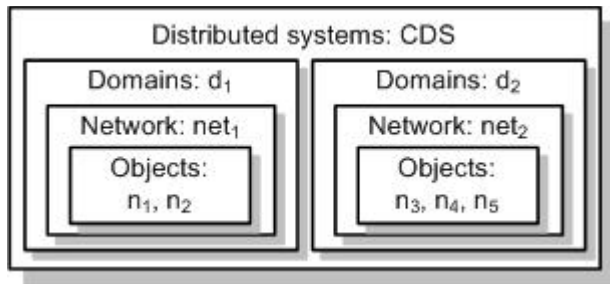


Fig. 9. The architecture of CDS

In normal case, all of monitored objects $\{n_1, n_2, n_3, n_4, n_5\}$ are running smoothly, set of information of objects in CDS contains in $NODES_{CDS}$ which consists of system resource information $NODES_{S_{CDS}}$, information about operations $NODES_{O_{CDS}}$ and error information $NODES_{A_{CDS}}$. Because the CDS has not any error, $NODES_{A_{CDS}}$ has not any description. Suppose that objects n_5 is down or overload, then $NODES_{A_{CDS}}$ contains down state or overload state of n_5 . Base on $NODES_{A_{CDS}}$, we will monitor all of errors or abnormal of CDS.

The connection and communication information of objects in CDS such as $LINKS_{CDS}$, $PORTS_{CDS}$, $port$, $communication$, etc will support us in building algorithms to display network visualization which consists of communication operations and link diagrams of nodes, networks and domains.

B. The Initial Experimental Results

Based on the model is presented in the previous sections, we designed a MCDS (Monitoring for Complex DS) system that consists of a set of monitoring entities (ME_MO , ME_MS , ME_MD , ME_DS as Fig. 6) for monitored objects, group of monitored networks, monitored domains and global system. The goal of MCDS is that monitor the architecture and operations information of devices on the VMSC3 system (a network system of VMS company at Vietnam), in which the architecture of monitored system can be displayed in hierarchical architecture as Fig. 2; operations information consists of local and communication operations (as Fig. 1) of monitored objects in VMSC3 system such as process, communication ports, etc.

The initial experimental results are shown in Fig. 10, in which presents some monitoring forms of MCDS such as

group of forms about basic architecture of objects and group of objects in VMSC3, as well as general description information about objects in system such as devices name, IP,...; group of forms about the communication and local operations information of monitored objects such as system information (descriptions, locations, OS...), hardware information (Cpu, Ram, I/O, ...) and information on the operations of the processes, status, communication, etc. This information is collected by ME_MO and will be used to send to other monitoring entities (ME_MS , ME_MD and ME_DS) by the message passing mechanism.

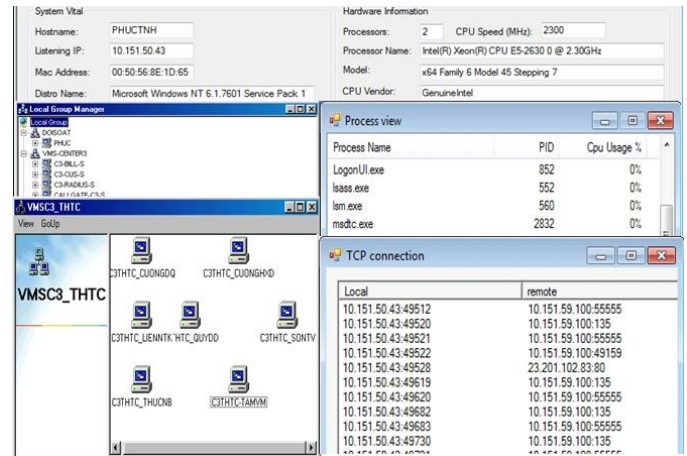


Fig. 10. MCDS for the monitored object in VMSC3

In order to evaluate this monitoring model for hierarchical architecture of distributed systems, we use some notation such as M_{our} for our model; M_{GM} for monitoring models is mainly based on tools (OS, utilities). Some evaluations as follows:

- **Monitoring presentation:** Because built-in tools only run object itself or by remote, so discrete monitoring information, independent of each device. Therefore, M_{GM} focuses on presenting monitoring information directly of objects MO in DS, it is only local part of DS. M_{our} presents monitoring results as the hierarchical architecture such as objects, networks, domains and global DS, so the presentation results consist of local part and global system, it provides an overview on monitored DS for administrators and is more appropriate for architecture of complex DS in the practical environment.
- **Monitoring function:** Solutions in SM group are only pay attention to solve special monitoring issues between MOs (computing, traffic, etc). M_{GM} focuses on general operations of MOs (devices, components) in DS, however this solution provides discrete monitoring information and have some disadvantages that we presented in [4,8]. M_{our} monitors general operations of MOs in DS with multi-levels, so it provides multi-level monitoring reports (local objects, networks, domains, etc). Therefore, M_{our} supports administrators in managing monitored objects in system more advantage and quickly detecting errors, potential risks arising during operation of DS based on elements of model at each level.

- **Implementation time:** Because most of the built-in tools in M_{GM} monitor DS by using discrete tools (OS or device vendors utilities), we have to type (or select) one or more commands respectively. M_{our} monitor based on MEs and MEs communicate to each other by the message passing. So, monitoring time with M_{our} will take less than M_{GM} . In order to evaluate for monitoring time, basic monitoring time for object O_i between M_{GM} and M_{our} are expressed as follows:

$$t_{GM}(O_i) = t_{Remote}(O_i) + \sum_{i=1}^n t_T(C_i) + \sum_{i=1}^n t_{R_P}(C_i) \quad (15)$$

$$t_{our}(O_i) = t_{mes}(O_i) + \sum_{i=1}^n t'_{R_P}(C_i) \quad (16)$$

Where: $t_{GM}(O_i)$ and $t_{our}(O_i)$ are monitoring time of M_{GM} and M_{our} for object O_i ; $t_{mes}(O_i)$, $t_{Remote}(O_i)$, $t_T(C_i)$, $t_{R_P}(C_i)$ and $t'_{R_P}(C_i)$: time for clicking function and monitoring message to object O_i , remote to object O_i , typing (or selecting) command C_i , running-presentation results of command C_i respectively.

Suppose that $t_{R_P}(C_i) = t'_{R_P}(C_i)$ for same monitoring function in M_{GM} and M_{our} with the same O_i .

The experimental results are implemented in VMSC3, in which nodes work on MS windows environment. Result consists of some cases as follows:

Monitoring implementation in object itself: $t_{Remote}(O_i) = 0$, $t_{mes}(O_i) \approx 2 \div 3s$ (clicking), $t_T(C_i) \approx 10 \div 20s$; with components as Fig. 2, we use about 7 commands respectively ($n=7$), so $\sum_{i=1}^7 t_T(C_i) \approx 70 \div 140s$, hence $t_{GM}(O_i) > t_{our}(O_i)$ (approximately $68 \div 137s$)

Monitoring implementation for a remote object on LAN: $t_{Remote}(O_i) \approx 20 \div 30s$, $t_{mes}(O_i) \approx 3 \div 4s$, $t_T(C_i) \approx 11 \div 22s$, so $\sum_{i=1}^7 t_T(C_i) \approx 77 \div 154s$, hence $t_{GM}(O_i) > t_{our}(O_i)$ (approximately $94 \div 180s$)

Therefore, we are easily aware that $t_{GA}(O_i) > t_{our}(O_i)$

When monitoring implements for a group of m objects on a network S_i : $t_{GA} = \sum_{i=1}^m t_{GA}(O_i)$ for M_{GM} and time for M_{our} is $t_{our} = \sum_{i=1}^m t_{our}(O_i) = t_{our}(ME_MS_{S_i})$, where: $ME_MS_{S_i}$ is a ME_MS for S_i and can be seen as a network object. Hence, the bigger monitoring time t_{GM} compares with t_{our} ($t_{GM} \gg t_{our}$) when the bigger m is.

The hierarchical architecture model of monitored objects in DS and experimental result show that our proposed model is feasible and will overcome the disadvantages of specific built-in tools in monitoring hierarchical architecture of DS, as well as actively support administrators in managing DS in according

to multi-level such as object level, network, domain and global DS level. Some actively results of above proposed model are presented in Table III.

TABLE III. SOME RESULTS BETWEEN BUILT-IN TOOLS (GM) AND MCDS

Issue	Specific built-in tool	MCDS
Monitoring function	Monitoring for general operations of MOs in DS, based on tools that developed by device vendors side or operating systems side	Monitoring for general operations of MOs in DS with multi-levels, based on set of monitoring entities: objects, networks, etc.
Implementation of monitoring requirements	Administrators must have good skill to use all support tools (utilities) integrated with monitored objects and OS of MOs.	Administrators only run monitoring requirements in MCDS by click on menu.
Implementation method	Manual method, based on remote connection and tool is manually executed.	Automatic method, based on implementing of monitoring agents.
Monitoring scope	Discrete, objects, local	Local, global, large scale DS
Monitoring time	Depending on skill of the administrators and network infrastructure.	Depending on monitored network infrastructure.
Error detection	Manually	Automatic warning
Diagnosing, and evaluation	Manually, depending on the skill of the administrator, local.	Automatic, multi-level: objects, networks, domains,...

IV. CONCLUSION

The modeling has an important role in the development of efficient algorithms for the monitoring problems in DS. This paper proposes a modeling method for the basic architecture of objects in DS, the monitoring solution for hierarchical architecture of DS. With the proposed models, we develop the MCDS solution that supports administrators for monitoring information visually such as the DS topology, the operations and status information of objects in the system, etc. Based on the monitoring entities, we easily develop extensions for these monitoring entities to provide complete online architecture information that effectively support for administrators, as well as allow storing monitoring data into database for the synthesis, evaluation and analysis of historical monitoring data later. This information is actively useful for the appropriate management decisions and controlling actions the monitored system.

In order to effectively deploy the monitoring solution for the distributed systems, we continue investments to complete the solution and optimize for monitoring algorithms, the dynamic management model and effective communication model for monitoring entities, as well as the analyzing techniques that optimize the computations for the large number of monitoring information in the large-scale systems.

REFERENCES

- [1] Christos G. Cassandras, Stéphane Lafortune, "Introduction to Discrete Event Systems", 2nd edition, Springer, 2008.
- [2] Gabriel A. Wainer, Pieter J. Mosterman, "Modeling and simulation theory and applications", CRC Press, 2011.
- [3] Gerard J. Holzmann, "Design and validation of computer protocols", Prentice Hall, 1991.

- [4] Phuc Tran Nguyen Hong, Son Le Van, "An online monitoring solution for complex distributed systems based on hierarchical monitoring agents", The Fifth International Conference on Knowledge and Systems Engineering, pp 191-202, 2013.
- [5] Son Le Van, Phuc Tran Nguyen Hong, "Researching on an online monitoring model for large-scale distributed systems", Proceedings of the 13th National Conference in Information and Communication Technology, Hungyen, Vietnam, 2010.
- [6] Weilong Hu, Hessam S. Sarjoughian, "A co-design modeling approach for computer network systems", Proceedings of the 2007 Winter Simulation Conference, 2007.
- [7] Yannick Pencolé, marie-odile cordier, Laurence Rozé, "A decentralized model-based diagnostic tool for complex systems", International Journal on Artificial Intelligence Tools (IJAIT), 2002.
- [8] Phuc Tran Nguyen Hong, Son Le Van, Huy Nguyen Xuan, "The technical overview report on some monitoring solutions for distributed systems", The technical survey report, Danang University of Technology, The University of Danang, Vietnam, 2014.
- [9] George Coulouris, Jean Dollimore, Tim Kindberg and Gordon Blair, "Distributed systems concepts and design", 5th Edition, Addison Wesley Press, 2011.
- [10] <http://corbatrace.sourceforge.net>
- [11] <http://monalisa.caltech.edu/monalisa.htm>
- [12] <https://www.ietf.org/rfc/rfc792.txt>
- [13] Jeffrey Joyce , Greg Lomow, Konrad Slind, Brian Unger, "Monitoring Distributed Systems", ACM Transactions on Computer Systems, 5(2), pp. 121-150, 1987.
- [14] R.Hofmann, "The Distributed Hardware Monitor ZM4 and its Interface to MEMSY", Universit'at Erlangen, IMMD VII, 1993.
- [15] Sheng-Yuan Yang, Yi-Yen Chang, "An active and intelligent network management system with ontology-based and multi-agent techniques", Expert Systems with Applications,38(8), 2011.
- [16] Phuc Tran Nguyen Hong, Son Le Van, "Monitoring of large-scale distributed systems based on SNMP development", The Journal of Science and Technology, Danang University, 1(8),79-84, 2012.
- [17] Xavier Logean, "Run-time Monitoring and On-line Testing of Middleware Based Communication Services", PhD dissertation, Swiss Federal, 2000.
- [18] Kwang-Hui Lee, "A Distributed Network Management System", Global Telecommunications Conference, IEEE,1994.
- [19] Aman Mahajan, Haresh Joshi , Sahil Khajuria , Anil k Verma, "ICMP, SNMP: Collaborative Approach to Network Discovery and Monitoring", International Journal of Smart Sensors and Ad Hoc Networks (IJSSAN) ISSN No. 2248-9738 Volume-1, Issue-3, 2012.

Ipv6 Change Threats Behavior

Firas Najjar

National Advanced IPv6 Center (Nav6)
Universiti Sains Malaysia
Penang, Malaysia

Homam El-Taj

Computer Science
Tabuk Univesity
Tabuk, Saudi Arabia

Abstract—IPv4 address pool is already exhausted; therefore, the change to use IPv6 is eventually necessary to give us a massive address pool. Although IPv6 was built with security in mind, extensive research must be done before deploying IPv6 to ensure the protection of security and privacy. This paper firstly presents the differences between the old and new IP versions (IPv4 and IPv6), and how these differences will affect the attacks, then the paper will show how the attacks on IPv4 and IPv6 will remain mostly the same; furthermore, the use of IPv6 will give rise to new types of attacks and change other types' behavior.

Keywords—Computer Attacks; IPv4; IPv6; Security

I. INTRODUCTION

Internet Protocol (IP) is a set of technical rules that define how computers communicate through networks [1], IP address is just like a home address or telephone number. In computer network; all devices in the same network must have a unique IP address to exchange data between them, without a well configured IP address, the communication with other devices in the network will be broken.

Nowadays most commercial and governmental information systems are connected through the Internet, using new technology like IPv6 at the time being might seem risky because it isn't be fully tested, which make it possible to attack. These systems must be protected from unauthorized access that may expose critical information, this can be done by detecting any suspicious anomalies in the network traffic patterns due to Distributed Denial of Service (DDoS) attacks, worm propagation [2] [3], viruses, Trojans and other kinds of malicious programs that introduce more panic into network society. Based on these attack types, securing such networks infrastructure has become a priority for most researchers.

The first IP address system widely deployed is Internet Protocol Version 4 (IPv4); IPv4 has proven to be robust, easily implemented, and interoperable. It has stood up to the test of scaling an internetwork to a global utility, the size of today's Internet, this is a tribute to its initial design[1], but the huge growth of using internet leads to the exhaustion of the IPv4 address pool [4], as a result, public IPv4 addresses have become relatively scarce, forcing many users and some organizations to use a Network Address Translation (NAT) [5]; to map a small number of public IPv4 addresses to multiple private IPv4 addresses. Although NATs promote the reuse of the private address space, they violate the fundamental design principle of the original Internet that all nodes have a unique, globally reachable address; additionally the growth of using the internet insures the reduction of IPv4 public addresses.

In 2011, Internet Assigned Number Authority (IANA), which is the main authority for IP address allocation announced the exhaustion of its free pool of IPv4 addresses [6], in addition, on 14th of September 2012 the Europeans Network Coordination Centre (RIPE NCC) which is responsible of addresses in Europe and in the middle east began to allocate IPv4 address space from the last /8 address pool of IPv4 address space it holds. Table I and Figure 1 show the exhaustion dates of IPv4 pool addresses.

TABLE I. PROJECTED RIR ADDRESS POOL EXHAUSTION DATES [6]

RIR	Exhaustion Date	RIP Pool/8
APNIC	19-Apr-11	0.8180
RIPE NCC	14-Sep-2012	0.8535
LACNIC	19-Jan-2015	1.4427
ARIN	12-Feb-2015	1.5558
AFRINIC	24-May-2022	3.4479

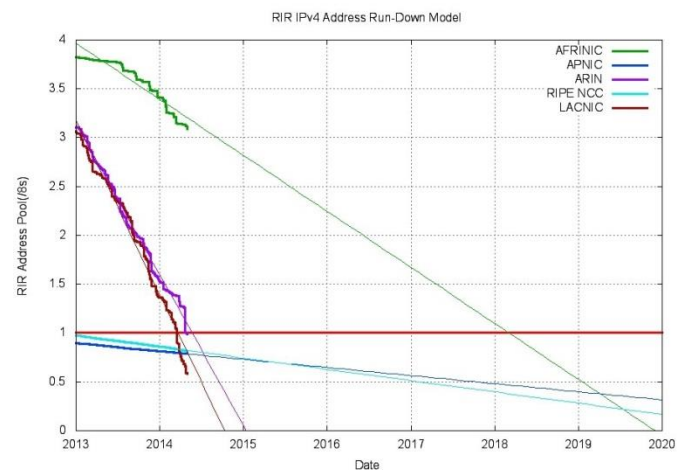


Fig. 1. Projection of consumption of Remaining RIR Address Pools [6]

Internet Protocol Version 6 (IPv6) [7] was deployed to overcome IPv4 address exhaustion limitation. IPv6 intended to replace IPv4 that still carries the vast majority of Internet traffic 2013. In December 2013, the percentage of users reaching Google services over IPv6 surpassed 2.7% [8], for that we must prepare ourselves to the next generation of addressing system IPv6.

The rest of this paper will be organized as following: Section 2 will cover an overview on network system to produce basic knowledge about network concepts, Section 3 shows how the differences between IPv4 and IPv6 will affect

the security of networks, furthermore how these differences affect the types of the attacks, and does IPv6 reduce the attacks?

II. NETWORK OVERVIEW

Networks are simply two or more computers connected to each other through medium to exchange data between them. In order to exchange data there must be some protocol or model that organizes the transmission between computers, for that International Organization for Standard (ISO) produced a conceptual model that characterizes and standardizes the internal functions of a communication system by partitioning it into abstraction layers called Open Systems Interconnection (OSI) model [9].

Each layer of OSI model serves the layer above it, and served by the layer below it, Table II shows the OSI model layers with main function and example protocol from real world.

TABLE II. OSI MODEL LAYERS WITH MAIN FUNCTIONS AND PROTOCOLS

NO.	Layer Name	Function	Protocols
7	Application	Provide service protocol to applications	FTP, HTTP
6	Presentation	Data representation, encryption and decryption	SSL,TLS
5	Session	Control Conversations/sessions between application	PPTP,RTP
4	Transport	Reliable delivery of packets between points on a network	TCP, UDP
3	Network	End to end Delivery	IP, ICMP
2	Data Link	Reliable direct point-to-point data connection.	PPP
1	Physical	Media Interface Transmission Method	

A. Internet Protocol Suite

Internet protocol suite is a suite of protocols, which were first designed for the Defence Advanced Research Project Agency (DARPA) network, which was called the (ARPAnet) during the early 1970s [10].

In the early 1980s, it was included as an integral part of Berkeley's UNIX version 4.2. Today, it is the protocol used by ARPAnet, MILnet and many other networks. The Internet Protocol suite is also commonly called TCP/IP protocol suite, because the most two important protocols in it: the transmission control protocol (TCP) and the Internet protocol (IP), these were also the first two protocols in the suite to be developed. If we compare Internet Protocol suite with OSI model, Internet Protocol suite contains four layers:

a) *The Internet application layer includes OSI Model application layer, presentation layer, and most of the session layer.*

b) *Transport Layer includes the graceful close function of the OSI session layer as well as the OSI transport layer.*

c) *Internet layer is a subset of the OSI network layer.*

d) *Link layer includes the OSI data link and physical layers, as well as parts of OSI's network layer.*

1) IPv4

IPv4 [1] is the fourth version of the Internet Protocol (IP) used to address the devices on the network to identify them, Internet Protocol is one of the major protocols in Internet Protocols suite, this protocol works at Network layer of OSI model and at Internet layer of Internet Protocol model. IPv4 is the first version of internet protocol widely used [11], IPv4 packet header consists of 14 fields, of which 13 are required, and the 14th field is optional.

IP protocol is responsible for the identification of hosts based upon their logical addresses and to route data between them over the underlying network, additionally IP provides uniquely identification mechanism to host by IP addressing scheme. IP does not guarantee the delivery of packets to destined host, but it will do its best to reach the destination.

IPv4 uses 32-bit addresses, which limits the address space to 4294967296 addresses, and because the exhaustion of these addresses, Internet Engineering Task Force developed, a new version called Internet Protocol Version 6 (IPv6), that uses 128-bit addresses, which is a very huge number of addresses.

2) IPv6

IPv6 is the latest version of the Internet Protocol (IP). IPv6 developed by the Internet Engineering Task Force (IETF) to overcome IPv4 address exhaustion. IPv6 is an Internet Layer protocol for packet-switched internetworking and provides end-to-end datagram transmission across multiple IP networks. Compare IPv6 to IPv4, IPv6 uses simplified header format in seven fields instead of 13 fields in IPv4, with fixed length header of 40 bytes only even that the IPv6 header contains two 128 bit addresses (source and destination IP address).

Figure 2. shows the differences between header formats for both protocols. IPv6 packet header contains fields that facilitate the support for true Quality of Service (QoS) for both differentiated and integrated services, to provide better support for real-time traffic like Voice over IP. IPv6 also includes labeled flows in its specifications to recognize the end-to-end packet flow through routers [12]. Due to the large address space, IPv6 uses stateless address auto configuration to auto configure addresses to hosts. IPv6 is not that different from IPv4, they use the same routing protocol, layer 4 unchanged, and Layer 2 also remain unchanged.

To summarize the changes between IPv4 and IPv6, there only three major changes:

- Fixed Header Length.
- Larger IP Address space.
- Address Resolution Protocol (ARP)[13]replaced with Neighbor Discovery Protocol (ND)[14].

The following list summarizes the features of the IPv6 protocol:

- New header format.
- Large address space.
- Stateless and stateful address configuration.

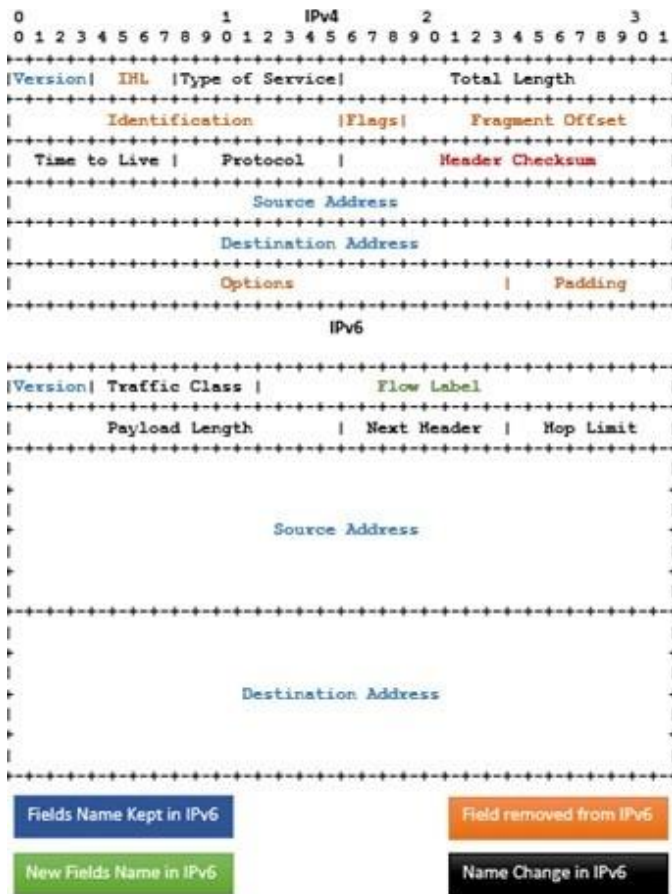


Fig. 2. Comparing IPv4 and IPv6

Internet Engineering Task Force (IETF) standards for IPv6 protocol stack functionality, includes the following:

- IP security (IPsec)[15] header support required. Better support for prioritized delivery.
- New protocol for neighboring node interaction.
- The IPv6 header [16].
- Unicast, multicast, and anycast addressing [17].
- The Internet Control Message Protocol for IPv6 (ICMPv6) [18].
- Neighbor Discovery Protocol (NDP) [19].
- Multicast Listener Discovery (MLD) [20] and MLD version 2 (MLD v2) [21].
- Stateless address auto-configuration [22].

Until IPv6 completely supplants IPv4, many mechanisms produced to make communication between IPv4 and IPv6 networks, by translating complete headers between IPv4 and IPv6 headers or by tunneling IPv4 packets in IPv6 packets [45]. These mechanisms are beyond the scope of this paper.

III. IPV4 AND IPV6 DIFFERENCES AND ATTACKS

IPv4 and IPv6 differences change the types of attacks; IPv6 substantially changes how IP interacts with the link layer, in

particular host. NDP will replace ARP, which is ICMPv6 based, and the use of protocols such as Secure Neighbor Discovery (SEND) [23] is a must to secure NDP or we will fall prey to the same class of attacks we faced in IPv4 over networks[44].

This Section outlines the common known attacks against IPv4 and then compares how these attacks might affect an IPv6 network, new types of attacks will rise and other will change their technique.

A. Reconnaissance

Reconnaissance attacks used to gather information as much as possible about the victim network when the adversary has no specific target. These attacks include port scanning and IP scanning using methods to establish a range of IP addresses which map to live hosts called PING SWAP tools.

The adversary uses PING SWEEP (also known as an ICMP sweep) to determine which of a range of IP addresses map to live hosts like computers or servers, whereas a single PING will tell you whether one specified host exists on the network or not.

PING SWEEP tools consists of Internet Control Message Protocol (ICMP) ECHO requests sent to multiple hosts, if a given address were live, it would return an ICMP ECHO reply. Ping sweeps are among the older and slower methods used to scan a network. After identifying reachable hosts, the adversary can systematically probe these hosts on any number of Layer 4 ports scanning to find services both active and reachable, by discovering hosts with active services, the adversary can then move to the next phase of attacks, this is why these attacks called passive attacks.

1) IPV4 Reconnaissance Attack

In IPv4, it is feasible to scan host address space of a specific network. If we have network address space of 16 bits (class B network) which represents 65536 hosts, the adversary can scan the whole network within less than two hours if the scan uses 10 addresses per second. This makes scanning usable mean for reconnaissance in IPv4 networks.

2) IPV6 Reconnaissance Attack

In IPv6, the situation is more complicated, the usual subnet size is 64 bits and with the same speed of scanning IPv4 subnet, it would take 60 billion years to scan all addresses, this makes scanning techniques impossible unless an adversary uses different approaches. As T. Chown [24] mentioned, some techniques will reduce the subset size, as if the adversary knows the Ethernet vendor prefix, the search space will reduce to 48 bit, and furthermore, if the adversary knows the Ethernet vendor, the search space may be reduced to 24 bits. Network Mapper (NMAP)[25] which is a tool that can perform all these scan types at the same time, produces new techniques to find all the hosts who use IPv6 on a target network:

- Targets-ipv6-multicast-echo sends an ICMPv6 echo request packet to the all-nodes link-local multicast address (ff02::1), collect the IPv6 addresses that come from and mark those hosts as potential scan targets

- *Targets-ipv6-multicast-invalid-dst* sends an ICMPv6 packet with an invalid extension header to the all-nodes link-local multicast address. Any hosts replying with an ICMPv6 parameter problem packet can be marked as up and available for potential scanning.
- *Targets-ipv6-multicast-mlt* attempts to discover available IPv6 hosts on the LAN by sending an MLD (multicast listener discovery) query to the link-local multicast address (ff02::1) and listening to any responses.
- *Targets-ipv6-multicast-slaac* sends an ICMPv6 router acknowledgment packet with a random address prefix, causing hosts to begin stateless address auto-configuration (SLAAC) and send a solicitation for their newly configured address.

These new techniques will help the adversary to identify a reachable host in victim's network to make the next step without spending much time like brute force scan, after identifying reachable systems; the adversary tries to find active ports and services that used for its next step of the attack.

B. ARP and DHCP Attacks

ARP Spoofing is a type of attack in which adversary tries to link a legitimate Media Access Control (MAC) host to adversary IP address. Once the adversary MAC address is connected to an authentic IP address, the adversary will begin receiving any data that is intended for that IP address.

Furthermore, the adversary can simulate network servers like Dynamic Host Configuration Protocol (DHCP) server, with this action the adversary will be able to reply to DHCP request before the real DHCP server; because it is closer to the client host. It will configure the Client host with IP address of that subnet, but it will also give a false Default Gateway address to host and maybe even false DNS server address.

1) IPV4 ARP AND DHCP ATTACKS

ARP spoofing can enable adversary parties to intercept, modify, or even stop data in-transit. ARP spoofing attacks can only occur on local area networks that utilize the Address Resolution Protocol. Cisco implemented a new technique called snooping [26] to overcome DHCP identity thief, by allowing certain ports to send DHCP server messages.

2) IPV6 ARP AND DHCP ATTACKS

The situation significantly changes in IPv6; ARP protocol replaced by Neighbor Discovery Protocol (ND), similar attack is still possible through Neighbor Solicitation/Advertisement Spoofing [27]. To verify sender ownership of claimed IP address, SEcure Neighbor Discovery (SEND) is used, which is a security mechanism used to secure ND from attacks, based on Cryptographically Generated Addresses (CGA) [28] and asymmetric cryptography. SEND uses cryptographically generated addresses to verify the sender's ownership of a claimed address. CGAs are IPv6 addresses in which part of the address is generated by applying a cryptographic one-way hash function based on a nodes public key and auxiliary parameters. The hash value can then be used to verify the binding between the public key and a nodes address. By default, a SEND-enabled node should use only CGAs for its own addresses. The

basic purpose of CGAs is to prevent the stealing or spoofing of existing IPv6 addresses. While SEND is a robust mechanism for verifying sender ownership, it is difficult to implement because it's based on Public Key Infrastructure (PKI), and most popular hot operating systems do not support SEND [29] [30].

C. Smurf attack

Smurf attacks were one of the first network-based denial-of-service attacks. The name Smurf came from the name of the source code (Smurf.c). The Computer Emergency Response Team (CERT) first issued Smurf attacks in January 1998.

1) IPV4 Smurf attack

In Smurf attacks, the adversary sends an echo-request message (ping) with a destination address of a subnet broadcast and a spoofed source address using the host IP address of the victim; this causes all the devices on the subnet to respond to the spoofed source IP address and flood the victim with echo-reply messages.

A ping allows remote systems to quickly determine whether another system is live on the network. If system X wants to "ping" system Y, it sends an ICMP echo request packet with a source address of X and a destination address of Y. When Y receives the echo request, it reads the source address (in this case, X) and sends an ICMP echo reply message back to the originating host. These replies quickly add up and, when repeated, can overwhelm the victim system, causing a denial of service.

Many Broadcast Amplification attacks are easy to disable by simply disabling directed broadcast forwarding [31].

2) IPV6 Smurf attack

In IPv6 the concept of an IP broadcast is removed, there is no implementation of traditional IP broadcasting in IPv6; there are only multicast, unicast and any-cast.

To mitigate these attacks in IPv6; A.Conta and S.Deering [32] states that: an ICMPv6 message should not be generated as a response to a packet with an IPv6 multicast destination address, a link-layer multicast address, or a link-layer broadcast address. On the other hand, even nodes are compliant to RFC 2463, the smurf attack can use the generated "Parameter problem ICMPv6 message" error messages in response to a packet destined to a multicast group [33], and it may use the packets, which were used in multicast video stream, because multicast video stream required allowing path maximum transmission unit (MTU) discovery. E. Vyncke, S. Hogg [33] stated: this opens the door to an amplification attack in the same shot. In addition, to mitigate this problem they advise to apply rate limiting to those ICMP messages: They should be rare in every network so that a rate limit (10 messages/sec) can permit the correct use of those messages (path MTU discovery) while blocking the amplification attack.

D. Flooding attack

Flooding is a type of Denial of Service (DoS) attack, which attempts to cause a failure in network communication by sending many requests to a network hosts, too many requests cause the attacked host to collapse.

Flooding attack is one of the most frequent attack types present in IPv4 networks, this type of attack can also affect the IPv6 networks by sending Router Advertisement packets and forcing operating systems to create IPv6 addresses in response to every packet it receives. By flooding the network with enough RAs, the host machines will consume more CPU time as the Stateless Auto Configuration process tries to configure the addresses [35].

E. Application Layer Attack

An application-layer attack targets application and operating systems causing a fault in applications and operating systems. This results in the adversary gaining the ability to bypass normal access controls and takes advantage of this situation to gain control of the application, operating system, or network. Some known types of these attacks are: buffer overflow, web application attacks, viruses and worms.

Most of these attacks are not affected by moving to use IPv6, because it is difficult if not impossible to recognize these attacks on Network layer, especially when using IPsec, because IPsec would make it impossible to read encrypted data. However, the advantage of IPsec implementation would make it easier to trace back to the adversary, because of mandatory authentication. Without IPsec, the source address can be spoofed.

The only change in Application-Layer attack is the propagation of worms. Traditionally worms make local and wild scanning to find victim hosts, which make it unlikely to succeed in IPv6 environment, but as we discussed earlier; taking advantage of local knowledge and patterns in address-space assignment, the attack program can cut the search space considerably.

There is a number of strategies worms could use in an IPv6-based Internet to find new targets:

- Routing Tables, many organization run routing protocol internally such routing protocol (RIPng) [36] worm would be able to consult the host routing table [37].
- Multicast, is a fundamental part for IPv6 which can be abused for target discovery by worm.
- Server Logs, servers must log incoming mail server, website, DNS server; these logs are valued information for the worms to spread out.
- Server Addresses, IPv6 addresses are very hard to remember, most administrators tend to select easily memorized IP, which can be exploited by worms.
- Search Engine, for worms that target Web server, search engine is the best source of information; A.kamra [38] shows that DNS worm in IPv6 could spread as fast as an IPv4 address scanning worm.

F. Sniffer Attack

A sniffer attack is an application or device that can read,

monitor, and capture network data exchanges and read network packets. If the packets are not encrypted, a sniffer provides a full view of the data inside the packet. Even encapsulated (tunneled) packets can be broken open and read unless they are encrypted and the attacker does not have access to the key.

IPv6 provides fundamental technology preventing sniffing attacks with IPsec and Internet Key Exchange Protocol Version 2 (IKEv2) [39].

G. Rogue Devices

Rogue device is an unauthorized node on the network; rogue device can be a router, switch, or simply a laptop, which acts as DHCP or any server type. When a client enters the network, both legal and rogue servers will offer services for the client. For example, DHCP servers will offer IP addresses, default gateways and other services, if the client accepts services from rogue DHCP, it may lead to sniff all client data or the client cannot access the network resources which lead to denial of services.

Rogue DHCP servers can be toppled by means of intrusion detection systems [40] with appropriate signatures, as well as by some multilayer switches, which can be configured to drop the packets. In addition, we can use 802.1X as a way of preventing entry and IPsec as a way of preventing access; it becomes evident that in order to attempt to solve the rogue machine problems in different ways we have to analyze our threats, consider our risk stance, and choose the appropriate way to protect our system[41].

IV. CONCLUSION

IPv6 is the future for sure; the main reason for migrating to use IPv6 is the exhaustion of IPv4 address pool, not any security issues. The security concerns between IPv4 and IPv6 are largely the same, packet transporting techniques are almost unchanged, and the upper-layer protocols: the application layer and transport layer are not affected, therefore, most of the attacks on IPv4 can be applied on IPv6, the concept of the attacks remain the same, but types and attacks' behavior are changed.

IPsec is mandatory in IPv6, which make it more secure, on the other hand, network administrator will be blind, because all the data are encrypted, and network administrators cannot apply network policies between any two IPv6 nodes.

Many organizations got IPv6 running on their networks and they do not even realize it; because many computer operating systems by default enable both IPv4 and IPv6, which could cause security vulnerabilities if one of them is less secure than the other. IPv6 security vulnerabilities currently exist, as the popularity of the IPv6 protocol increases, the number of threats increases too, Table III. proves that most tools used in IPv4 attacks have new versions that work on IPv6, which mean; IPv6 didn't eliminate the attacks, it just change the behavior and techniques for the attacks.

TABLE III. IPV6 ATTACK TOOLS [42][43]

Attack	IPv6 Attack Tool
Reconnaissance	NMAP6, Dnsdict6, Alive6, Thcping6
Flooding	6tunnel6, Flood-router6, Flood-advertize6
Smurf	Smurf6, rsmurf6
Rogue Device	Fak-router6
Man In The Middle	Redir6, Parasite6, Toobig6

REFERENCES

- [1] J Postel, Internet Protocol, DARPA Internet Program Protocol Specification (September 1981), RFC 791.
- [2] Christos Douligeris, Aikaterini Mitrokotsa, DDoS attacks and defence mechanisms : classification and state-of-the-art ,Computer Networks: The International Journal of Computer and Telecommunications Networking, Vol. 44, Issue 5 , pp: 643 - 666, 2004.
- [3] Z. Chen, L. Gao, K. Kwiat, Modeling the spread of active worms, Twentyv, Second Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Vol. 3, pp. 1890 1900, 2003.
- [4] R. L. Mitchell. The grill: John Curran. Computer-World, Apr. 2010.
- [5] K. Egevang, P. Francis, The IP Network Address Translator (NAT), RFC 1631, May 1994.
- [6] G. Huston, GeoHuston <http://www.potaroo.net/tools/ipv4/index.html>, DEC. 2013.
- [7] S Deering, R Hinden, Internet Protocol, Version 6 (IPv6) Specification, RFC 2460, December 1998.
- [8] Google Statistics, www.google.com/intl/en/ipv6/statistic.html, 2012.
- [9] International Organization for standard ISO/IEC 7498-1
- [10] V. Cerf, The Internet Activities Board, RFC 1160, May 1990.
- [11] BGP Analysis Reports Retrieved, Jan 2014.
- [12] J. Rajahalme, A. Conta, B. Carpenter, S. Deering, IPv6 Flow Label Specification, RFC 3697 March 2004.
- [13] David C. Plummer, An Ethernet Address Resolution Protocol, RFC 826, NOV 1982.
- [14] P. Nikander, J. Kempf, E. Nordmark, IPv6 Neighbor Discovery (ND) Trust Models and Threats, RFC 3756, May 2004.
- [15] S. Kent, K. Seo, Security Architecture for the Internet Protocol, RFC 4301, December 2005.
- [16] S. Deering, R. Hinden, Internet Protocol, Version 6 (IPv6) Specification, RFC 2460, December 1998.
- [17] S. Deering, R. Hinden, IP Version 6 Addressing Architecture, Feb 2006, RFC 4291.
- [18] A. Conta, S. Deering, M. Gupta, Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification, RFC 4443, March 2006.
- [19] T. Narten, E. Nordmark, W. Simpson, H. Sliman, Neighbor Discovery for IP version 6 (IPv6), RFC 4861, September 2007.
- [20] S. Deering, W. Fenner, B. Haberman, Multicast Listener Discovery (MLD) for IPv6, RFC 2710, October 1999.
- [21] R. Vida, L. Costa, Multicast Listener Discovery Version 2 (MLDv2) for IPv6, RFC 3810, June 2004.
- [22] S. Thomson, T. Narten, T. Jinmei, IPv6 Stateless Address Autoconfiguration, RFC 4862, September 2007.
- [23] J. Arkko, J. Kempf, B. Zill, P. Nikander, Secure Neighbor Discovery (SEND), RFC 3971, March 2005.
- [24] T. Chown, IPv6 Implications for Network Scanning, RFC 5157, March 2008.
- [25] NMAP.Org, NMAP IPv6 Tool, Retrieved 2013.
- [26] Cisco, Understanding and configuration DHCP snooping, December 2012.
- [27] P. Nikander, J. Kempf, E. Nordmark, IPv6 Neighbor Discovery (ND) Trust Models and Threats, RFC 3756, May 2004.
- [28] CISCO, IPv6 Brief, White Paper, Oct 2011.
- [29] T. Chown, S. Venaas, Rogue IPv6 Router Advertisement Problem Statement, RFC 6104, Feb. 2011.
- [30] T. Aura, Cryptographically Generated Addresses, RFC 3972, March 2005.
- [31] E. Guttman, L. Leong, G. Malkin, Users Security Handbook, RFC 2504, February 1999.
- [32] A. Conta, S. Deering, Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification, RFC 2463, December 1999.
- [33] E. Vyncke, S. Hogg, IPv6 Internet Security for Network, Cisco Press, JUN 2009.
- [34] C. Kaufman, P. Hoffman, P. Eronen, Internet Key Exchange Protocol Version 2 (IKEv2), RFC 5996, SEPTEMBER 2010.
- [35] T. Chown, S. Venaas, Rogue IPv6 Router Advertisement Problem Statement, RFC 6104, Feb 2011.
- [36] G. Malkin, R. Minnear, RIPng for IPv6, RFC 2080, January 1997.
- [37] A. Kamra, H. Feng, V. Misra, A. Keromytis, The Effect of DNS Delays on Worm Propagation in an IPv6 Internet, IEEE INFOCOM, March 2005.
- [38] C. Zou, D. Towsley, W. Gong, S. Cai, Routing Worm: A Fast Selective Attack Worm Based on IP Address, Workshop on principles of Advance and Distributed Simulation, June 2005.
- [39] C. Kaufman, Y. Nir, P. Eronen, Internet Key Exchange Protocol Version 2 (IKEv2), RFC 5996, Sep. 2010.
- [40] H. Eltaj, F. Najjar, H. Alsenawi, M. Najjar, Intrusion Detection and Prevention Response based on Signature-Based and Anomaly-Based: Investigation Study, International Journal of Computer Science and Information Security, June 2013 .
- [41] I. Halil, Detecting and Preventing Rogue Devices on the Network, SANS Institute, Aug 2007.
- [42] NMAP.org, Network Mapper, May 2012.
- [43] Thc.org, thc-ipv6, Dec 2013.
- [44] Supriyanto, Iznan Husainy Hasbullah, Raja Kumar Murugesan and Sureswaran Ramadass, "Survey of Internet Protocol Version 6 Link Local Communication Security Vulnerability and Mitigation Methods", Vol 30, no 1, pp 64-71, Jan-Feb 2013.
- [45] Ala Hamarshah, "Assuring Interoperability between Heterogeneous (IPv4/IPv6) Networks without using Protocol Translation", Vol 29, no 2, pp. 114-32, Mar-Apr 2012.

Emotional Engagement and Active Learning in a Marketing Simulation: A Review and Exploratory Study

Kear Andrew, Bown Gerald Robin
The Business School
University of Gloucestershire, Gloucestershire, UK

Abstract—This paper considers the role of emotional engagement during the use of a simulation. This is placed in the context of learning about marketing. The literature highlights questions of engagement and interactivity that are entailed in the use of these simulations. It is observed here that both the anticipation of and the process of engagement with the simulation generate emotional responses. The evidence of emotional anticipation was collected through the use of vignettes and a short survey. The production of negative emotions before and after the activity was observed and considered. The particular occurrence of these emotions on the development of understanding is then discussed. There is general evidence for the mundane reality of such simulations that support learning and group engagement. The connection with activity theory was explored and proposed as a potential theoretical fit with the evidence.

Keywords—component; Learning; Simulations; Feedback; Emotional Learning Scenarios; Emotional Anticipation; Deep Learning; Vignette Research

I. LEARNING AND ENGAGEMENT

The authors have been involved in the delivery of a marketing simulation exercise to undergraduate students at a British university. This course forms one module of the undergraduate programme, often delivered in the second or third year. It is traditionally considered that the value of these simulations resides in the ability to give a different learning experience to the participants [32]. The same authors consider that the student's characteristics should be taken into account. Importantly for this paper these characteristics are considered to be multivariate. The perspective of this paper is to examine this assumption and to increase the understanding of the processes involved in using simulations in a learning environment. The proposition examined in this paper is that the achievement of active learning can be dependent on the mundane nature of situated understanding [20]. The emotional placement prior to the simulation will also be explored. The appearance of negative emotional responses and the connection with the generation of increased attentiveness will be further explored and elaborated in this paper.

In the past the structure of the learning environment was often taken for granted. During the early 1990's studies revealed that students found the traditional, didactic lecture to be their least favourite [4]. Students were found to be more willing and enthusiastic about taking a more active role in the lecture according to Ref. [36]. The teaching and learning strategies

adopted here are meant to assist the student in becoming a 'critical being' [3]. In this view the student attains the ability to determine the critical reason for a particular course of action or method, achieves a critical level of self-reflection and review of contextual history of the self, and finally becomes able to undertake a critical action that is based upon problem solving at the skills level. As a result of this problem based learning students were required to take a deep approach to learning [24]. As these ideas developed the idea of interactive lectures was found to be more enjoyable than traditional lectures [9]. Whilst the didactic lecture is still with us the development of technology has an impact in improving engagement, both within the lecture format [16], and also in the production of deeper learning [33]. Ref. [19] suggests that experiential simulations and class room games are an effective way to increase the attentiveness of students. This increased awareness and elaboration of their own unrehearsed behaviour may cause them to rethink their engagement through the importance of tactical decisions and overall strategies employed. This leads to the generation of further ideas about the idea of engagement which is now more prominent in these discussions than the idea of criticality mentioned above. Technology is something that we are required to interact with, but it has been argued as more than a system for delivery; there is still the need to build relationships for learning [11]. Ref. [12] suggests that the increase in student engagement can be attained through creativity and emotional engagement.

According to Ref. [19] the effectiveness of simulations is primarily achieved in engaging the whole student and not just their intellect and analytic powers. This emotional engagement is often considered in terms of psychological approaches to emotional and cognitive learning. Previously there had been some reluctance within higher education to include the engagement of emotions into the learning space where it had been described as 'inappropriate territory' [22]. This coincided with a challenge to the primacy of rationality in learning, thus preventing what some see as the development of the transcendental nature of rational understanding in a life world [2]. This previously 'inappropriate territory' has more recently become an important dimension of learning that can significantly enhance students' engagement with learning and assessment [27]. It may be the case that the preference for such learning is driven by novelty but it can be noted that a significant amount of research points to the value of an emotional content in the development of learning.

Starting from the tenet that we need to improve and refine the overall learning experience of students this creates an obligation to consider further some fundamental approaches to learning. One of these is the concept of interactivity. It is generally considered that improving learning can be achieved through the promotion of either interactivity [9] or transformative learning [26]. The idea of interactivity often evades a definition but is commonly used in the discussion of computer mediated environments. Ref. [38] in discussing interactivity, consider the device- and message-centric approaches to this idea before deriving from the literature a four-factor taxonomy of bi-directionality, timeliness, mutual control and responsiveness. Interactivity, in their formulation is generated through computer mediated *communication* (CMC) thus indicating the assumptions of this approach. This communicative framing of the concept of interactivity in this context needs to be developed in tandem with a learning perspective. In considering transformative learning [35] has previously noted that the empirical basis for transformative theory requires some further investigation into the contexts where it is fostered. The approach of transformative learning is operationalized as 'an epistemology of evidential and dialogic reasoning' [26]. The fit with interactivity is made in that this epistemology emphasises the dialogic/communicative aspects of learning. According to Ref. [31], in their findings offer support that tools for transformational learning are providing students with learning experiences that are 'direct, personally engaging and stimulate reflection upon experience' [36]. The authors here suggest that there may be some transformative currency in the use of a simulation but it is considered that the process of transformational learning requires some mediation for it to develop. This mediation can be the inspirational lecture, or through peer participation, but often some form of artefact is needed to enable this learning [15].

II. LEARNING AND SIMULATIONS

The artefact in this case is the computer simulation which often presents itself as a constructed object as the focus of the activity. There have been some studies on learning with simulations. Research by Ref. [18] found that students did in fact learn by participating in a simulation and the simulation was both enjoyable and perceived to be worthwhile. According to [21] students perceive simulations as being (1) engaging, (2) useful, (3) effective learning tools, and (4) effective in promoting teamwork. Students' perceptions of computer-based simulation team dynamics and their positions on the use of simulations and simulation performance was researched by [1] who found that student team cohesion and student team independence strongly influenced their perceptions of the use of computer based simulations. The other aspect that is worthy of note here is the role of the instructor or facilitator, who provides interpretation for the groups when requested. There is available an instructor in the simulation forum who can be considered to be both a narrator and also a guide. The latter role becomes prominent in the respect that they have travelled this way before. The specific role of instructors is likely to differ among simulation types but could be considered as providing feedback and guidance on the simulation and how to use it. In relation to the particular marketing simulation here, Markstrat, the role is to draw student's attention to the outcomes of their decisions and

subsequent performance relative to their competitors; often in a narrative form. So the dialogic learning here occurs in two directions, the intra group dialogue and the dialogue between the group and the facilitator. Research into the use of other marketing simulations such as Capism [6] offer an opportunity to operate virtual companies making decisions concerning marketing, production, finance, human resources, TQM and ethics. Here some students were taught how to use the simulation and coached by a professional. This increased performance when compared with a control group that were left to make decisions on their own. The findings indicated that the use of a business professional into the classroom improves the ability of students to make decisions. This seems to confirm the expectation that an effective instructor will improve learning.

It may be that the development of interactivity is now linked to the preferred mode of learning for a particular generation. According to Ref. [34] current undergraduate business students as members of generation Y have shorter attention spans and desire interaction and stimulation resulting in student engagement becoming more important. So in this respect a challenging environment provides students with an arena whereby they can thrive [34]. Generation Y are further characterised by having lived their entire lives with technology such as computers, mobile devices and video games. An increased level of "gaming", which is deemed to be characteristic of the activities expected by this generation, results in the need for educators to develop forms of engagement which are suitably commensurate with this mode of being.

Returning to the view of the simulation as an artefact based activity the ideas of activity theory consider 'that object orientated actions are always, implicitly or explicitly, characterised by ambiguity, surprise, interpretation, sense-making and potential for change' [8]. Central to this view is that 'an activity system is by definition a multi-voiced formation' [7]. In considering the role of artefacts in collaborative virtual environments (CVE's) [30] proposes that much can be learnt through the division of the learning environment into outlook, structure, and roles. The authors of this paper develop this idea in terms of an immersive computer environment which relies on the employment of artefacts for communication facilitation and task accomplishing. The simulation discussed here creates its own artefacts, such as screen presentations and reports, but in the occupation of a physical classroom. The participants are also collected into groups for the performance of the mediated task which affects the communicative and multi-voiced aspects of the CVE. In seeking to enhance this idea of active engagement in the development of learning this paper will develop the idea of interactive communicative environments (ICE).

The roles occupied by the participants of this research are to a significant extent driven by their approach to the task. Whilst Ref. [14] asserts that active learning can generate a positive emotional response that instigates an 'attentiveness that enriches understanding' and which can enhance self-esteem and a sense of empowerment [17]. There is, however, little research that investigates the effect of negative emotional response and its impact on attentiveness that enriches understanding. The attentiveness here is not considered to be in the transcendental domain of consciousness but to contain the elements of activity and engagement often developed by the incidence of ambiguity.

The structure of this simulation is the development of interactivity with the artefact. This interactivity can be an attribute that is more closely associated with the 'reality of business life'; however defined. Ref. [28] suggested that computer technology such as simulations can be used to create authentic assessments that mirror real world scenarios. Related to active learning is the generation of deep learning as a result of the simulated experience and linking a taught theory such as segmentation, targeting and positioning (STP) to the 'real' world resulting in the challenging of preconceptions [14]. According to Ref. [24] the type of test that is anticipated will determine the level of learning undertaken by the students. As such it is deemed beneficial to explore the thoughts of the participants before they interact with the simulation. It is worth noting at this point the nature of business activity, which is often characterised by periodic reporting of performance and the business simulation offering the same process of periodic reporting. This develops a rational and reflective processing of information, within a dialogic group dynamic as discussed earlier. Other programmes such as role playing games offer continuous incremental feedback.

It is suggested that active learning can create an enhanced affective response [14]. This is supported by the findings of Ref. [31]; one participant in their research stated that 'You can't develop a real understanding until you experience it personally.' According to Ref. [5] getting students committed, via small group exercises, and other active strategies, to owning the material. This, it is proposed, results in the students becoming more effective learners who are more likely to achieve the learning outcomes of the lecture. The learning is often considered to reside at the level of skills, which are often demanded by employers [10]. Therefore the structured approach to teaching and learning strategies is strongly based upon active learning and the implicated engagement with artefacts. The extent to which active learning generates an affective response and is moderated by the use of artefacts and dialogic reasoning will be considered in this piece.

III. BACKGROUND TO MARKSTRAT (SIMULATION)

The simulation used was the marketing simulation Markstrat. This is employed in a 2nd year undergraduate module in Marketing management. The simulation can be operated with 4, 5 or 6 companies operating in 2 product markets within consumer electronics. The students work in teams of between 4 and 6 team members and are assigned to control and make decisions for one of the companies for 6 rounds (the equivalent of 6 years). With the Markstrat simulation there are clear and visible outcomes to the decisions made via a brand map displaying the position of customer segments, the position of the company's brands and the key dimensions for improving positioning. The rounds last for approximately 1 hour and 30 mins and the feedback of the decision making of the teams is then analysed and reported back visually via a projector. This meant that all teams could compare their decision making with their peers who are running competing firms. Furthermore the students were prepared for the simulation by undertaking a number of case analyses that aimed to build the students' knowledge in relation to the type of decisions that they would be making. Whilst there are many uses for simulations in higher education Ref.[28] explored how simulations could be used as

assessment tools. They suggest based on their findings that simulations work by helping students to master knowledge and skills and that they work well with formative assessments. In the Markstrat simulation assessments are based upon the student's perceptions of the outcomes of their decision making and overall understanding of key concepts in relation to their performance. The same authors find that a conceptual framework and supporting learning materials are necessary to support student learning. The Markstrat simulation is underpinned by the marketing and branding conceptual framework on which the literature is vast. The Markstrat simulation is a simulation with a web-based interface and a brief summary of its operation is given here. The teachers / experts provided expertise on how to run the simulation and where to find and how to use typical marketing data to underpin their decisions. The students are in groups of 4 or 5 and have to operate a company and its brands in 2 markets with a total of 8 segments (5 segments in market 1 and 3 segments in market 2). All teams start in identical points regarding performance and positioning. As such the students have to make tactical decisions for each of their brands. The results of these decisions are presented periodically and include profits (overall and per brand), turnover, market shares, stock price index, inventory costs, awareness etc. The results are also communicated in a comparative form, group performance is judged to be better or worse than the other participants. The simulation had been previously found to enhance learning through active engagement and as such the authors wished to identify why and how the use of a simulation could improve the active learning within the business school.

The overall aim was to illuminate the decisions that have to be made to manage marketing in a competitive context. The aim and description of the simulation can be taken from the website:

"Markstrat offers MBA students and professionals a risk-free platform for testing theories and making decisions. From competitive forces to the effects of sales, distribution, R&D and advertising, every aspect is real ... The competition is real, but so is the teamwork."[23]

This description nicely sets up this research project as it identifies the 'reality' of the game and the elements of teamwork that are integral to its operation.

IV. RESEARCH METHODS

It was decided to approach this research by trying to discover the emotional placement of the students in respect of the simulation exercise. The idea was to explore the mundane nature of situated understanding [20] which is conceived to be a transcendence of the emotional and intellectual division. The anticipation of the exercise has also been noted as a significant factor. If this simulation incorporated the elements of activity theory discussed above the potential for it to create this ambiguity and surprise would be anticipated by the participants. The methodology was designed to produce an evoked awareness of the situation that the participants were about to enter. To explore this situation of created anticipation a student cohort were, in the first instance, given some scenario sketches [37] about imaginary predecessors on this particular course. In order to generate a response that could address a situated understanding, the students were given a sheet which contained 4 descriptions of students that might be facing this simulation

exercise. The descriptions were about 100 words long and presented the ideas of complication, speed and timeliness, group work, management of unpredictability, and a sufficing approach.

This methodology allowed students to adopt a description that suited their feelings and also allowed a sensitisation to the arising issues in the conduct of the simulation. The scenarios or vignettes, were also designed to be short thus enabling a quick completion of the task. The rapid response was designed to facilitate an emotional rather than an intellectualising response and such vignettes have been used in illustrative research [25].

The students were asked to see which description matched their own situation and comment on their own words. They were also asked after the simulation whether they still agreed with what they said before, and what advice they would give to the subsequent group. There were 91 students in the cohort, 61 gave a response and 30 agreed with one of the descriptions offered.

Description	Number	Percentage
Stephen	6	20%
Kirsty	1	3.33%
Jayne	16.5	55%
Kevin	6.5	21.67%
Total	30	

V. FINDINGS AND ANALYSIS

The written comments were now studied. More students had provided comments than had chosen a scenario (61 to 30). From an initial visual inspection of the words used in the comments it seemed that the responses could be grouped under six headings. These were; The ability to achieve results, the enjoyment of participation/interaction, the problem solving focus, the interesting/engaging nature of the task, the uncertain nature of the task, and the idea of task novelty. The initial textual analysis found that the first and the last were areas where the least comments were made. The few comments on the ability to achieve were 'we're going on to win', 'how I will achieve success' and 'interactions are vital to achieve success'. The comment about personal success in a group assignment is interesting and the comments about success have been incorporated in a revised analysis. It is possible to conclude that the elements of overt competition were not articulated to a great extent, and this might provide support for the elements of anticipation and ambiguity, so not quite knowing what to expect. The idea of the different nature of the simulation garnered few comments that were difficult to separate from other areas. Several comments about this task being different due to its practical nature seemed to say something more than merely contrasting the difference from the standard learning situation. There was one comment about the nature of this task being different because it required the group to brief the next group but there were few comments of this nature.

There were a number of comments about the uncertain nature of the task. These contained comments such as 'working outside the box' and 'thrown in at the deep end'. It seemed that the comments that indicated difficulties with control belonged here. The task was 'overwhelming' and 'unpredictable'; 'I do not know what's coming' 'not knowing what to do' 'what to expect' 'will I be able to cope'. The time was a factor in the espoused uncertainty 'not much time' 'complicated and speedy

decisions' were required. 'I am overwhelmed about the task ahead'. It is possible to see these comments in terms of anticipation.

Under the category of interesting several of the comments were of the form that this was 'an adventure' 'something new' 'looking forward to what it can teach me' 'it is active not reflective'; they saw the novelty of the exercise. The task to some seemed also to be created to conceal. It was 'unclear' 'there were a lot of what ifs' 'I didn't really understand'. It was also said that 'you can use the knowledge you have gained beforehand'. It was an opportunity therefore to use some of the instructional teaching in a 'practical' forum. This practical artefact-based activity with the simulation becomes explorative; the exploration of being effective in the situation given.

The problem solving approach seemed to focus on decision making in what was perceived to be a real situation. The 'real' nature of the task was mentioned by a significant number of the participants. So it was 'unpredictable – more realistic' 'theory and problem solving in situ' 'real business decisions – real life'. The judgmental nature of the task was identified; 'put in a position where you have to make a judgment call' 'the effect of one decision on the next' 'having to make decisions' 'simulation is a different concept of learning – an idea about decisions'. One of the participants talked about the role of luck in decisions but a later view, after the simulation, talked about 'an informed prediction' being a better description. One of the scenarios talked about man vs machine and this was picked up by one who felt that the machine had been programmed by man. So it was a constructed situation.

Those that expressed the view that this was enjoyment through interaction were all those who enjoyed 'making decisions in a group'. This exercise was seen as 'clever and exciting' presumably more so than the tasks they were currently faced with. This was the sort of activity that was 'open to different views', this needed a consensus to make it work. The collaborative nature of the task here was more important than the decision making.

The six initial interpretations which were discernable in this research seemed to reside in the same domain as those given for activity theory. There was evidence that the participants looked forward to the dialogic nature of learning. As most of these participants were from Generation Y who are considered to value independence [13] there was a significant level of apprehension about the exercise, which seemed to be in contrast to theory based exercises. Ref. [13] also explores the incidence of group think within this generation and how shared perspectives develop in these circumstances. Given that a number of these responses could contain emotional elements it was decided that a follow-up exercise would be useful in exploring the idea of emotional engagement further. There would be indicated studies of both expectation of, and response to, the task. This second investigation would explore the ideas of uncertainty and response to feedback, exemplified in the response to unexpected outcomes of the simulation. This uncertainty was likely to be highlighted where the initial results from the first round of the simulation were not as good as the participants would have hoped. The participants were given a

short questionnaire which was designed to elicit open comments to the questions given below:

Based on your performance how do you feel now?

Slightly better-- Yes slightly-- confused but start to understand the aim of the game-- ok, however unhappy we didn't win because we entered emergent markets only-- like some loser-- confused and negative as I have no idea what went wrong and how to improve-- No I feel the knowledge I had was obviously wrong-- confused, but start to understand the aims of the game-- I could have done better-- Not amazing, as the system failed to recognise our inputs at one point which meant our drastic actions that could have made us win came into action too late-- No because there is no way of applying it to a real company, it's hard to get a feel for anything when it's just numbers on a screen-- I felt fine after the first few rounds, but disappointed in round 3 results-- I feel much more confident in using simulations and how to change certain aspects to influence the final result-- Not that great but still confident that our stock price can increase-- Feel that our strategy was wrong and we should have perhaps taken a more aggressive approach--Disappointed after a strong start - demotivated with the continuous decline in stock price and market share - Annoyed as our market share has plummeted drastically, also demotivated - Not good we have decreases in market share massively - Disappointed - Disappointed, we started well but our strategy wasn't as aggressive as others - Tired hungry numb soulless - Disappointed and frustrated - Disappointed about our group's performance - I feel very confused - Very disappointed with the overall position of the company - Distracted that we haven't managed to turn the company around, still struggling to understand where it all went wrong.

What was your response to poor performance (if any) of your company?

Unhappy yet we had a very strong position in the market and should have come first-- review previous decisions, amend these and invest more, see where we went wrong and make changes-- mostly confusion as we felt our ideas would have benefited the organisation and the consumer. I felt that the other groups should have shared their views and decisions, It was very frustrating as our decisions seemed to have no effect despite us putting a lot of time and effort into our decisions-- Horrible I want to cry-- We need to relook at our strategy and think about changes we make more carefully. Perhaps look at data and analyse more. Do better-- My response was negative, as I was not aware, or did not have the ability to put it right or improve the performance in the next round - How did it occur how can we fix it - Reduce costs, look at competitors - Disappointment - Tried to cut costs down - Increase production - Panic/Recklessness[sic]/ find cause of problem and solve - Frustration - Proved the importance of market share positioning, doing well from a poor position - From stone cold bottom to 3rd in the last year - Focused on a few key brands and markets pouring more money into advertising and sales force for those - Pleased to have pulled off a third place finish.

Do you understand what has gone wrong?

Didn't really think about the numbers-- We did not use any budgets, targeting too many markets-- Yes apparently we did the

wrong thing-- we understood that we needed to improve sales and brand-- No ☹ -- Yes, communication breakdown as simulation got more complicated-- Not really-- After miscalculating the first period, I feel that this led to our downfall and that problems with the system that didn't allow us to access the R&D section until much later in the process-- We didn't evolve with the market quickly enough-- We were too conservative early on-- Yes we didn't do anything to improve our standing, but just in maintaining what we had. We accidentally withdrew a product-- Yes I invested a bit too much in a declining market which didn't work out--To a certain point- Wrong market segment changes in operations incurring unnecessary costs while not getting revenues -costs, more aggressive competition, targeting wrong segment - Our costs are too high - Totally rebranded when we didn't need to, people clicking without thinking, not looking at all the graphs - We targeted the wrong markets in the wrong rounds, also we sold out of units multiple times - Some parts yes - Yes not enough units sold - Although we may have made some bad decisions throughout the process I feel the group previously did not stand us in good stead - We didn't produce enough products and look at forecast - We didn't pay enough attention to the marketing forecasts provided - I believe the group we took over from are the main cause of our problems due to the wrong target markets. In addition I believe we didn't look into the forecasts in much detail causing the issues to multiply.

What do you need to improve?

Careful analysis and planning-- We have to do more on power-- We need to target the price and power, target fewer markets-- brand positioning, covering all the markets-- change decisions-- brand and sales-- I feel that I really do not understand the fundamentals of marketing. I therefore need to read and learn about marketing in general-- R&D in relation to target audience. Didn't fully invest in market when we had the opportunity to dominate that market-- not sure-- be more organised etc-- enter the new market-- We need to focus on one segment with our side product, rather than split it between two segments-- How to convert numbers and refer to numbers from the last period in order to make a good decision for the next period-- Enter a new market-- Think I need to look into market positioning more and what can affect it -Understand the market segments better - Define target segment accurately, run operations according to demands of the target segment - Reduce costs and increase sales to be able to increase ROI and market share, target another market - On spending according to our budget - Do not totally change target market when one brand is doing well - Pay more attention to the market forecast - Base future experiences on problems that were faced today - critical and strategic thinking - We need to think about our decisions more and not change our target market in the closing stages of the simulation - Don't try and fix something that isn't broken - Put more money into advertising, stayed with a successful product - More detailed discussions and look into the details of the task closer in order to make more strategic decisions.

Did your understanding improve?

Yes slightly (2 times) - Yes (17 times)- No, I feel the knowledge I thought I had was obviously wrong - it depends, in certain areas as some problems require lateral thinking while

other need to be thought about [in] more detail – No because there is no way of applying it to a real company. It's hard to get a feel for anything when it's just numbers on a screen – Yes because this gives me a bit more of an insight to real world marketing and the decisions taken in a business – It has a little, how quickly price product and market segments can change so quickly – Yes understanding of figures and relation between sales/price (positioning maps) – Not really, [I] think it will do with time but after period one has not had enough time to see changes – disappointed after a strong start – Yes a lot – I am starting to understand all the decisions businesses face – Definitely improved – To a certain extent.

VI. DISCUSSION AND CONCLUSIONS

The results of this investigation into participants' responses go some way towards elaborating the nature of student engagement in the simulation as a form of situated learning. It employed an exploratory and iterative methodology that enabled the initial questions to be refined and developed. As a result of the initial exploration of the learning situation it is noticeable that there is certainly an expressed anticipation about what might happen in the use of the simulation. The evidence of performance is more visible and immediate in the conduct of such a public exercise which may increase the uncertainty, risk of failure, and intergroup competitiveness. Often the results of learning in taught sessions are communicated and assessed sometime after the event. In this exercise there is this element of immediacy attached to the results which require a quick response. There is constructed a competitive simulation environment whereby no one wants to lose at least visibly, those that are losing may become even more attentive. This can be interpreted from the comments received in the questionnaire that highlighted the response to perceived poor performance. The replies to the question 'Based on your performance how do you feel now?' generated a number of negative views. These were expressed as disappointment, demotivation, being distraught, and confusion. One participant found that it was hard to get a feel for anything. These comments seem to fit in with the concepts of activity theory of ambiguity, surprise and interpretation introduced above. It is this participation in the time focused activities of the simulation that gives rise to its mundane nature. Being thus 'of the world' gives rise to the reality of such simulations.

In general there was a tendency to offer mainly analytical comments to the questions about understanding and the actions needed to improve. Particular comments made were "...be more organised; understand the market segments better; we need to think about our decisions". So it may be inferred that there is some move towards the development of expertise in response to this question and a consequent need to develop understanding further. So whilst the activity system in Ref. [7] used an expert activity system here there seemed to be the development of expertise although the emotional responses to the exercise were negative.

It can be noted that the phrasing of the questions tended to govern the nature of the response. 'What was your response' tended to generate personal comments, whereas 'do you understand' tended to encourage reflections on the group performance; but both of these were present. One said there was

'communication breakdown' but general reflection on dialogic communication was not overt. Perhaps because these were people who had undertaken other group activities before this simulation exercise. Perhaps the nature of the simulation as an artefact tended to discourage the group communicative aspects. It was the technology that drove the interaction as opposed to the members of the group. In contrast to the comments about understanding there was a large amount of negative emotion connected with the question about their response to poor performance. Confusing, frustration and disappointment were statements that were noticeable. One respondent wanted to cry. There seemed to be a difficulty in knowing what to do here – in this activity – which could have generated the negative emotions. So as such the operation and interaction with the simulation generated this aporia. As such the engagement of emotions 'negatively' in the above scenario was found to be a vehicle to utilise this important dimension of learning [27] and to generate attentiveness to enrich understanding [14]. As an artefact the simulation was noticeably generative of this situation which would not have occurred in a traditional lecture format. This quotation that said 'it's hard to get a feel for **anything** when it is **just** numbers on a screen' seems to illustrate the point well. Yet when the students anticipated their performance in the activity it was the reality of it that was prominent. This reality was dissipated by a perception of computer problems and perhaps more importantly by the acceptance of poor results. Failure was not due to the reality; reality could be where anyone fails, but often put down to poor judgment.

The understanding of what went wrong was generally limited to negative comments usually about the group performance. 'What we did wrong' was the form of the comments. There did seem to be a group emphasis in this learning and the collective expression of poor performance could be evidence for dialogic learning as discussed above. It seems also to fit with the Generation Y need to feel connected with group decisions [13]. One group confessed to 'communication breakdown'. There was a view by one respondent that system problems detracted from any learning. This seems to reinforce the interpretation of this exercise in terms of activity theory. Artefacts should be accessible by all; they should be used directly rather than a pseudo-object; and they could be available in diagnostic and explanatory models [7].

In response to the question about the improvement of understanding there was a general positive response. A majority of the respondents said their understanding had improved. A couple of those that did not blame the artefact; 'there's no way of applying it to a real company; it's just numbers on a screen. These students seem to see it as a pseudo-object in the above terms. It is not really typical of real business.

This has been an exploratory study in the use of a marketing simulation in an undergraduate course. It has explored the concepts of dialogic reasoning, the use of artefacts in active learning and the exploration of a 'constructive reality'. It has also gathered some evidence for the occurrence of negative feedback that is generated in this type of forum. This has contributed to the agenda for further research identified by [34] in exploring why simulations were not viewed as more effective

than case studies, service learning, and in-class discussions. It can be proposed that this emotional impact during the performance of the activity has a significant effect on learning that is not present in the activities above. This paper offers insights to the emotional reflection on negative comparative feedback via a simulation in a common arena. The proposition is that negative feedback (publicly) will increase attentiveness and learning beyond that of traditional feedback mechanisms that are not within the dialogic space. It has contributed to the call for further research on the contextualised nature of learning and the potential for ethnographic studies in this context [29]. Using the techniques of scenarios and survey has enabled an empirical exploration of the student engagement in this type of learning with technological artefacts. It was noted earlier that simulations could be used as assessment tools and that they work well with formative assessments. There was no attempt here to discuss the impact of the simulation on the quality of the assessment. This research was not designed to determine this but to concentrate on the process of engagement within the arena constructed by the assessment.

VII. FUTURE RESEARCH

The aim of this research was to explore the role emotional engagement relating to a simulation learning experience. The findings highlight the emotional anticipation and responses resulting in negative emotions that confirm a deep level of learning. It is proposed that further research could focus on the custodial nature of the learning where there occurs an intergroup transfer of the company, and the influence of this on emotional engagement, perceptions of reality, and emotional acceptance of relative failure. In addition the perceived importance of making sure the next stages of performance of their company by takeover group are successful. This will make a further contribution to the intra-group dynamics of learning literature.

REFERENCES

- [1] Anderson, J., (2003). The relationship between student perceptions of team dynamics and simulation game outcomes: An individual-level analysis. *Journal of Education for Business*, 81(2), 85-89
- [2] Ashworth, P., & Lucas, U., (2000). Achieving empathy and engagement: A practical approach to the design conduct and reporting of phenomenographic research, *Studies in Higher Education*, 25(3), 295-308
- [3] Barnett, R., (1997). *Higher education: A critical business*. Buckingham: Open University Press.
- [4] Butler, J. A. (1992). The use of teaching methods within the lecture format, *Medical teaching* 14(1)
- [5] Crowe, C., & Pemberton, A., (2000). 'But that's your Job!: Peer Assessment in Collaborative Learning Projects'. *Proceedings of the 3rd Effective Teaching and Learning at University Conference, 9-10 November 2000*. Brisbane: University of Queensland.
- [6] Dickenson, J. B., & Dickenson, C. D., (2012) The effect of Introducing coaching from an experienced business professional on performance in a computer simulation classroom exercise, *Journal of Instructional Pedagogies*, Vol.8, July,
- [7] Engestrom Y., (1992), *Interactive Expertise: Studies in Distributed Working Intelligence*, Research Bullitin 83, Department of Education, University of Helsinki
- [8] Engestrom, Y., (2009), *Expansive Learning* in Illeris K., (Ed), 2009, *Contemporary Theories of Learning*, London, Routledge
- [9] Exley, K., & Dennick, R. (2004) *Giving a Lecture: From Presenting to Teaching*, Taylor & Francis Ltd
- [10] Germain, L., (2009). Global: MBA still in Demand despite recession. *University World News*, Issue 104, Dec, Accessed <http://www.universityworldnews.com>
- [11] Hannon, J., & Bretag, T. (2010). Negotiating Contested Discourses of Learning Technologies in Higher Education. *Journal of Educational Technology & Society*, 13(1), 106-120
- [12] Hermann, M (2010). Harnessing Students' Creativity and Imagination as a Means to Effective Engagement in Sustainable Education, *Learning and Teaching in Higher Education (LATHE)*, Issue 5
- [13] Hogg, D. (2013). Application of Groupthink to Generation Y Decision Making Processes within a Professional Services Context in New Zealand. *International Journal Of Business & Management*, 8(8), 69-78
- [14] Hope, M. (2009). The importance of direct experience: a philosophical defence of fieldwork in human geography, *Journal of Geography in Higher Education*, Vol.22, no.2, pp.169-19
- [15] Jones O., McPherson A., & Thorpe R., (2010). *Entrepreneurship & Regional Development*, Vol. 22, Nos. 7-8, pp649-673
- [16] Jones, S.E., (2007). Reflections on the lecture: outmoded medium or instrument of inspiration?, *Journal of Further and Higher Education*, 31(4), 397-406
- [17] King, K. P (2000), The adult ESL experience: facilitating perspective transformation in the classroom, *Adult Basic Education*, 10(2), 69-90
- [18] Klassen, K., and Willoughby, K. (2003) In-class simulation games: assessing student learning. *Journal of Information Technology Education*, 2, 1-13
- [19] Kuhn, J. W. (1998) Emotion As Well As Reason: Getting Students Beyond "Interpersonal Accountability", *Journal of Business Ethics*, 17, 295-308
- [20] Kupers, W., (2005), Phenomenology of embodied implicit and narrative knowing, *Journal of Knowledge Management*, 9(6), 114-133,
- [21] Lainema, T., and Lainema, K. (2007) Advancing acquisition of business know-how: Critical learning elements. *Journal of Research on Technology in Education*, 40(2), 183-198
- [22] Lucas, 1998. Cited in Mortiboys, A., (2002), *The emotionally intelligent lecturer*, Birmingham, SEDA.
- [23] Markstrat, (2012), at; <http://www.stratxsimulationsonlinehome.aspx>
- [24] Marton, F., & Saljo, R. (1976), On qualitative differences in learning – Outcome and process, *British Journal of Educational Psychology*, 46, 4-11
- [25] McLean, R., Oliver, P. G., & Wainwright, D. W. (2010). The myths of empowerment through information communication technologies. *Management Decision*, 48(9), 1365-1377
- [26] Mezirow, J. (2009), An overview on Transformative Learning, in Illeris K., (Ed), 2009, *Contemporary Theories of Learning*, London, Routledge
- [27] Mortiboys, A., (2002), *The emotionally intelligent lecturer*, Birmingham, SEDA.
- [28] Neely, P., and Tucker, J (2012) Using Business Simulations As Authentic Assessment Tools, *American Journal of Business Education*, 5(4).
- [29] Phan, H. P., Maebuta, J., & Dorovolomo, J. (2010). The Relations between Personal Epistemology and Learning Approaches in Sociocultural Contexts: A Theoretical Conceptualization. *International Journal of Learning*, 17(5), 465-478.
- [30] Prasolova-Førland, E. (2004). Virtual spaces as artefacts: Implications for the design of educational CVEs. *International Journal of Distance Education Technologies*, 2(4), 94-115.
- [31] Rose, J., Fuller, M., Gilbert, L., & Palmer, S., (2010) Transformative Empowerment: stimulating transformations in Early Years Practice, *Learning and Teaching in Higher Education*, Issue 5, pp56-71
- [32] Salas, E., Wildman, J. L., & Piccolo, R. F. (2009). Using Simulation-Based Training to Enhance Management Education. *Academy Of Management Learning & Education*, 8(4), 559-573
- [33] Tan, K., Tse, Y., & Chung, P. (2010) A plug and play pathway approach for operations management games development. *Computers and Education*, 55(1), 109-117
- [34] Tanner, J.R., Stewart, G., Totaro, M.W., & Hargrave, M (2012) Business Simulation Games: Effective Teaching Tools Or Window Dressing?, *American Journal of Business Education*, 4(2), 115-128
- [35] Taylor, E. W. (2007). An update of transformative learning theory: A critical review of the empirical research (1999-2005) *International Journal of Lifelong Education*, 26, 173-191.

- [36] Williams, E (1992) Student Attitudes towards approaches to Learning and assessment. *Assessment and Evaluation in Higher Education*, 17(1), 45-58
- [37] Willis, P. (2011). Utopian scenario sketching: An imaginal pedagogy for life giving civilisation. *Australian Journal of Adult Learning*, 51(3), 479-497
- [38] Yadav, M. S., & Varadarajan, R. (2005). Interactivity in the electronic marketplace: An exposition of the concept and implications for research. *Academy of Marketing Science Journal*, 33(4), 585-603

Modeling and Simulation Analysis of Power Frequency Electric Field of UHV AC Transmission Line

Chen Han¹

College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Yuchen Chen²

College of Electronic and Electrical Engineering
Shanghai University of Engineering Science
Songjiang District, Shanghai 201620, China

Abstract—In order to study the power frequency electric field of UHV AC transmission lines, this paper which models and calculates using boundary element method simulates various factors influencing the distribution of the power frequency electric field, such as the conductor arrangement, the over-ground height, the split spacing and the sub conductor radius. Different influence of various factors on the electric field distribution will be presented. In a single loop, using VVV triangular arrangement is the most secure way; in a dual loop, the electric field intensity using reverse phase sequence is weaker than that using positive phase sequence. Elevating the over-ground height and reducing the conductor split spacing will both weaken the electric field intensity, while the change of sub conductor radius can hardly cause any difference. These conclusions are important for electric power company to detect circuit.

Keywords—boundary element method; electric field distribution; power frequency electric field; UHV

I. INTRODUCTION

The world's first UHV AC with tower of 1000 kV in dual loop-Anhui Power Transmission was officially put into operation on September 25, 2013. Thus the State Grid has built two 1000 kV AC and two 800 kV DC projects so far marking the great new achievements of China's UHV construction. Liu Zhenya said, "National Grid is building two 800 kV DC projects and at the same time developing the technology and equipment of 1100 kV DC whose transmission capacity can be 13.75 million kilowatts and economic transmission distance is 5000 km. Such projects make contributions to constructing trans-regional, transnational and transcontinental transmission channels. For example, if Africa and the Middle East can be connected, South America can build up a large power grid."

With the rising voltage level and increasing current of UHV transmission line, the effects of its power frequency electric field on the surrounding environment and ecology have attracted growing concerns. UHV power frequency alternating electric field can not only bring environmental issues like the radio interference and the audible noise but also has rather great harm on people's health and spirit [1]. Now the power frequency electric field distribution near the UHV AC transmission line has become one of the issues that scholars all over the world focus on and discuss.

This paper based on boundary element method which is used for modeling and calculation on the influence of various factors, the power frequency electric field distribution, such as the arrangement of conductors, the simulation of the height, split spacing and sub conductor radius.

II. THE CALCULATION METHOD OF THE POWER FREQUENCY ELECTRIC FIELD

The boundary element method is an accurate and effective engineering numerical method which uses the boundary integral equation defined on the boundary as governing equation. It transforms boundary element interpolation into algebraic equation sets through discrete method [2-4]. This essay uses line model when simulating the power frequency electric field of UHV transmission lines, thus reducing the difficulty of modeling as well as improving the operational efficiency.

Supposing that all radius of transmission line conductor are r_0 , the charge in the center line of wire is line charge, the charge is in the form of surface charge distribution on the surface of other devices and its density is θ , the electric potential of a point in this space can be expressed as

$$\varphi = \int_l \frac{\lambda dl}{4\pi\epsilon P} + \int_s \frac{\theta dS}{4\pi\epsilon P} \quad (1)$$

In the formula, P is the space between source point and field point, S is the surface integral region and l is the line integral path. When this formula is discrete, field and source units may be either line element or surface element. Thus, formula (1) can produce 4 kinds of discrete form. Supposing the field unit is the line unit and source unit is the surface unit, the formula can be written as

$$\sum_e \sum_j \sum_i \int_{l_e} N_j N_i \varphi_i dl = \frac{1}{4\pi\epsilon_0} \sum_e \sum_{e'} \sum_j \sum_i \int_{l_e} N_j \int_{S_{e'}} \frac{N_i \theta_i}{P} dS' dl \quad (2)$$

In the formula above, $i, j=1, 2, \dots, m, \dots, n$, corresponding to different discrete nodes, where m is the number of conductor discrete nodes, n is the number of discrete nodes of calculation model, and N_i together with N_j is the interpolation function. φ_i is the node potential in which θ_i is the destiny of node surface, e and e' are respectively the number of field unit and source unit, l_e represents the line integral path of field unit and $S_{e'}$ is

the surface integral area of source unit. If making some corresponding modification to the integral path and electric charge destiny in the formula (2), we can get the other 3 kinds of discrete forms. Making vectors:

$$B = [\lambda_1, \lambda_2, \dots, \lambda_m, \theta_{m+1}, \theta_{m+2}, \dots, \theta_n]^T$$

$$u = [\varphi_1, \varphi_2, \dots, \varphi_m, \varphi_{m+1}, \varphi_{m+2}, \dots, \varphi_n]^T$$

And the matrix

$$A_{ij} = 4\pi\epsilon \sum_e \int_{l_e} N_j N_i dl \quad (3)$$

$$C_{ij} = \sum_e \sum_{e'} \int_{l_e} N_j \int_{S_{e'}} \frac{N_i}{P} dS' dl \quad (4)$$

In the formulas above, $\lambda_1, \lambda_2, \dots, \lambda_m$ and $\varphi_1, \varphi_2, \dots, \varphi_m$ are respectively the linear density and electric potential of conductor node charge; $\theta_{m+1}, \theta_{m+2}, \dots, \theta_n$ and $\varphi_{m+1}, \varphi_{m+2}, \dots, \varphi_n$ are respectively surface density and electric potential of conductor node charge on the equipment's surface. Then the formula (1) can be written as follows:

$$CB = Au \quad (5)$$

Under the condition of knowing the device and the conductor potential, the surface density θ and the line destiny λ of conductor equivalent charge can be obtained. We can simulate different conditions of UHV transmission lines according to the formula and after obtaining θ and λ , can use the integral formula of the electric field to get the electric field intensity of any space point

$$E = \frac{1}{4\pi\epsilon} \sum_e \left[\iint_{S_{e'}} \frac{\theta(r-r') dS}{|r-r'|^3} + \int_{l_e} \frac{\lambda(r-r') dl}{|r-r'|^3} \right] \quad (6)$$

III. CALCULATION AND ANALYSIS OF POWER FREQUENCY ELECTRIC FIELD AROUND THE UHV TRANSMISSION LINE

A. Parameters of 1000kV AC UHV transmission line

China's 1000kV UHV demonstration project Jindongnan-Nanyang-Jingmen transmission line is all set up in single loop. Its full-length is 654 kilometers, transmission capacity is 600kVA, system nominal voltage is 1,000KV and its maximum operating voltage is 1,100KV. Setting up such a program has far-reaching influence on national energy security and reliable power supply. At present, UHV transmission line systems of our country mainly include single loop (IVI horizontal arrangement, VVV horizontal arrangement, IVI triangle arrangement and VVV triangle arrangement) and the common-tower double loop (I series vertical arrangement and V series vertical arrangement).

B. The factors influencing the power frequency electric field

According to the power frequency electric field calculation formula, the charge number on the wire surface, the distance between wires and other factors directly determine the electric field strength of the points in space [5]. Apart from the voltage of both sides, the number of charge on wire surface has relation

to the arrangement, types and sizes of the tower wire. Therefore, further researches are needed about the relationship between the distribution of power frequency electric field under the transmission line and several factors as the arrangement of wires, the over-ground height, the split spacing and the sub wire radius.

1) Influence of arrangement of wires

a) Different Arrangement of Single Loop

Supposing the over-ground height of the phase conductor $h=22m$, with the height 1.3m and the distance from the center wire 100m, the electric field distribution curve under four different arrangement of wires can be presented as follows:

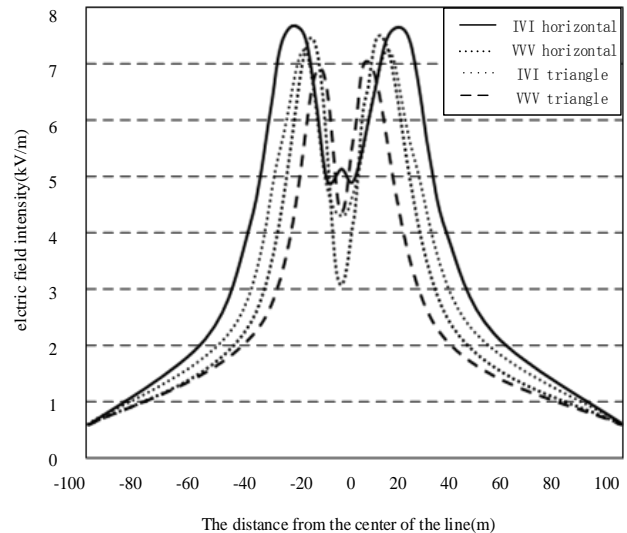


Fig. 1. The power frequency electric field of different arrangement of single loop

As can be seen from Fig.1, regardless of the arrangement, the electric field strength all reaches the maximum below the phase conductor and gradually weakens to both sides. Although the maximum field strength of IVI triangle arrangement is a little more than that of VVV horizontal arrangement, but the corresponding high field strength area of surpassing 4kV/m is smaller than the latter. In general, no matter what the maximum power frequency electric field and the corresponding high field strength area of under 4kV/m is, VVV triangle arrangement is always better than the other three arrangement. Therefore, using triangle arrangement can reduce the maximum field strength of the electric transmission line as well as the area under the cover of high field strength, thus reduces the construction risk and hidden danger.

b) Different Phase Sequence Arrangement of Double Loop

Double circuit arrangement of UHV AC can be subdivided into four types: positive and reverse phase sequence vertical arrangement of I and V. Supposing the over-ground height of the phase conductor $h=22m$, with the height 1.3m and the distance from the center wire 100m, the electric field distribution curve under four different arrangement of wires can be presented as follows:

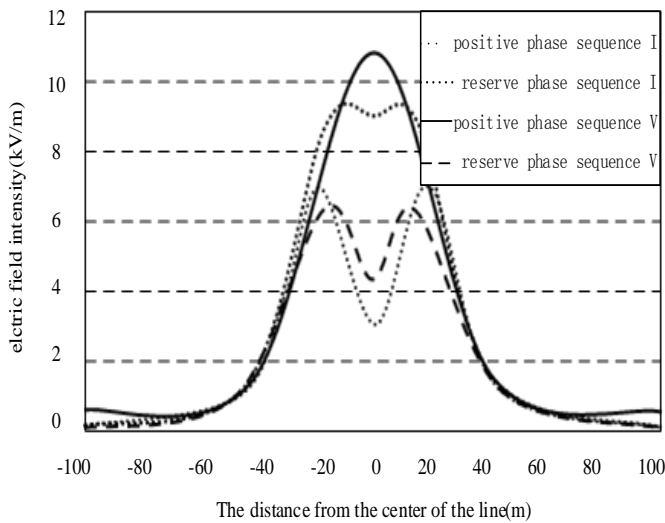


Fig. 2. The power frequency electric field of different arrangement of double loop

As we can see from Fig.2, the maximum field intensity of reverse phase sequence I is 2.2kV/m less than that of positive phase sequence [6-9]. The maximum field intensity of reverse phase sequence V is 4.5kV/m less than that of positive phase sequence. No matter using vertical arrangement of I or V, the maximum field intensity and the areas under high field strength of reverse maximum field strength are much lesser than those of positive maximum field strength. Therefore, in double circuit arrangement, using reverse phase sequence arrangement is an effective way to reduce the line strength and the area under high field strength.

2) The influence of conductor over-ground height

The changes of the power frequency electric field of UHV transmission line under different conductor over-ground height will be analyzed taking IVI horizontal arrangement for example. Supposing the conductor over-ground height h are 20m, 22m, 24m and 26m, the distribution curves of the power frequency electric field 1.3m above the ground can be presented as follows:

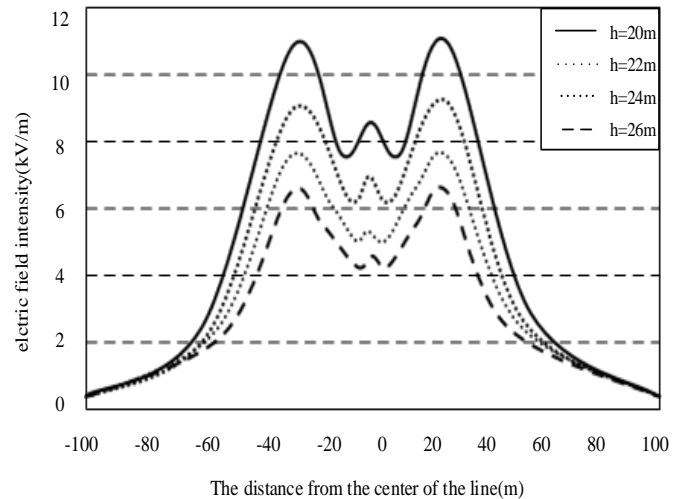


Fig. 3. The power frequency electric field of different conductor over-ground height

As we can see from Fig.3, the electric field intensity of transmission line becomes more and more weak with the increasing conductor over-ground height [10-12]. When the conductor over-ground height increases from 20m to 22m, the maximum electric field intensity reduces by 1.93kV/m. if the conductor over-ground height increases 2m each time, the maximum field intensity will reduces 1.45kV/m and 1.07kV/m in turn. Therefore, the decreasing amplitude of the electric field intensity is weaker. When the conductor over-ground height h is not so high, increasing the height has obvious influence on reducing the electric field intensity. With the increasing h , the economic input will gradually increase in order to reduce the same electric field intensity. Thus, we must take all the relevant factors into consideration when designing the lines and choose the appropriate conductor over-ground height.

3) The Influence of Split Spacing

The changes of the power frequency electric field of UHV transmission line under different split spacing will be analyzed using IVI horizontal arrangement as an example. As is known that the phase conductor of 1000kV UHV transmission line has

8 division structure, supposing the over-ground height of the phase conductor h is 22m and the split spacing are 0.3m, 0.4m, 0.5m and 0.6m, the distribution curves of the power frequency electric field 1.3m above the ground is presented in the following figure. As can be seen, the power frequency electric field intensity of UHV transmission lines decreases with decreasing split spacing. The maximum electric field intensity decreases from 0.42 to 0.54kV/m while the split spacing reduces every 0.1m. Therefore, reducing the conductor split spacing can decrease the electric field intensity of the transmission lines.

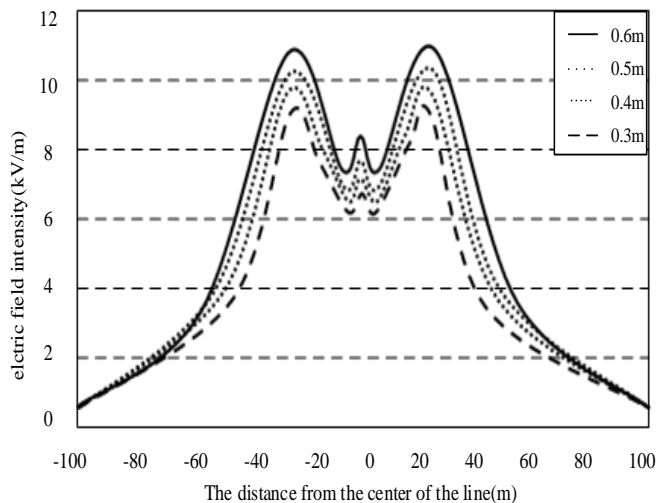
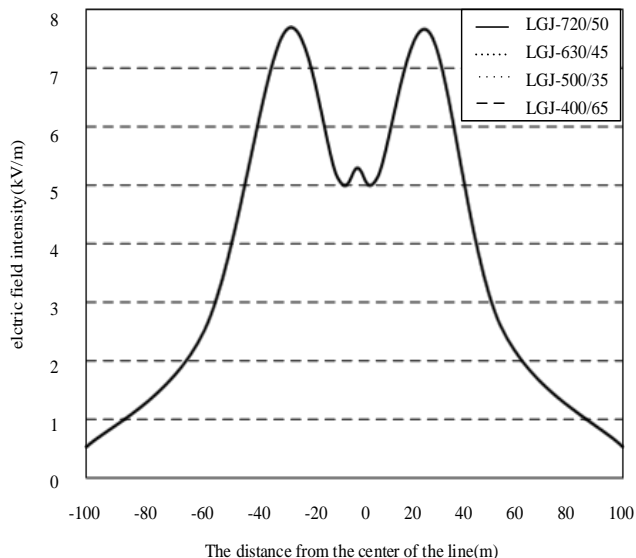
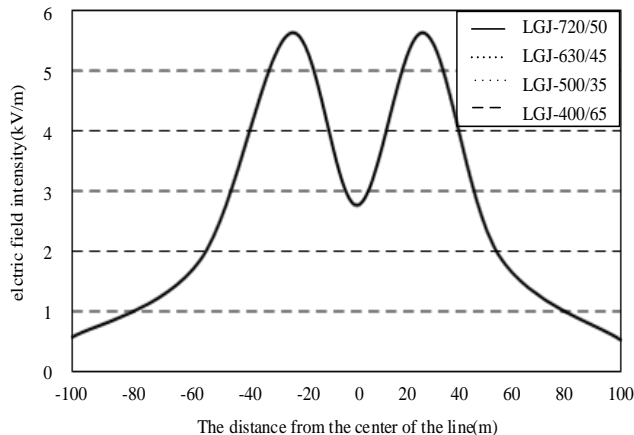


Fig. 4. The power frequency electric field of different split spacing



(a) IVI horizontal arrangement



(b) IVI triangle arrangement

Fig. 5. The power frequency electric field of different sub conductor radius

4) The influence of sub conductor radius

The relationship between conductor radius and the electric field strength will be analyzed using four types of sub conductor: LGJ-720/50, LGJ-630/45, LGJ-500/35 and LGJ-400/65. In IVI horizontal arrangement and IVI triangle arrangement, supposing the conductor over-ground height $h=22m$, the distribution curves of the power frequency electric field 1.3m above the ground is presented in the following figure. As is shown in the figure, the influence of different conductor radius on the power frequency electric field under the transmission lines can be ignored in the two different arrangement [13].

IV. CONCLUSION

Using boundary element method to simulate and calculate the power frequency electric field of the 1000kV UHV AC transmission line and analyze the distribution of electric field under different conditions, the following conclusions can be drawn:

- A. Different factors have different effects on the power frequency electric field distribution of UHV transmission lines. The conductor arrangement, the over-ground height and the split spacing all have a certain influence on the electric field while the influence of the sub conductor radius on the power frequency electric field can be nearly ignored. Adopting triangle arrangement in single loop and reverse phase sequence arrangement in double loop all have great influence on reducing the ground field strength. The over-ground height of phase conductor has great effect on the field strength and the higher the over-ground height is, the smaller the ground field strength will be. However, simply raising the conductor height will definitely increase the investment and difficulty of the project. Therefore, in order to decrease the ground field strength and reduce the construction cost as much as possible, we must choose the reasonable conductor over-ground height.

B. The influence of the splitted conductors of UHV transmission lines on the power frequency electric field mainly depends on the conductor radius while the equivalent radius is mainly decided by the conductor split spacing. The electric field intensity will decrease with the decreasing split spacing. The sub conductor radius basically has no influence on the power frequency electric field.

REFERENCES

- [1] Shu Yinbiao. Development and execution of UHV power transmission in China[J]. Electric Power, 2005, 38(11): 1-8.
- [2] Kovalev V, Panibratets A, Volkova O et al. The equipment for the AC 1150 kV transmission line[Z]. Moscow: All-Russian Electrotechnical Institute (GUP VED), 2005.
- [3] Shu Yinbiao. Current status and development of national grid in China[C]. IEEE/PES T&D Conference, Dalian, 2005.
- [4] Okamoto H. System design in 1000kV AC transmission conducted by TEPCO[Z].
- [5] Wu Jingru, Xu Yongxi. Development prospect of UHV AC power transmission in China[J]. Power System Technology, 2005, 29(3): 1-4.
- [6] Shao Fangyin. Phase conductor configuration and power frequency electromagnetic environment of UHV transmission lines in China[J]. Power System Technology, 2005, 29(8): 1-7.
- [7] Zhou Hao, Yu Yuhong. Discussion on several important problems of developing UHV AC transmission in China[J]. Power System Technology, 2005, 29(12): 1-9.
- [8] Zhang Wenliang, Wu Weining, Hu Yi. Study of UHV transmission technology and developing of power system in China[J]. High Voltage Engineering, 2003, 29(9): 16-18.
- [9] Chen Yong. Discussions on UHV conductors and tower structure in China[J]. High Voltage Engineering, 2004, 30(6): 38-41.
- [10] He Jiali, Li Yongli, Li Bin, et al. Relay protection for UHV transmission lines: Part two disposition of relay protection. Automation of Electric Power Systems, 2002, 26(24): 1-6.
- [11] Dong Xinzhou, Su Bin, Bo Z Q, et al. Study of special problems on protective relaying of UHV transmission line. Automation of Electric Power Systems, 2004, 28(22): 19-22.
- [12] Ferrero RW, Rivera JF, Shahidepour SM. ADynamic Programming Two-stage Algorithm for Long-term Hydro-thermal Scheduling of Multireservoir Systems[J]. IEEE Transactions on Power Systems, 1998, 13 (4) : 1534 -1540.
- [13] M. Duvall and E. Knipping. Environmental Assessment of Plug-in Hybrid Electric Vehicles[R]. New York: EPRI, 2007: 23-27.

A Proposal of SNS to Improve Member's Motivation in Voluntary Community Using Gamification

Kohei Otake/ Yoshihisa Shinozawa/ Akito Sakurai
School of Science for Open and Environmental Systems,
Keio University
Kanagawa, Japan

Makoto Oka
Industrial & Management Systems Engineering
Tokyo City University
Tokyo, Japan

Tomofumi Uetake
School of Business Administration
Senshu University
Kanagawa, Japan

Ryosuke Sumita
Hitachi Solutions Ltd
Kanagawa, Japan

Abstract—Recently, the number of voluntary communities such as local communities and university club activities are increasing. In these communities, since there are various types of members and there are no binding forces, it is usually difficult to maintain and improve member's motivation. To maintain and improve member's motivation, most of these communities use social networking services (SNSs). However, since existing SNS offer few functions for voluntary community, it is difficult to solve this problem. This research focused on the concept of gamification and proposed an SNS to improve member's motivation of voluntary community. First, the authors analyzed the current conditions and members of a voluntary community. Based on this analysis, the authors found that an SNS to improve member's motivation of voluntary community requires functions which support member's personal activities and also functions which increase social activities. Next, the authors built an SNS that had these functions by applying the concept of gamification. The authors implemented the SNS for a University club's activities for one month and showed the effectiveness of our SNS.

Keywords—Gamification; Voluntary Community; Motivation Management

I. INTRODUCTION

Recently, the number of voluntary communities such as local communities and university club activities are increasing. In these communities, since there are various types of members and there are no binding forces, it is usually difficult to maintain and improve member's motivation. Under this situation, companies, schools, and municipalities are quite actively using social networking services (SNSs) with the aim of revitalizing communication within communities formed in the real world, along with maintenance and improvement of motivations among members. However, application of SNSs in non-profit and voluntarily-formed communities has found it difficult to maintain and improve their motivation. This is because of a feature of voluntary community: users who belong to voluntary community have different motivation. There are various types of members (beginner, expert or highly motivated users, unmotivated users) in a voluntary community. Moreover, the purpose of belonging to voluntary communities is ascribed not to financial profit, but to a sense of satisfaction and amusement.

In addition, most of the SNSs run by voluntary communities do not have binding forces, where each individual user has different motivation. In order to manage users' motives effectively in voluntary community, users probably need to keep motivation by means of certain methods other than some binding forces. The authors thought that SNSs had potentials to solve the problem. However, since existing SNS offers few functions for voluntary community, it is difficult to solve this problem.

Gamification has attracted much attention recently as a method to help users to maintain and improve their motivation. Gamification is defined as "to use gaming elements, such as concept, design, and mechanics of a game, for social activities or services other than the game itself." This idea became widespread after 2010, and has empirically been applied to course design in university education [1], rehabilitation activities in the medical field [2], communication on the network [3], and e-learning [4] with the aim of maintaining and improving users' motivations. This application has produced beneficial effects. At the same time, however, this field is still in the sprouting stage. Only a few attempts have been made so far at verifying the effectiveness of implementing and running an SNS system.

This paper is organized as follows. In section 2, the purpose of this research is described. In section 3, previous researches on profit-oriented communities are reviewed and gamification is explained. In section 4, the results of analysis of the data obtained from a questionnaire and interview study are described. Based on the results, in section 5, the SNS with four functions, owing to the concept of gamification that maintains and improves motivation are proposed. In section 6, results of analysis of the experiment are described. Section 7, concludes the paper.

II. THE PURPOSE OF OUR RESEARCH

In this research, the authors focused on the Senshu University Philharmonic Orchestra (SUPO), a voluntary community, as the experimental subject. The authors built and evaluated an SNS called f-simo for users to maintain and improve their motivations based on gamification; more specifically for each member to maintain and improve their motivation for practice. Our ultimate goal is to build a motivational support and im-

provement SNS that fosters a better environment for community members to participate. The authors built an SNS that has these functions by applying the concept of gamification. The authors implemented the SNS for the university club activities for one month and showed the effectiveness of proposed SNS.

III. PREVIOUS RESEARCHES

A. Existing SNS for Profit-Oriented Communities

Use of SNSs in profit-oriented communities with the purposes of information sharing [5], knowledge management [6], and communication support has achieved certain positive results and contributed to profit-earning activities.

In voluntary communities, however, although SNSs are used with the purposes similar to those of profit-oriented communities, similar achievements are not always guaranteed due to the differences in responsibilities and the binding force for belonging to a community. In voluntary communities, it is especially difficult to care for the amount of website traffic and also retain active users continuously [7]. It is necessary to develop functions in SNS for motivation improvement based on the features of a voluntary community. Moreover, it is important to propose appropriate functions considering attributes of the user who belongs to voluntary community.

Therefore, this research focused on gamification that has attracted much attention recently as a method to help users to maintain and improve their motivations through an SNS. It is possible to improve user's motivation and activity using gamification.

B. What is Gamification?

Gamification is defined as "to use gaming elements, such as concept, design, and mechanics of a game, for social activities or services other than the game itself." The following seven methods are included as the gamification.

- 1) *Honorific badges or titles are given according to achievements*
- 2) *Names and scores of competitors are displayed on a real-time basis*
- 3) *The graphic interface shows the progress of each task*
- 4) *Virtual currency is introduced to promote purchases of virtual goods*
- 5) *Rewards such as coupons or gifts are provided*
- 6) *Assignments that encourage users to collaborate together are presented*
- 7) *Simple games are prepared between activities in order to keep users from being bored*

In voluntary communities members do not always maintain and improve their motivations simply by financial rewards. For example, they do so by confirming their growth. As a motivation that aims toward personal growth, humans have a need for their achievements to be recognized or commended by others [8][9], in addition to the convenience that can be acquired after growth. Incorporation of gamification in SNSs is likely to meet

the need for recognition of each individual user, maintain and improve the individual motivation, and enhance the entire community performance.

IV. ANALYSIS OF THE CURRENT CONDITIONS

A. Analysis of the Experimental Subjects

The experimental subject, the Senshu University Philharmonic Orchestra (SUPO), is an amateur orchestra consisting of volunteer students. All the members are different in their position in this orchestra. For example, a section leader exists in each musical instrument section. As for analysis of the current conditions, in this research the authors conducted a questionnaire survey regarding motivations of the members of SUPO in order to clarify the motivational problems that this orchestra had.

First, the authors conducted questionnaire and interview surveys regarding "Problems associated with maintenance and improvement of motivations" for 25 male and female members of SUPO. As a result of the survey, existence of the following two problems became clear.

- It is difficult for members to realize the benefits of self practice
- Achievement levels of practice of other members are not supplied

With respect to the problem of difficulty for members to realize the benefits of self practice, the authors considered the visualization of the amount of practice. Performing technique of musical instruments is evaluated significantly based on sensations. In other words, it is very difficult to measure performing skills of each individual player on a quantitative scale. For this reason, each individual cannot realize his/her own growth that much, and this makes it difficult for him/her to maintain and improve motivation for practice. Humans are motivated when they realize their own growth, and their motives are improved [10]. Based on this reason, the visualization of the amount of practice can probably solve this problem.

With respect to the problem of not being able to know the achievement level of practice of other members, the authors examined methods for sharing information regarding the achievement level of practice which could be effective and understood intuitively. During weekdays, each individual member of SUPO practices voluntarily. This makes it difficult for each of them to know the achievement level of performance of other members. Specifically, the section leaders devote their practice time to observe each member to understand their achievement level of performance. Therefore, these problems could be solved if an SNS could show the achievement level of performance of other members in a form by which all members can understand it. This means that the section members can understand the achievement level of practice of each other, and since information sharing produces a competition, this encourages members to improve their motivation for practice. As a result, the amount of practice probably increases.

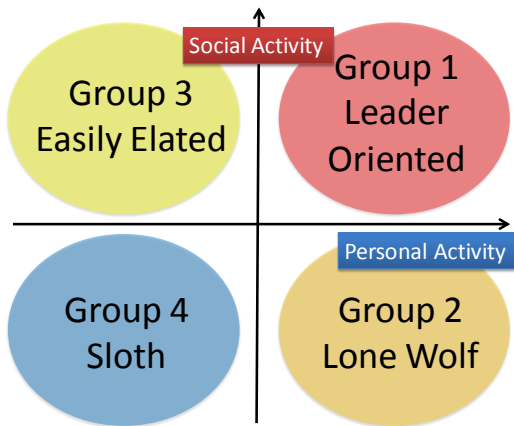


Fig. 1. Grouping by personal activity and social activity

This research propose functions that incorporate the gamification elements, the visualization of achievement level of one's own practice and a method for the sharing of information regarding the achievement level of performance of other members, in order to maintain and improve users' motivation effectively.

B. Analysis of the member who belongs to SUPO

It is important to propose appropriate functions after considering the attribute of the user who belongs to voluntary community. Therefore, the authors classified the members of SUPO by the attributes. It became clear that the members can be grouped into four types from the viewpoint of individual practice and orchestral rehearsal (personal activity) and information sharing (social activity) as concluded from the questionnaire and interview survey the authors conducted before the experiments (Fig 1).

The feature of each group is described below.

- Group 1 Leader Oriented: Members who did more practice and information sharing positively. The presence of Leader Oriented members in the community makes it more active.
- Group 2 Lone Wolf: Members practice much but they do not contribute to information sharing positively. Although they are serious and important for voluntary communities, they are sometimes isolated.
- Group 3 Easily Elated: Members contribute to information sharing positively but they do not practice actively. Members who belong to the Easily Elated group might become good or bad.
- Group 4 Sloth: Members who do not contribute to information sharing positively and they do not practice actively. When members of Sloth group are large, it is not rare that a voluntary community collapses. Therefore, it is important to reduce them in number in a voluntary community.

The purpose of this research for SUPO is to increase practice time and to make more members shift to Group 1(Leader Oriented). In SUPO, all of the members perform same music at the concert. All of the members have to raise their performance

skill. Therefore, it is very important to increase the practice time of members who belongs to Group 3(Easily Elated) and Group 4(Sloth).

V. PROPOSAL OF AN SNS FOR MOTIVATION MAINTENANCE AND IMPROVEMENT

With the analysis results of the current conditions, this research proposed an SNS called fortissimo (f-simo) that supports members to maintain and improve their motivation for practice by visualizing the practice achievements and understanding the achievement level of performance of each member (Fig. 2)[11]. In order to improve motivation, this research proposed and introduced methods incorporating gamification in f-simo. In addition, f-simo was made accessible from multi-devices, such as a personal computer (PC) and a smart phone. The authors prepared account for each member.

A. The Outline of Our Proposed SNS

This f-simo provides the following four functions.

- 1) Improvement of the Avatar Using the Experience Points
- 2) Graphical Representation of Practice Time
- 3) Presentation and Sharing of Rankings
- 4) Character Shimosuke that Grows Up by Cooperation of All Section Members

As for these functions, 1, 2, and 6 of gamification methods (Section III-B) were incorporated. Next section will describe each of the above functions. The figure below shows the system image of f-simo (Fig.2).

After practice, users enter their practice time and concentration level regarding practice based on a self-evaluation scale of one to five. The concentration level is required to be entered because the practice achievements could depend on the concentration level even though each member spent the same period of time of practicing. The experience point is calculated from these two numerical values. The experience point is used

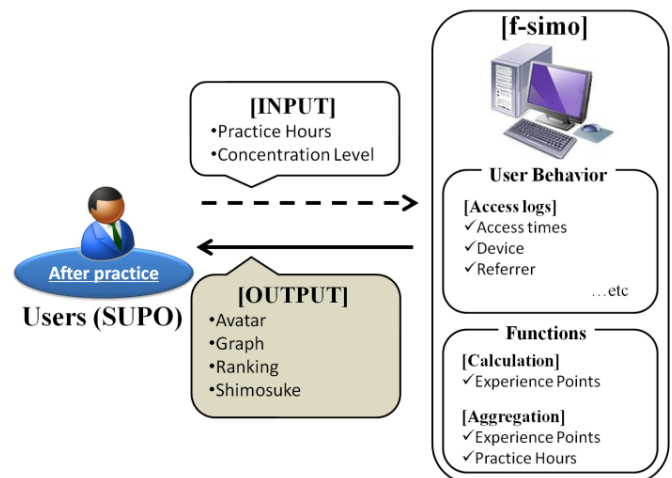


Fig. 2. System image of f-simo

as an index which is being accumulated after practice. These items, the practice time, the concentration level, and the experience point, are also used for improvement of the avatar's level

and the character Shimosuke that grows up based on the cooperation of all orchestra members. Avatars referred to by this research indicate characters of users used in our proposed SNS. Level is an index of the experimental point which is able to compare with other member. When the level is improved, the title, and the background color of the avatar changes. "Shimosuke" is the character that grows up by collaborative of users.

Our proposed SNS has four functions aimed at increasing personal activity and social activity. "1. Improvement of the Avatar Using the Experience Points," "3. Presentation and Sharing of Rankings" and "4. Character Shimosuke that Grows Up by Cooperation of All Section Members" corresponds to the improvement in personal activity. Moreover, "2. Graphical Representation of Practice Time" corresponds to the improvement in social activity. The feature of each function is described below. (IV-B and IV-C).

B. Functions to Improve Personal Activity

1) Improvement of the Avatar Using the Experience Points.

The authors set this function so that the user's avatar level increases by accumulating experience points and the avatar's title changes accordingly (Fig.3). The purpose of this function was to visualize the practice achievements not only with numerical values, but also with the title and change of color so that the user can realize his/her growth. This function was inspired by the first gamification method mentioned in III-B, "Honorific badges or titles are given according to achievements."

2) Character Shimosuke that Grows up by Cooperation of All Section Members.

The authors implemented a character that grows when all the members of each section spend a certain amount of time for practice (Fig.4). The aim of this function was to give a common assignment for all section members to work on so that they can strengthen their group ties and improve their motivation. This function was inspired by the sixth gamification method in section III-B, "Assignments that encourage users to collaborate together are presented."

3) Presentation and Sharing of Rankings

This function aims toward improving the users' motivation based on a sense of competition. There are two rankings, the practice time of the previous day and the accumulated practice time (Fig.5). These two different rankings were introduced in order to avoid a decline in the motivation for practice of those who cannot come to practice on weekdays. This function was inspired by the second gamification method mentioned in III-B, "Names and scores of competitors are displayed on a real-time basis."

C. Functions to Improve Social Activity

1) Graphical Representation of Practice Time

The authors implemented the graphical representation function so that each member can understand the transition of their practice time intuitively. This function also enables members to compare their own practice time with their average practice

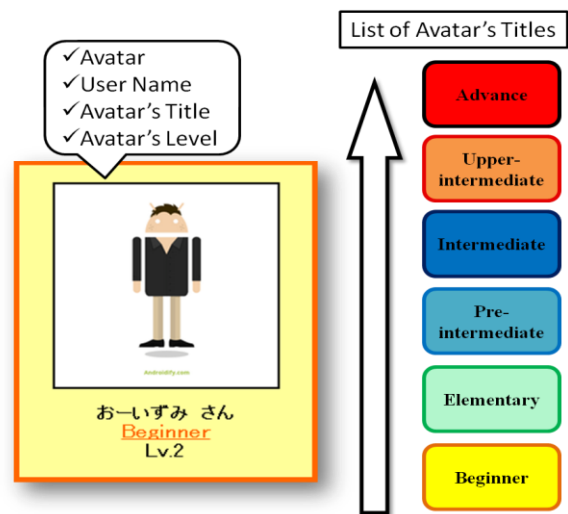


Fig. 3. The example of usage and list of avatar's titles

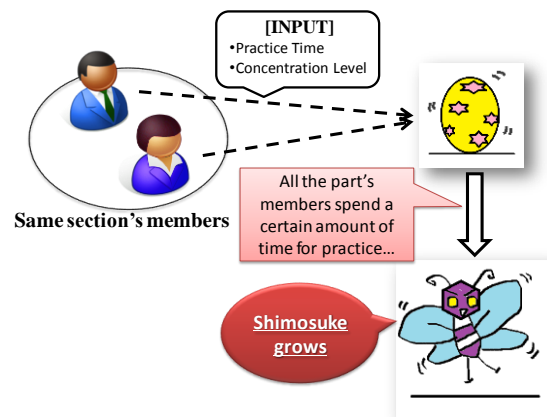


Fig. 4. The image of the cooperation which uses Shimosuke



Fig. 5. An example of a graphical representation of Ranking

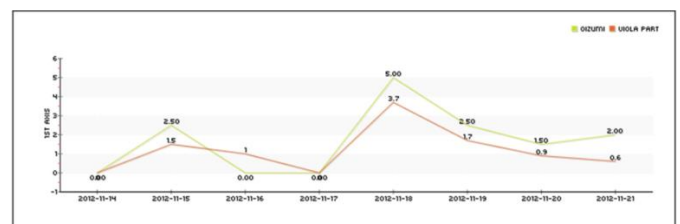


Fig. 6. A user's average practice time and the practice time of other members within "viola" section

time and the practice time of other members within the same section (Fig.6). This function aims toward improving the users' motivation based on a sense of competition. This function was inspired by the second gamification method mentioned in III-B, "Names and scores of competitors are displayed on a real-time basis."

VI. ANALYSIS OF EXPERIMENTAL RESULTS AND ACCESS LOG

In order to verify the effectiveness of our proposed SNS, the authors conducted an experiment with 25 male and female SUPO members as the experimental subjects by having them use our proposed SNS for one month, from November 25 to December 15 of 2012. In this verification experiment, the authors studied the effectiveness of our proposed SNS regarding the two items:

- 1) *Members' motivation for practice*
- 2) *Evaluations of each function*

Item 1 above was determined by quantifying the number of practice sessions of members per day and the transition of practice time before and after our proposed SNS was introduced (TABLE I). Item 2 above was determined based on the results of the questionnaire survey conducted after the experiment (TABLE II). Members evaluated our proposed functions on a scale of 5 from the following questions (TABLE II). The tables below summarize the experimental results.

TABLE I shows that the number of practices and practice hours increased each week. With respect to the point as to whether all section members improve their motivation for practice, as shown by TABLE II, the average number of practice sessions and time per day increased as the weeks went by. With respect to the evaluations for each function, in all functions, evaluation values of questionnaire of after 3rd week was higher than that of after 2nd week. Moreover, the authors interviewed section leaders after the verification experiment and the authors got opinion that management of the section member became easy using visualization functions.

On the other hand, a low average evaluation value of questionnaire was confirmed with respect as to whether Shimosuke could serve to strength the section ties. This is likely due to the short experimental period, just one month, so that the change of the character based on the collaborated assignment could only slightly be confirmed. Actually at the end of the experiment, the character brought up by the members of only one of four sections was confirmed to have grown from the initial stage. The characters of the other three sections remained the same as the initial stage. This point shows that the parameters related to the character growth needs adjustment. When compared to the evaluation value of questionnaire of after 2nd week, however, the evaluation value of questionnaire of after 3rd week did increase. This shows that continuous use of this function could strengthen the ties within the section.

TABLE I. COMPARISON OF THE NUMBER OF PRACTICE SESSIONS AND PRACTICE TIME BEFORE AND AFTER OUR PROPOSED SNS WAS INTRODUCED

	Before the introduction	After 2nd week	After 3rd week
Number of practice sessions (average times)	0.4	0.6	0.8
Practice time (average hours)	1.0	1.2	1.7

TABLE II. EVALUATIONS FOR EACH FUNCTION

Question	After 2nd week	After 3rd week
Did the change in your avatar make you realize your growth more than ever before?	3.16	3.36
Did graphic representation of practice time and ranking presentation increase your sense of competition?	3.24	3.36
Did Shimosuke serve to strengthen the ties within your section?	2.60	2.80

A. Analysis of Practice Time Based on Classification

In this section, the authors describe the detailed analysis of practice time and access time based on user classification.

First, the authors classified the SUPO members using individual practice and orchestral rehearsal (personal activity) and information sharing (social activity). In this research, social activity refers to information sharing, such as a practice situation and practice contents. Generally, communication is being activated in the community with many members who contribute information sharing positively. Information sharing is very important to maintain and improve member's motivations in voluntary communities. The classification was performed using access time of 1st week from the experimental start and practice time of 1st week. In Fig. 7, the authors show a scatter-plot of users where the vertical axis is social activity and the horizontal axis is personal activity.

The authors classified the SUPO members into four groups along with four quadrants of Fig.7. The number of members belonging to each group is shown in TABLE III. The members of Group 3 (Easily Elated) and Group 4 (Sloth) are with short practice time at the start of the experiment. TABLE III shows that there are 11 members in Group 3 (Easily Elated) and Group 4 (Sloth). Focusing on these 11 members, the authors analyzed effectiveness of gamification functions using access log, by comparing practice time in the 1st week to the 2nd week

TABLE III. THE NUMBER OF MEMBERS WHO BELONGS TO THE RESPECTIVE GROUPS

Group Name	Number of Users
Group 1 : Leader Oriented	9
Group 2 : Lone Wolf	6
Group 3 : Easily Elated	3
Group 4 : Sloth	8

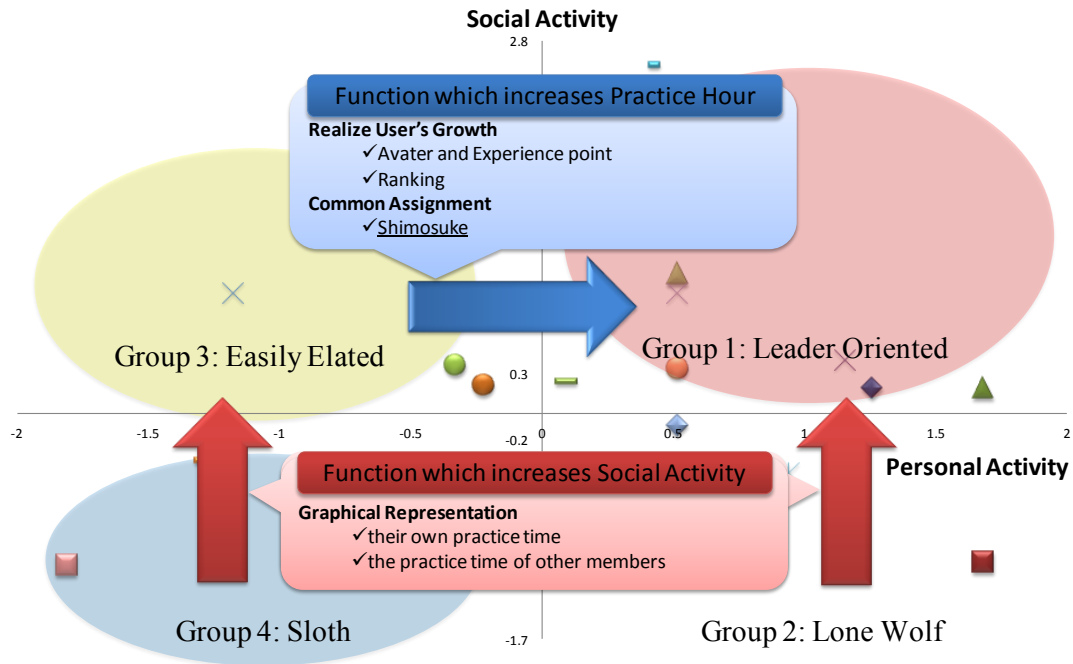


Fig. 7. The scatter diagram which set social activity on vertical axis against personal activity on the horizontal axis, and the placing by the proposal functions

TABLE IV. AVERAGE OF MEMBER'S ACCESS TIME WHO INCREASED PRACTICE TIME

	1st week	2nd week	3rd week
Avatar	8.4	11.6	9.2
Graph	1.6	4.6	1.0
Graph other members	0.4	1.6	0.2
Ranking	2.0	3.6	2.4
Shimosuke	0.8	2.0	0.8

TABLE V. AVERAGE OF MEMBER'S ACCESS TIME WHO DECREASED PRACTICE TIME

	1st week	2nd week	3rd week
Avatar	9.8	5.0	5.0
Graph	0.8	0.0	0.3
Graph other members	0.8	0.0	0.5
Ranking	2.0	0.3	0.5
Shimosuke	1.3	1.5	0.5

and in the 2nd week to the 3rd week.

First, the authors performed comparison of the 2nd week with the 1st week. 9 out of 11 members showed upward tendency at practice time. Moreover, about 70 percent of the members whose practice time improved also increased the number of times of accesses. Focusing on members who increased practice time and the number of times of accesses, on the 2nd week, they more often accessed pages of graphical representation of their own practice time and the practice time of other members and Shimosuke, compared with members who decreased practice time (TABLE IV and TABLE V). Through the result, members who belong to Group 3 (Easily

Elated) and Group 4 (Sloth), i.e., members who did not practice positively, increased, from the 1st week to the 2nd week. Moreover, they often used the gamification functions on our proposed SNS.

Second, the authors performed comparison of the 3rd week with the 2nd week. Focusing on the members whose practice time improved and whose number of times of accesses increased on the 2nd week, it became clear that about 80 percent of users decreased practice time a little from the 2nd week to the 3rd week. However, compared with the 1st week, all the members increased practice time. Moreover, members who decreased practice times also decreased the number of times of accesses. Analyzing the access log, it was found that they decreased the number of times of accesses to pages of gamification. On the other hand, these numbers of times of accesses are as large as those of members who decreased practice time (TABLE IV and TABLE V).

Results of analysis of comparing the 1st week to the 2nd week and of the 2nd week to the 3rd week, the effect and problem of our proposed SNS became clear. First, from the 1st week to the 2nd week, positive correlation was seen between the number of times of accesses, and practice time. Moreover, many of the users who increased practice time, used functions using gamification. The authors performed independent t-test, to compare the mean of two different members. As results, p-value was below 0.10 (10%), we could reject the null hypothesis and conclude that there is a statistically significant difference between the two population means with significance level 10%. From these results our proposed SNS gave a certain effect on the user who often used this SNS.

On the other hand, some members did not use our proposed SNS. The authors need to reconsider our proposed SNS so that it becomes easier to use. Moreover, from the 2nd week to the

3rd week, the downward tendency was seen in the number of time of accesses and practice time. Restriction of practice time is considered as one of the cause of the result. The member who belongs to SUPO is a university student, and the practice time which can be spared for orchestra activity is restricted, i.e. each member's practice time may have reached its limit of being spared. Another cause is the problem with setting of parameters of gamification. For example, at the end of the experiment the character Shimosuke brought up by the members of only one of four sections was confirmed to have grown from the initial stage. In the interview after the end of an experiment, there are some options such as "I want to raise a level more" and "I want to know how to make Shimosuke will grow up." When using gamification in a voluntary community, it is important how to change outputs, i.e. avatar, level and character. Change outputs attract the user's interest and leads to improvement of the rate of utilization. Moreover, it is necessary to also consider game balance and outputs simultaneously. For example, in social games, users can get experimental points easily at the beginning. Therefore growth is easy to realize. Members may find difficulties to realize growth and become weary to the function as a result. The authors need to develop more interesting functions, and to make users to use them continuously.

Through these experimental results, the authors consider that the effectiveness of our proposed SNS was successfully verified. These results clarify that the following gamification methods are effective for voluntary communities: 1. Badges obtained according to the achievement level, or level determination, and 2. Presentation of the names and scores of the current competitors on a real-time basis. When it comes to method 6, "Assignments that promote collaborative work," the continuous use of this function could increase the effectiveness of this system to a greater extent.

VII. CONCLUSION

In this research, the authors proposed an SNS for improving motivation by utilizing gamification, targeting one university club, the Senshu University Philharmonic Orchestra, as an example of voluntary communities. The purpose of this research was to maintain and improve the motivations of each individual orchestra member for practice. In voluntary communities, since there are various types of members and there are no binding forces, it is usually difficult to maintain and improve member's motivation. To maintain and improve member's motivation, most of these communities are using SNS. However, since existing SNS offers few functions for voluntary community, it is difficult to solve this problem. First, the authors analyzed the current conditions and members of voluntary community. Based on this analysis, the authors found that an SNS to improve member's motivation of voluntary community requires the functions which support member's personal activities and the functions which increase social activities. Next, the authors built an SNS that visualized practice achievements and enabled sharing of information among section members, while applying the concept of gamification in order to reinforce these functions. The effectiveness verifica-

tion experiment conducted for one month verified the effectiveness of our proposed SNS with respect to the following functions: "1. Honorific badges or titles are given according to achievements." and "2. Names and scores of competitors are displayed on a real-time basis." With respect to function 3, "6. Assignments that encourage users to collaborate together are presented," continuous use of our proposed SNS will probably strengthen the ties between members.

The authors analyzed access logs and practice time further. Focusing on members who did not practice positively at the start of the experiment, we analyzed effectiveness of gamification functions using access log, by comparing practice time in the 1st week to the 2nd week and in the 2nd week to the 3rd week. About 70 percent of the members whose practice time improved also increased the number of times of accesses and positive correlation was found between the number of access times, and practice time in the 1st week to the 2nd week. Moreover, they often used the gamification functions on the proposed SNS. On the other hand, the downward tendency was found in the number of time of accesses and practice time in the 2nd week to the 3rd week. They may be attributed two facts. One is that since the member who belongs to SUPO is a university student, the practice time which can be spared for orchestra activity is restricted. The other is a problem with setting of parameters of gamification. The proposed SNS need to be improved to be easier to use.

REFERENCES

- [1] Y. Kishimoto and K. Mikami, "About effectiveness of university education utilizing Gamification," Journal of Digital Games Research Association Japan, In An annual general meeting 2012, Japan, 2013. (in Japanese)
- [2] H. Matsuguma, S. Fujioka, A. Nakajima, K. Kaneko, J. Kajiwara, K. Hayashida and F. Hattori, "Research and Development of Serious Games to Support Stand-up Rehabilitation Exercises," Information Processing Society of Japan, Vol53, No3, 2012, pp. 1041-1049. (in Japanese)
- [3] Y. Yano, Y. Muramoto, K. Kitahara and M. Okubo "A Proposal of SNS for activation Physical Community," Information Processing Society of Japan, 2013, pp. 153-155. (in Japanese)
- [4] T. Matsumoto, "Possibility of e-Learning by using Gamification," Japanese Society for Information and Systems in Education, 27(3), 2012, pp. 34-40. (in Japanese)
- [5] Chisokukan, <http://jp.fujitsu.com/group/fst/services/chisokukan/> (2014-11-28 author checked)
- [6] T. Sabetto, M. Kotani, "Utilizing the enterprise social network for knowledge management," The journal of Information Science and Technology Association, 62(7), 2012, pp. 296-301.
- [7] R. Yamaguchi, F. Toriumi and K. Ishii, "Analysis of user behavior in SNS," Information Processing Society of Japan, Proceeding, 2009, pp.69-74.
- [8] H. Ota, "Recognition and motivation [the proved effect]," Dhobun sha shuppan, 2011.
- [9] Harvard Business Review Anthology: Power to motivate - Theory of motivation, and practice, 2009.
- [10] H. Frederick, "Motivation to Work," John Wiley & Sons Inc, 1959
- [11] K. Otake, R. Sumita, M. Oka, Y. Shinozawa, T. Uetake and A. Sakurai, "A Proposal of a Support System for Motivation Improvement Using Gamification," the International Conference, SCSM 2014, 2014, pp. 571-580.

Fast Vertical Mining Using Boolean Algebra

Hosny M. Ibrahim

Information Technology Department
Faculty of Computer and
Information, Assiut University
Assiut, Egypt

M. H. Marghny

Computer Science Department
Faculty of Computer and
Information, Assiut University
Assiut, Egypt

Noha M. A. Abdelaziz

Information System Department
Faculty of Computer and
Information, Assiut University
Assiut, Egypt

Abstract—The vertical association rules mining algorithm is an efficient mining method, which makes use of support sets of frequent itemsets to calculate the support of candidate itemsets. It overcomes the disadvantage of scanning database many times like Apriori algorithm. In vertical mining, frequent itemsets can be represented as a set of bit vectors in memory, which enables for fast computation. The sizes of bit vectors for itemsets are the main space expense of the algorithm that restricts its expansibility. Therefore, in this paper, a proposed algorithm that compresses the bit vectors of frequent itemsets will be presented. The new bit vector schema presented here depends on Boolean algebra rules to compute the intersection of two compressed bit vectors without making any costly decompression operation. The experimental results show that the proposed algorithm, Vertical Boolean Mining (VBM) algorithm is better than both Apriori algorithm and the classical vertical association rule mining algorithm in the mining time and the memory usage.

Keywords—association rule; bit vector; Boolean algebra; frequent itemset; vertical data format

I. INTRODUCTION

Data mining is defined as “The non trivial extraction of implicit, previously unknown and potentially useful information from databases” [1]. Association rules mining is an active research topic in the data mining field, which is the key step in the knowledge discovery process [2, 3]. Mining frequent itemsets (FIs) is the most important task in mining association rules. Furthermore, frequent itemsets detection can be used in other data mining tasks like classification and clustering [4-6]. Therefore a lot of mining frequent itemsets algorithms has been proposed. None of these methods can outperform other methods for all types of datasets with every minimum support [7, 8]. The well-known (FIs) mining algorithms are based on either horizontal or vertical data structures. Some of the horizontal based algorithms are Apriori, AprioriTid and FP-growth. Apriori is a level-wise algorithm that adopts an iterative method to discover frequent itemsets, in which k frequent itemsets is created by joining k-1 frequent itemsets, and then remove itemsets that contain non-frequent items. Non frequent items are detected by scanning the database once for each itemset to calculate its support. This is the most important shortcoming of Apriori algorithm [9]. AprioriTid has been proposed to improve Apriori algorithm’s efficiency by reducing the overhead of I/O by scanning the database only once in the first iteration [9]. FP-growth algorithm mines frequent item sets by scanning the database only two times without candidate generation. It also compresses the data set into a data structure called FP-tree. FP-

growth finds all the frequent item sets by searching the FP-tree, recursively [10]. Eclat, BitTableFI and IndexBitTableFI are some examples of vertical based (FIs) mining methods. Eclat uses a structure called Tidset, which store the transaction identifiers for each itemset. The support of an itemset X can be fast derived as the cardinality of the Tidset of the itemset. Thus, the support(X) = |Tidset(X)|. It also proposed the way of computing Tidset(XY) by the intersection operator between Tidset(X) and Tidset(Y). That is, Tidset(XY) = Tidset(X) \cap Tidset(Y) [11]. In BitTableFI [12] each item occupied |T| bits, called a bit vector, where |T| is the number of transactions in D. The bit vector of a new itemset XY from the two itemsets X and Y could be easily derived by the AND operation on the two bit vectors of X and Y. Because the length of the two bit vectors was the same, the result would be a bit vector with the same length of |T| bits. Dong and Han used the BitTable to mine frequent itemsets based on the level-wise concept in the Apriori algorithm [9]. Their approach was named BitTableFI [12]. Note that in the Apriori algorithm, the supports were computed by re-scanning databases, while in the BitTableFI approach, they were derived by the intersection of bit vectors. The support of an itemset could be found by counting the number of ‘1’ bits in its corresponding bit vector. Later, in Song et al. [13], Index-BitTableFI employed index array to improve the algorithm. From the above discussion, the following points can be concluded: vertical association rule algorithms conquer some disadvantages of horizontal ones, vertical association rule algorithms need memory space too much when the dataset is too large. In order to overcome this issue, in this paper, a proposed algorithm that depends on a simple representation of frequent itemsets, which is, compressing the support sets bitmap of data itemsets that to be sent to memory, so as to save the space required by the algorithm. It contributes to reducing not only the execution time but also the required memory. The rest of this paper is organized as follows; Section II briefly revisits some association rule background information. The difference between vertical and horizontal data formats are listed in Section III. Boolean Algebra rules and theories are given in section IV. In Section V, the new algorithm, VBM algorithm is proposed. Section VI analyzes the performance of the proposed algorithm and conclusion is given in Section VII.

II. BASIC CONCEPTION

Association rule mining involves detecting items which tend to occur together in transactions and the association rules that relate them [14].

Consider $I = \{i_1, i_2, \dots, i_m\}$ as a set of items. Let D , the task relevant data, is a set of database transactions where each transaction T is a set of items such that T is a subset of I . Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An Association Rule is an implication of the form $A \Rightarrow B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \emptyset$.

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules.

The rule $A \Rightarrow B$ holds in the transaction set D with support s , where s is the percentage of transactions in D that contain $A \cup B$ (i.e., the union of sets A and B , or say, both A and B). This is taken to be the probability, $P(A \cup B)$. That is,

$$\text{Support}(A \Rightarrow B) = P(A \cup B). \quad (1)$$

The rule $A \Rightarrow B$ has confidence c in the transaction set D , where c is the percentage of transactions in D containing A that also contain B . This is taken to be the conditional probability; $P(B|A)$. That is,

$$\begin{aligned} \text{Confidence}(A \Rightarrow B) &= P(B|A) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)} \\ &= \frac{\text{Support_Count}(A \cup B)}{\text{Support_Count}(A)} \end{aligned} \quad (2)$$

The definition of a frequent pattern relies on the following considerations. A set of items is referred to as an itemset (pattern). An itemset that contains K items is a K -itemset. The set $\{X, Y\}$ is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency or the support count of an itemset [15]. An itemset satisfies minimum support if the occurrence frequency of the itemset is greater than or equal to the minimal support threshold value defined by the user [16]. The number of transaction required for the itemset to satisfy minimum support is therefore referred to as the minimum support count. If an itemset satisfies minimum support, then it is a frequent itemset.

A minimum support threshold and a minimum confidence threshold can be set by users or domain experts. Rules that satisfy both a minimum support threshold (min_support) and minimum confidence threshold (min_confidence) are called strong. The objective of association rule mining is to find rules that satisfy both a minimum support threshold (min_support) and minimum confidence threshold (min_confidence). Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

In general, the problems of association rules can be divided into two sub ones [17, 18]:

- 1) Find out all the frequent itemsets in database D according to the minimum support.
- 2) Generate association rules from frequent itemsets with the limitations of minimal confidence.

Since the second step is much less costly than the first and the overall performance of mining association rules is

determined by the first step, here we are concentrating only on the first step.

III. VERTICAL ASSOCIATION RULES MINING

Horizontal and vertical data formats are two common kinds of data formats to be adopted in frequent itemsets mining. Horizontal structure is the data distribute way by most association rules mining algorithm, its dataset is made up of a series of transactions, each of them includes transaction's ID TID and relevant transaction's inclusive itemsets. However, vertical structure is that the dataset is made of a series of items; each of the items has its TID-list that is the ID list including all transaction of this item [19]. Table I shows transaction database. Fig. 1 and 2 show respectively horizontal and vertical structures for database in Table 1.

TABLE I. A TRANSACTION DATABASE

TID	Item
1	A, C, D
2	A, B, D, E
3	B, C, E
4	A, B, C, D
5	C, D, E
6	A, B, C, D, E

1	A	C	D		
2	A	B	D	E	
3	B	C	E		
4	A	B	C	D	
5	C	D	E		
6	A	B	C	D	E

Fig. 1. Horizontal structure

A	1	2	4	6	
B	2	3	4	6	
C	1	3	4	5	6
D	1	2	4	5	6
E	2	3	5	5	

Fig. 2. Vertical structure

Algorithms for mining frequent itemsets based on the vertical data format are usually more efficient than those based on the horizontal, because the former often scan the database only once and compute the supports of item sets fast [20].

IV. BOOLEAN ALGEBRA

Boolean algebra which was developed by George Boole in 1854, is an algebraic structure defined by a set of elements, B (i.e. B is defined as a set with only two elements 0 and 1 in two

valued Boolean algebra), together with two binary operators, (+) and (\bullet), providing that the following postulates are satisfied[21]:

Note:

- Here is listed only the postulates which are of interest to that work not all Boolean algebra postulates.
- In two valued Boolean algebra, zero and one define the elements of the set B, and variables such as x and y merely represent the elements.

1. (a) The element 0 is an identity with respect to +; that is, $x + 0 = 0 + x = x$.

1. (b) The element 1 is an identity with respect to \bullet ; that is, $x \bullet 1 = 1 \bullet x = x$.

2. (a) The structure is commutative with respect to +; that is, $x + y = y + x$.

2. (b) The structure is commutative with respect to \bullet ; that is, $x \bullet y = y \bullet x$.

3. For every element $x \in B$, there exists an element $x' \in B$ (called the complement of x) such that (a) $x + x' = 1$ and (b) $x \bullet x' = 0$.

Duality principal of Boolean algebra states that: every algebraic expression deducible from the postulates of Boolean algebra remains valid if the operators and identity elements are interchanged, simply interchange OR and AND operators and replace 1's by 0's and vice versa as shown in parts a and b in the above postulates.

Some important theorems that were derived from the above postulates:

- 1: (a) $x + x = x$.
(b) $x \bullet x = x$.
- 2: (a) $x + 1 = 1$.
(b) $x \bullet 0 = 0$.
- 3: involution $(x')' = x$.
- 4: DeMorgan (a) $(x + y)' = x'y'$.
(b) $(x y)' = x' + y'$.

V. VERTICAL BOOLEAN MINING ALGORITHM (VBM)

This section divided into three subsections the first subsection V.A. describes the schema of bitmap used and its compression function. The intersection methods of the compressed vectors are described in V.B. Finally the detailed steps of the algorithm are illustrated in the last subsection V.C.

A. Schema of bitmap used and its compression function

This algorithm is based on vertical data format but instead of representing each item with a bit vector of fixed length equal to the total number of transactions, it uses compression function that works as described below.

The primary goal of the compression function is to make each vector starts and ends with consecutive zeros and then it gets rid of these zero bits to compress the vector.

For each bit vector in the bitmap the compression function examines the start and the end of that vector. Three cases could be found:

Case1: the vector starts and ends with sequence of zeros.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
0	0	0	1	1	1	0	0	1	0	0	0	0	0	0
Flag	rem	Data												
0	11	1	1	1	0	0	1							

Case2: the vector starts and ends with sequence of ones.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
1	1	0	1	0	1	0	0	1	1	1	1	1	1	1
Flag	rem	Data												
1	10	1	0	1	0	1	1							

Case3 (a): the vector starts with sequence of ones and ends with zeros.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
1	1	1	1	1	1	1	0	1	1	1	1	0	0	0
Flag	rem	Data												
1	111	1	0	0	0	0	1	1	1					

Case3 (b): the vector starts with sequence of zeros and ends with ones.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
0	0	0	0	0	0	0	0	0	0	1	0	1	1	1
Flag	rem	Data												
0	1010	1	0	1	1	1								

Fig. 3. Example of transactions before and after compression

In case 1, the compression function does not change the vector form. In case 2, the compression function sets the vector in its complement form to make it starts and ends with sequences of zeros as in case 1. In the third case the compression function counts the number of sequential zeros and the number of sequential ones in the front and tail of the vector. The compression function leaves the vector in its original form if the number of zeros counted is greater than number of ones and puts it in the complement form in the opposite case.

For all three cases given in Fig. 3, the algorithm uses a new data structure for vectors. The vector consists of three elements. The first element, *flag*, Boolean value which indicates either the vector is in the original form (i.e. when *flag*=0) or complement form (i.e. when *flag*=1). The second element, removed (abbreviated as *rem*), binary value representing the number of zeros or ones removed from the beginning of the bit vector.

The third element, *data*, list of bits representing the remaining bits after removing sequences of zero bits or one bits at the front and the tail of the old vector.

B. How to intersect two compressed bit vectors and calculate their support

Fundamental idea: In order to enhance algorithm's operation speed after compressing bitmap, the algorithm makes use of Boolean algebra's rules and postulates to intersect two compressed bit vectors and to calculate their support fast without making any decompression operation for itemsets' bit vectors.

During intersecting two compressed bit vectors according to the schema illustrated above one of the following three cases will be occurred:

Case 1: the two vectors are in the original form (i.e. $flag1=flag2=0$).

Initially, the decimal equivalents of the "rem" values of the two vectors are compared and the larger is used as the "rem" value of the resulting vector. Then AND operations are performed for the *data* parts of the two bit vectors. These operations start at position zero for the vector of larger "rem". The starting point for the AND operation in the other vector is the difference between the decimal equivalents of vectors' "rem" values. If an initial resulting value is 0, then the "rem" value of the outcome vector is increased by 1 until the first non-zero resulting value is reached. Next, from the position of non-zero bit, all the resulting bits by the AND operation are kept in the outcome vector's *data* part except the last consecutive zero bits. Finally the resulting vector's *flag* value is set to zero indicating that the result of intersecting two original bit vectors is a vector in the original form. An example is given below to illustrate the intersection operation on case 1. Assume there are two vectors in the original form: {0, 11, {1, 1, 1, 0, 0, 1}} and {0, 111, {1, 1, 0, 1, 0, 1, 1, 1}} and their intersection is to be found. Both vectors are in the original form because their flags equals to 0. Because the "rem" value (111) of the second vector which is corresponding to 7 in decimal is larger than that (11) of the first that means 3 in decimal system, the AND operation then begins from position (7-3= 4) of first vector and position (0) of the second, at which the result of 0 and 1 is 0. The resulting "rem value" increased by one to be 8. Then, the result of next bits 1 AND 1 is 1 not equals to zero. The rest bits of the second vector are automatically removed because they haven't corresponding bits in the first vector which means that those bits in the first vector were zero bits so that the compression function removed them, and the results are all 0. The resulting vector is then {0, 1000, {1}}.The process is shown in Fig. 4.

Case 2: one vector is in the original form and the other is in the complement form (i.e. $flag1=0, flag2=1$ or vice versa).

The result of intersecting two vectors in different forms as in this case is a vector in the original form so the outcome vector's *flag* value is set to zero. The "rem" value of the resulting vector initially equals to that of the vector in the original form whether it is the larger or not. Then the decimal equivalents of the "rem" values of the two vectors are compared, if the "rem" value of the original vector is the

larger, the AND operations start at position zero for *data* part of the original vector and from position equals to the difference between original vector's "rem" and complement vector's "rem" for *data* part of the complement vector. If an initial resulting value is 0, then the "rem" value of the outcome vector is increased by 1 until the first non-zero resulting value is reached. But if the complement vector's "rem" is the larger, the first complement vector's "rem" value minus original vector's "rem" value bits of the original vector are added to the *data* part of the resulting vectors as they are, because those bits are corresponding to bits of value 1 that were removed from the complement vector in its original form and according to Boolean algebra postulates element 1 is an identity element with respect to AND operation. Next, AND operations are performed between the bits of the original vector and the complement of bits in the complement vector and the resulting bits are kept in the outcome vector's *data* part except the last continuous zero bits. An example is given below to illustrate the intersection operation on this case. Two vectors are given the first is in the original form: {0, 11, {1, 1, 1, 0, 0, 1}} and the second is in the complement form {1, 111, {1, 0, 0, 0, 0, 1, 1, 1}} as shown by their *flags*, and their intersection is to be found. Initially the "rem" value of the resulting vector will be 3 because the "rem" value of the original vector is (11) that means 3 in decimal system. Then the first 4 bits in the original vector will be put in the *data* part of the resulting vector, because those bits in the original vector were actually corresponding to 4 ones in the complement

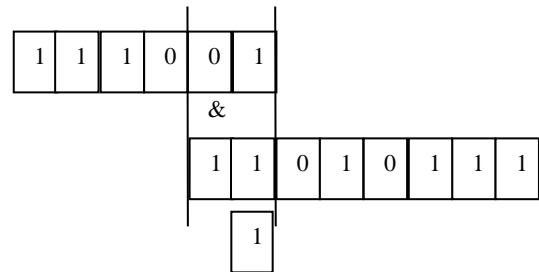


Fig. 4. Intersection Example

vector before compression (i.e. 7-3=4 where 7 is the "rem" of complement vector greater than 3, "rem" value of original one). Next the AND operation starts at bit number 4 in the original vector and position (0) of the complement one, at which the result of 0 and 1 is 0. The resulting position then moves backward to 8. Then, the result of next bits 1 and 1 is 1 not equals to zero. The rest bits of the second vector are automatically removed because they haven't corresponding bits in the first vector which means that those bits in the first vector were zero bits so that the compression function removed them, and the results are all 0. The resulting vector is then {0, 11, {1,1,1,0,0,1}}.The process is shown in Fig. 5 and the opposite is given in Fig. 6 for vectors {0,11,{1,1,1,0,0,1,1,1}}and {1,10,{1,1,1,0,1,0,1}}with result equals{0,101,{1,0,0,0,1,1}}.

Case 3: the two vectors are in the complement form (i.e. $flag1=flag2=1$).

In this case the resulting vector of intersecting two vectors in the complement form is a vector in the complement form, but in some cases this complemented vector may require

transforming to the original form according to some conditions that will be described in the algorithm in section V part c.

Depending on Demorgan theory illustrated in section IV, the algorithm follows steps that are exactly the opposite to those that were followed in case 1. Case 3 intersection steps are illustrated here by an example shown in Fig. 7.

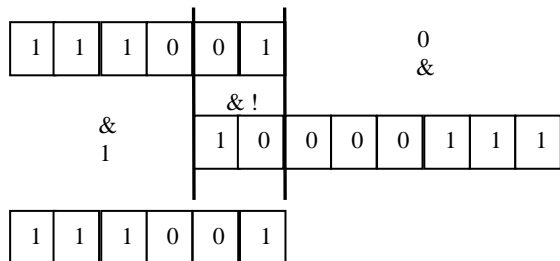


Fig. 5. Intersection Example

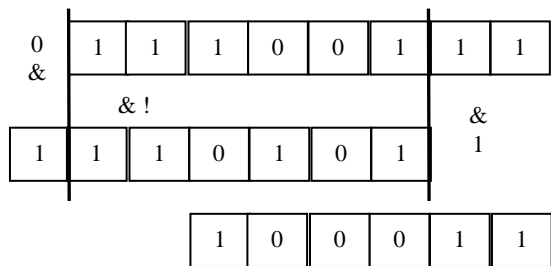


Fig. 6. Intersection Example

Note that: In Fig. 5 and 6 the algorithm automatically puts any bit above sign $\binom{\&}{1}$ in the *data* part of the resulting bit vector and removes any bit below sign $\binom{0}{\&}$ without actually making AND operation to those bits with 0 and 1, depending on Boolean algebra rules illustrated above in order to save execution time.

Fig. 7 shows $\{1,11,\{1,0,1,0,1,0,0,1\}\}$ and $\{1,100,\{1,1,0,0,1,0,1\}\}$ two compressed bit vectors in the complement form assuming that the original length of the bit vectors before compression is 15 bit.

The result is to be obtained by the following steps:

1) The “rem” value of the resulting vector equals to the smallest “rem” value of the two vectors

As $rem1=11 < rem2=100$ (i.e. $3 < 4$ in decimal system), therefore the resulting “rem” value=3.

2) The $(rem2-rem1)$ first bits of the vector with the smallest “rem” value are placed as they are in the data part of the resulting vector according to postulate 1(a) in section IV, because those bits are actually corresponding to zeros in $rem2$.

As $rem2-rem1=4-3=1$ bit therefore first bit only of the first vector is to be put in the first bit of the *data* part of the resulting vector as shown in Fig. 7.

3) Since the intersection operation of two original vectors is accomplished through AND operations, therefore the opposite is done here according to Demorgan theory (i.e. $(x \cdot y)' = x' + y'$) aforementioned in section IV, using An OR

operations between each two corresponding bits till reaching the end of one of the bit vectors as shown in Fig. 7.

4) If the bits of one of the vectors finished before the other, the remaining bits of the longer vector will be placed as they are in the data part of the resulting vector, because those bits are actually corresponding to zeros of the shorter vector as discussed in step 2.

5) Finally the resulting vector is equals to $\{1,11,\{1,1,1,0,1,1,0,1\}\}$.

C. How VBM algorithm works

The main steps of the algorithm can be summarized as follows.

First Step: Scan the database once, obtain a compressed bit vector for each data item by the aforementioned method and set up the result in the structure that were described in section V.A, and calculate support of each data item to produce frequent 1-itemsets and its related compressed bitmap.

Hint: support of items represented by vectors in the original form is calculated by counting the number of set bits in the *data* part of that vector (i.e. number of “1” bits). But support of items represented by complemented vectors is

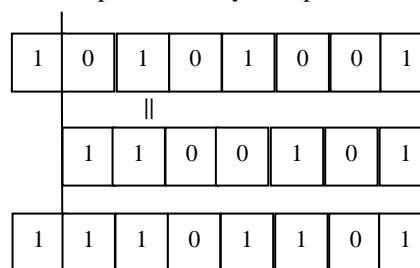


Fig. 7. Intersection Example

equals to the total number of transactions minus the number of set bits in the *data* part of the bit vector.

Second step: In order to get the higher order candidate and frequent $(k+1)$ -itemsets F_{k+1} for each $k > 1$, given a frequent k -itemset F_k , the algorithm uses Depth-first method to join each two itemsets if they have the same first $k-1$ items (excluding just the last item) and the last item of first k -itemset comes before the last item of the second k -itemset in F_k , and applies a modification of the intersection function, which works on three components of the bit vector consisting of the *flag* part, removed value and the *data* part, so that obtaining higher order frequent $(k+1)$ -itemsets does not require rescanning the database again.

Third step: the intersection function first checks *flags* of both bit vectors to be intersected to detect which type of intersection needs to be followed as illustrated in section V.B, in section V.B. cases 1 and 2 are straightforward but case 3 may return vector either in the original or complemented form. Case 3 returns vector in the complemented form if the *rem* values of both vectors aren’t equal to zero and the total of the length of the *data* part of the bit vectors plus the decimal equivalents of their *rem* values (i.e. number of removed zeros) is less than the total number of transactions, because this means that both original vectors were starting and ending with ones so the

result will for sure starts and ends with ones, so the resulting vector should be returned as it is (i.e. in the complement form). Case 3 returns vector in the original form (i.e. case 3 will return the complement of the complemented vector section IV Theorem3 involution) if one of the previous conditions aren't satisfied.

Examples on case 3:

a) intersection method returns complement vector:

the result of intersecting $\{1,11,\{1,0,1,0,1,0,0,1\}\}$ and $\{1,100,\{1,1,0,0,1,0,1\}\}$ equals $\{1,11,\{1,1,1,0,1,1,0,1\}\}$.

b) intersection method returns original vector: the result of intersecting $\{1,0,\{1,1,1,0,1,1,1\}\}$ and $\{1,101,\{1,1,0,0,1,0,1,1,1\}\}$ equals $\{0,11,\{1,0,0,0,1,1,0,1\}\}$.

Forth step: after detecting which type of intersection needs to be followed, the intersection operation is accomplished as illustrated in details in section V.B. to obtain the resulting bit vector. Then support count is calculated for the result. If support count \geq min_support the result is added to frequent k+1 itemset or removed otherwise.

Finally, this process will be continued until there aren't any longer frequent (k+1)-itemsets, then the algorithm ends.

The proposed VBM approach and the schema of bit vectors used consume less time for computing the intersection among compressed bit vectors and for counting the number of 1 bits in the resulting bit vector due to their shorter lengths so the number of bits to be checked is smaller than in the case of classical vertical association rule mining algorithms.

VI. EXPERIMENTAL RESULTS

All experiments were performed on an Intel Core 2 Duo (2x2 GHz), with 3GBs RAM of memory and running Windows vista and algorithms were coded using java programming language. Three real databases used previously in the evaluation of frequent itemsets mining algorithms [22, 23, 24] are used for the experiments, with their characteristics shown in Table II. Due to the huge amounts of the resulting frequent itemsets the method org.apache.commons.io.FileUtils.contentEquals from package commons-io-2.4.jar downloaded from apache library² is used to compare the results of the new algorithm with those of the Apriori algorithm and classical vertical association rule mining algorithm without compressed bitmap, to make sure that the results are correct.

Fig. 8 to 10 show the comparison of the execution time of the VBM algorithm, Apriori algorithm and the classical

vertical association rules algorithm without compressed bitmap, along with different minimum supports. It could be observed that the VBM algorithm was always faster than the other two in all the results.

Next, experiments were conducted to compare between the VBM total memory usage (in MBs) and the vertical association rules algorithm without compressed bitmap. The VBM algorithm compression percentage is also calculated. The results for the three databases under different min_support values are shown in Table III.

From Fig. 8 to 10 we can see that the mining time of VBM algorithm is far from Apriori algorithm but not faraway from the mining time of vertical association rules algorithm without compressed bitmap. But VBM decreased much in memory used by frequent itemsets bitmap than vertical association rules algorithm without compressed bitmap as illustrated in Table III.

Regarding execution time, the non-parametric wilcoxon significance test has been performed to proof the efficiency of the VBM algorithm for the three datasets. The results of the test are given in Table IV. The VBM algorithm showed significant results when compared to Apriori algorithm and the vertical association rules algorithm as p-value < 0.05 in all cases.

The given results show that the strength of the proposed algorithm (VBM) lies in its ability to decrease much in mining time than horizontal association rule mining algorithm and

TABLE II. CHARACTERISTICS OF DATASETS

Dataset	No. of transactions	No. of Items	Average Transaction Length
Chess	3196	75	37
Mushroom	8124	119	23
Connect	67557	129	43

decrease much in memory space than vertical ones. So the proposed algorithm is better than both of them.

As observed from results the reduction in memory and mining time of the proposed algorithm is significantly affected by the content of dataset. The reduction in memory & time cannot be achieved unless the records in the bitmap starts & ends with sequences of zeros and ones as illustrated in section V.A.

¹ <http://fimi.cs.helsinki.fi/data>

² <http://commons.apache.org/proper/commons-io/> [Accessed 19/7/2014]

TABLE III. MEMORY USAGE OF THE VBM AND VERTICAL ASSOCIATION RULE ALGORITHM

DataSet	Minimum Support	No. Frequent Itemsets	Memory Usage of Dataset in Vertical Association Rules Algorithm (MBs)	Memory Usage of Dataset in VBM Algorithm (MBs)	Compression Percentage
Chess	65%	111239	42.9	31.8	25.8%
	70%	48731	18.57	13.92	25%
	75%	20993	8	5.91	26%
	80%	8227	3.13	2.26	27.8%
	85%	2669	1.02	0.74	27%
Mushroom	10%	574431	556	350.28	37%
	20%	53583	51.89	34.14	34.2%
	30%	2735	2.6	1.76	32%
	40%	565	0.54	0.35	35.2%
	50%	153	0.15	0.104	30.5
Connect	86%	105047	845	591.5	30%
	90%	27127	218	156.31	28.3%
	94%	4223	34	24.14	29%
	98%	180	1.44	1.045	27.4%

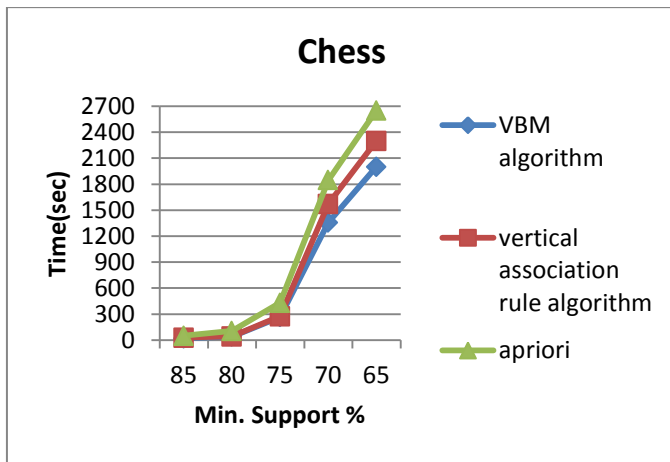


Fig. 8. Execution time of the three algorithms for chess dataset under different min_support values

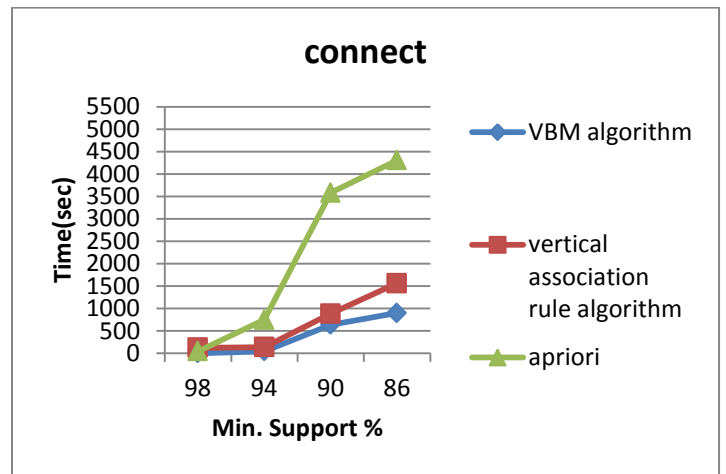


Fig. 9. Execution time of the three algorithms for connect dataset under different min_support values

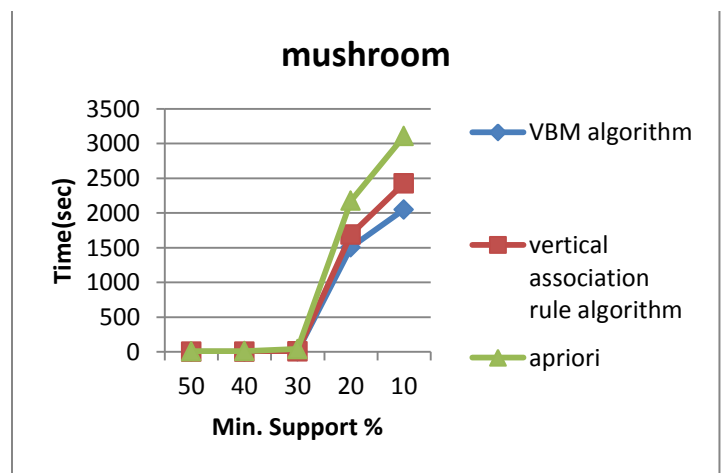


Fig. 10. Execution time of the three algorithms for mushroom dataset under different min_support values

TABLE IV. EXECUTION TIME SIGNIFICANCE TEST

P-value of:	Chess	Mushroom	Connect
VBM vs. Vertical	0.007	0.014	0.002
VBM vs. Apriori	0.001	0.008	0.001

VII. CONCLUSION

This paper proposes a new algorithm that uses a new data structure for compressed bitmap that allows fast computing of support count. So this algorithm relieves the contradiction between vertical association rules algorithm's run speed and memory space to a certain extent. The contributions could be divided into two parts. First contribution is using new data structure to compress bit vector of transaction list representing each frequent item set in only one database scan. Second In order to enhance algorithm's operation speed after bitmap compression, the algorithm makes use of Boolean algebra theories and postulates to perform bit vectors' intersection operation and calculate support count without need to decode the compressed bit- vectors. Therefore, frequent itemsets is generated quickly. The experimental results indicate that the proposed algorithm is much more efficient than Apriori and the classical vertical algorithm for mining association rules in terms of mining time and memory usage. When the database does not contain consecutive bits of zeros and ones at the start and the end of large number of its transactions, the VBM algorithm may suffer the problem of memory scarcity. So solving this memory problem will be the target addressed in one of our future works. We may use transaction partitioning to solve this mentioned problem or search for other techniques.

REFERENCES

- [1] J. Han and M. Kamber, *Data Mining: concepts and Techniques*. San Francisco, CA : Morgan Kaufmann Publishers, 8-131-20535-5, 2010.
- [2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington DC, pp. 207-216. May 1993.
- [3] Y. Tong, L. Chen, Y. Cheng, "Mining frequent itemsets over uncertain databases," Proceeding of the VLDB Endowment, Vol. 5(11), pp.1650-1661, Aug. 2012.
- [4] J. Han and M. Kamber, *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, 1-55860-901-6, 2006.
- [5] M. H. Marghny, R. M. Abd El-Aziz and A. I. Taloba, "An Effective evolutionary clustering algorithm: hepatitis C case study," Computer Science Department, Egypt, International Journal of Computer Applications, vol. 34, No.6, pp. 0975-8887, 2011.
- [6] M. H. Marghny and A. I. Taloba, "Outlier detection using improved genetic K-means," International Journal of Computer Applications, vol. 28, No.11, pp. 33-36, 2011.
- [7] M. H. Marghny, and A. A. Shakour, "Fast, simple and memory efficient algorithm for mining association rules," International Review on Computers & Software, 2007.
- [8] M. H. Margahny and A. A. Shakour, "Scalable algorithm for mining association rules," ICCST, 2006.
- [9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," Proceedings of the 20th International Conference on Very Large Data Bases, Sep. 1994.
- [10] J. Han, J. Pei, Y. Yin, "Mining frequent patterns without candidate generation," in Proceedings of the 2000 ACM SIGMOD international conference on Management of data, ACM Press, pp. 1-12, 2000.
- [11] M.J. Zaki, S. Parthasarathy, M. Ogihara and W. Li, "New algorithms for fast discovery of association rules," 3rd Int. Conf. Knowl. Disc. Data Min. (KDD), pp. 283-286, 1997.
- [12] J. Dong and M. Han, "BitTableFI: An efficient mining frequent itemsets algorithm," Knowl.-Based Syst., Vol. 20(4), pp. 329 – 335, 2007.
- [13] W. Song, B. Yang and Z. Xu, "Index-BitTableFI: An improved algorithm for mining frequent itemsets," Knowl.-Based Syst., Vol. 21(6), pp. 507-513, 2008.
- [14] A. T. Bjorvand, "Object Mining: A Practical application of data mining for the construction and maintenance of software components," Proceedings of the Second European Symposium, PKDD-98, pp. 121-129, 1998.
- [15] A. Tiwari, R. K. Gupta and D. P. Agrawal, "Cluster based partition approach for mining frequent itemsets," Journal of Computer Science, Vol. 9(6), pp. 191-199, 2009.
- [16] M. Houtsma and A. Swami, "Set oriented mining for association rules in relational databases," 11th International conference on Data Engineering, pp. 25-33, 1995.
- [17] T. Y. Lin, Hu. Xiaohua and E. Louie, "A fast association rule algorithm based on bitmap and granular computing fuzzy systems," FUZZ '03, Vol. 12(1), pp. 25-28 May 2003.
- [18] T. Karthikeyan and N. Ravikumar, "A survey on association rule mining," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3(1), pp. 5223-5227 Jan 2014.
- [19] Liu Yang and Mei Qiao, "A Bitmap compression algorithm for vertical association rules mining," 2008 International Symposium on Computer Science and Computational Technology (ISCSCCT). (IEEE), pp. 101-104, 2008.
- [20] V. Bay, H. Tzung-Pei and L. Bac, "Dynamic bit vectors: An efficient approach for mining frequent itemset," Scientific Research & Essays, Vol. 6(25), pp. 5358-5368, 2011.
- [21] M. Morris Mano & M. D. Ciletti : Digital Design, Chapter (2), Pearson Prentice Hall. 0-13-277420-8. pp. 38-43, 2013.
- [22] YU. Xiaomei , H. Wang, "Improvement of Eclat algorithm based on support in frequent itemset mining," journals of computer, Vol. 9(9), pp. 2116-2123, Sep 2014.
- [23] Y. Dejnouri, Y. Geraibia, M. Mehdi, A. Bendjoudi and N. Nouali-Taboudjemat, "An efficient measure for evaluating association rules," Proceeding of the 6th international conference of soft computing and pattern recognition (SoCPaR), IEEE explore, pp. 406-410, Aug. 2014.
- [24] Z. Khan, F. Haseen, S. T. A. Rizvi and M. ShabbirAlam, "Enhanced BitApriori algorithm: an intelligent approach for mining frequent itemset," Proceeding of the 3rd international conference on frontiers of intelligent computing: Theory and Application (FICTA), Vol. 327, pp. 343-350, Springer, 2015.

Investigation of Adherence Degree of Agile Requirements Engineering Practices in Non-Agile Software Development Organizations

Mennatallah H. Ibrahim

Department of Computers and Information Sciences
Institute of Statistical Studies and Research
Cairo University
Cairo, Egypt

Nagy Ramadan Darwish

Department of Computers and Information Sciences
Institute of Statistical Studies and Research
Cairo University
Cairo, Egypt

Abstract—Requirements are critical for the success of software projects. Requirements are practically difficult to produce, as the hardest stage of building a software system is to decide what the system should do. Moreover, requirements errors are expensive to fix in the later phases of the software development life cycle. The rapidly changing business environment is highly challenging traditional Requirements Engineering (RE) practices. Most of the software development organizations are working in such dynamic environment, as a result, either by or without their awareness agile methodologies are adopted in various phases of their software development cycles. The aim of this paper is to investigate the adherence degree of agile RE practices in various software development organizations that are classifying themselves as adopting traditional (i.e. non-agile) software development methodologies. An approach is proposed for achieving this aim and it is applied on five different projects from four different organizations. The result shows that even the non-agile software development organizations are applying agile RE practices by different adherence degrees.

Keywords—*agile methods; agile requirements engineering practices; requirements engineering*

I. INTRODUCTION

Software development in its own is a very complex process and if requirement is not stable and keep changing from the requirement gathering to the development phase, it becomes very difficult to implement [16]. The rapidly changing business environment in which most organizations operate is challenging traditional RE approach. The traditional RE approach focus on gathering all requirements and preparing requirements specification document early before the beginning of the design phase. Software development organizations mostly deal with requirements that are highly volatile (i.e., requirement that tend to evolve quickly and become useless even before project completion), as a result, the early requirements gathering and specification is not suitable as it leaves no room to accommodate changing requirements later in the development life cycle. Furthermore, many other factors make the traditional RE inappropriate for the dynamic context in which software development organizations operate as [9]: (1) rapid changes in competitive threats; (2) stakeholder preferences; (3) development technology; and (4) time-to-market pressures.

Agile methods seek to address the challenges faced by the software development organizations that operate in such dynamic context. Many agile methods advocate the development of code without waiting for formal requirements analysis and design phases, based on constant feedback from the various stakeholders; requirements emerge throughout the development process [10]. In particular, several agile practices deal with requirements in order to implement them correctly and satisfy the needs of the customer [2]. Agile RE practices focus on the continuous interaction between the software application developers and the stakeholders to address the requirements evolution over time, prioritize the requirements, and deliver the most valuable functionalities firstly.

The aim of this paper is to investigate the adherence degree of agile RE practices in traditional software development organizations. The paper answers two questions which are: (1) Do the traditional software organizations apply any of the agile RE practices? ; (2) To what extent agile RE practices are applied in such organizations? A proposed approach has been followed to investigate the adherence degree of agile RE practices in non-agile software development organizations. Such investigation approach has been applied to investigate five projects developed by four different organizations. The result of the investigation is finally concluded and analyzed.

The paper is structured as follows; it is divided into seven sections. Section II introduces the related work. Section III presents an overview on requirement engineering. Section IV introduces an overview on agile methodologies are. Section V introduces the frequently used agile requirements practices. Section VI introduces the proposed investigation approach. Section VII introduces the application of the proposed approach. Section VIII summarizes the main points discussed in the paper. Section IX introduces the future work.

II. RELATED WORK

Several studies and research works are conducted to address the issue of agile RE practices and their challenges. In [2], it is stated that agile RE differs from traditional RE in that agile RE takes an iterative discovery approach. Case studies in [2] were conducted on two types of organizations: (1) organizations that characterize themselves as involved in agile or high-speed software development but didn't explicitly

follow any specific agile methods and (2) organizations that used XP, Scrum, or both explicitly. It was revealed that agile RE practices are adopted in both types of organizations by various adoption levels. The set of adopted agile RE practices are Face-to-Face Communication, Iterative RE, Extreme Prioritization, Constant Planning, Prototyping, Reviews & Test, and Test-Driven Development.

A systematic literature review is conducted in [7] on agile RE practices and their challenges. Such systematic literature review reveals seventeen agile requirements engineering practices which are Face-to-face communication; Customers involvement and interaction; User stories; Iterative requirements; Requirement prioritization; Change management; Cross-functional team; Prototyping; Testing before coding; Requirements modeling; Requirements management; Review meetings and acceptance tests; Code refactoring; Shared conceptualization; Pairing for requirements analysis; Retrospective and continuous planning. Also eight challenges posed by the practice of agile requirements engineering are identified which are: minimal documentation; customer availability; budget and schedule estimation; inappropriate architecture; neglecting non-functional requirements; customer inability; contractual limitations; and requirements change. The research conducted in [6] also reveals the following challenges of agile RE practices: cost and schedule estimation; non-functional requirements; customer access and participation

The research delivered in [1] is concerned with discussing the problem of requirements engineering activities conduction and it suggests some improvements to solve some of the challenges caused by agile requirements engineering practices in large projects. The paper also discusses the requirements traceability problem in agile software development and as well as the relationships between the traceability and refactoring processes and their impact on each other.

The research done in [13] suggests guidelines to improve RE using agile methodologies which are: (1) considering various point of views while eliciting requirements; (2) using various interviewing techniques; (3) considering verification; (4) early consideration of non-functional requirements; (5) adapting requirements management practices; (6) separating environment setup and product development.

III. REQUIREMENTS ENGINEERING (RE)

Requirements are the basis for every software project as requirements. Requirements define what the stakeholders, users, customers, suppliers, developers, and businesses in a potential new system need from the software project and also what the software project must do in order to satisfy all the determined needs [3]. Generally, RE process can be defined as a systematic process of developing requirements through an

iterative co-operative process of analyzing the problem, documenting the resulting observations, and checking the accuracy of the understanding gained [12].

Requirement engineer should work with the following Objectives [16]: (1) Engineer needs to focus on understanding customers and all the stakeholders' desire and their requirement. They should create it and manage it and it will reduce the risk of failure of the software and it should full fill customer's demand; (2) Requirement engineer should give emphasis to know the relevant requirement, remove the conflict and create consensus among the stake holders if any for any requirement. Create unambiguous documentation with given standards and manage requirements systematically.

As shown in figure 1, here are three major activities of RE process which are [12]: (1) Requirements Elicitation; (2) Requirements Documentation/Specification; and (3) Requirements Validation. First, requirements elicitation (also called requirements acquisition) is the activity through which the system's requirements are discovered and elaborated through consultation with stakeholders, from previous documents, and from domain knowledge; the proposed system's boundary is defined during this activity [4].

Second, requirements documentation is the activity that results in producing the output of the RE process which is requirements specification. Generally, there is a wide variety of ways for expressing a requirements specification; such ways are ranging from informal natural language to more formal graphical and mathematical notations [12]. Third, requirements validation is the activity that detects possible problems in the requirements specification before it is being used for software development.

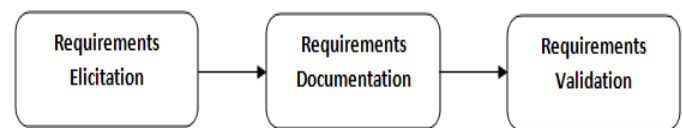


Fig. 1. Major activities of RE process

As mentioned before, requirements are the base of all software projects. However, their elicitation, management, and understanding are common problems for all development methodologies. Particularly, the requirements variability is a major challenge for all of the commercial software projects. According to a study of the Standish Group, five of the eight main factors for software project failure are dealing with requirements which are [2]: incomplete requirements; low customer involvement; unrealistic expectations; changes in the requirements and useless requirements (shown in TABLE I).

TABLE I. MAIN FACTORS OF PROJECT FAILURE [13]

Software Project Problems	
Problem	%
Incomplete requirements	13.1
Low customer involvement	12.4
Lack of resources	10.6
Unrealistic expectation	9.9
Lack of management support	9.3
Change in the requirements	8.7
Lack of planning	8.1
Useless requirements	7.5

IV. AGILE METHODS

Agile Methods are a family of development techniques which are designed to deliver products on time, within budget, and with high customer satisfaction [2]. This family includes several and very different methods. The common agile methods are Extreme Programming (XP) [11] and Scrum [11]. Agile methods embrace iterations where small teams work together with stakeholders to define quick prototypes, proof of concepts, or other visual means to describe the problem to be solved [15].

Agile Methods generally focus on the value of people to solve problems and share information, not on the process and a massive amount of documentation [2]. However, the people-orientation can represent a main weakness for Agile Methods since skills required to build good agile teams are not common [2]. Team members have to be excellent developers who are able to work in teams and have excellent communication skills since the team is self-organizing and cannot refer to a predefined process to share knowledge and solve problems.

The team tasks are to [15]: (1) define the requirements for the iteration; (2) develop the code; (3) create and run integrated test scripts. The users verify the iteration results. Verification occurs early in the development process allowing stakeholders to fine-tune requirements while the requirements still relatively easy to change. Figure 2 shows a generic agile development process features which are an initial planning stage, rapid repeats of the iteration stage, and some form of consolidation before release.

Different agile methods vary in practices and emphasis; but, they follow the same principles behind the agile manifesto which are [1]:

- Working software is delivered frequently (weeks rather than months).

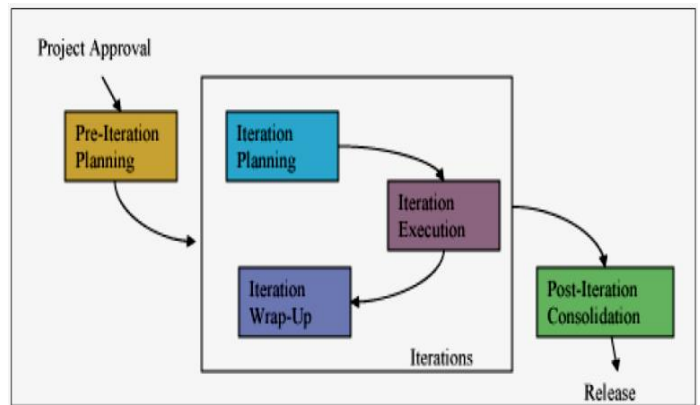


Fig. 2. A Generic Agile Development Process [5]

- Working software is the principal measure of progress.
- Customer satisfaction by rapid, continuous delivery of useful software.
- Even late changes in requirements are welcomed.
- Close daily co-operation between business people and developers.
- Face-to-face conversation is the best form of communication.
- Projects are built around motivated individuals, who should be trusted.
- Continuous attention to technical excellence and good design.
- Simplicity.
- Self-organizing teams.
- Regular adaptation to changing circumstances.

V. AGILE RE PRACTICES

Agile RE processes aren't centralized in one phase before development; they're evenly spread throughout development [10]. Several agile practices deal with requirements in order to implement them correctly and satisfy the needs of the customer [2]. Such practices focus on the continuous interaction between the stakeholders and the development team in order to overcome the problem of requirements volatility, incompleteness and vagueness. The common agile RE practices according to [10] are: Face-to-Face Communication, Iterative RE, Extreme Prioritization, Constant Planning, Prototyping, Reviews & Test, and Test-Driven Development (as shown in Figure 3). Each of such approaches has different aim that helps in performing requirement engineering processes effectively and efficiently.

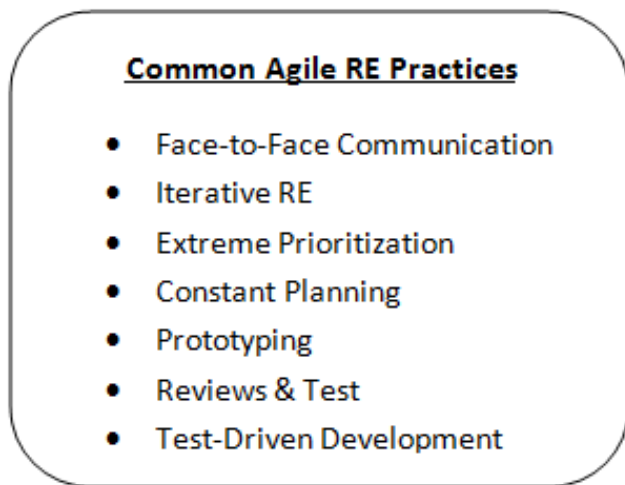


Fig. 3. Common Agile RE Practices

Face-to-Face Communication aims to effectively transfer requirements from the stakeholder to the development team directly without creating extensive documentation. Iterative RE aims to make agile RE continues at each development cycle where at the beginning of each cycle, the stakeholders meet with the development team to provide detailed information on a set of features that must be implemented. During this process, requirements are discussed at a great level of details.

Extreme Prioritization aims to implement the highest priority features at the beginning of the development process so that customers can gain the most business value early. The stakeholders prioritize their feature lists repeatedly during development life cycle as the stakeholder's and the developer's understanding of the project evolves, particularly when requirements are added or modified. Constant planning aims to accommodate requirements changes during the project development so that the system can be tuned easily to better satisfy customer needs. There are commonly two types of requirements changes [10]: (1) adding or dropping features; and (2) changing already implemented features. Generally, changes are easier to implement and cost less in agile development.

Prototyping aims to produces software application in the form of operational prototype, a refinement of the code created for experimentation with required features. This helps the organizations to rush to market as many of these organizations deploy these prototypes rather than wait for robust implementations. The ability to quickly deploy newer versions of the products on the Internet also contributes to this tendency.

Test-driven development is an evolutionary practice in which developers create tests before writing new functional code; such approach treats writing tests as part of a

requirements/design activity in which a test specifies the code's behavior [10]. Such practice aims to help developers team to write an explicit requirements specification. Review & Tests aim to take the advantage of frequent review meetings to validate requirements. At the end of each development cycle, a meeting is held between the developers, the stakeholder, quality assurance personnel and management personnel to validate the specified requirements.

The agile RE practices are usually used in combination with each other within the single organization. Agile RE practices provide benefits of improved understanding of customer needs and the ability to adapt to dynamic environment in which software development organizations operate. However, they pose several challenges to their adopting organizations. Therefore, such organizations should carefully compare the costs and benefits of agile RE practices in their projects.

VI. PROPOSED INVESTIGATION APPROACH

An approach for investigating the adherence degree of agile RE practices in non-agile software development organizations is proposed (shown in Figure 4). Such approach is consisting of seven steps. First step is to identify the list of agile RE practices that their adherence degree will be investigated. Second step is to design a questionnaire by formalizing the agile RE practices in form of questions in order to fit the purpose of our study; the simplicity of the designed questionnaire is considered. The questionnaire is used as a tool for collecting information from various participants (i.e., project managers and development team). We then checked the questionnaire against each of the identified agile RE practice to ensure that all of those practices are addressed within it.

Third step is to select the projects that will be subjected to the study where the selected projects are in different fields and by different organizations. Semi-structured interviews are held with various participants (i.e., project managers and one or more of the development team) in each project as a fourth step. If required, requirement documents of projects are reviewed to get more information about the projects requirements and sometimes to be an evidence of the information gathered from the participants and this is the fifth step.

In the sixth step, the information gathered from all participants in each project is analyzed to identify the adherence degree of each agile RE practice in every single project separately. In the seventh step, all the results (i.e., adherence degree) of each agile RE practices in all project are then aggregated in one table. The mean of all adherence degrees of an agile RE practice in all projects is calculated and then the agile RE practices are ranked according to the calculated mean.

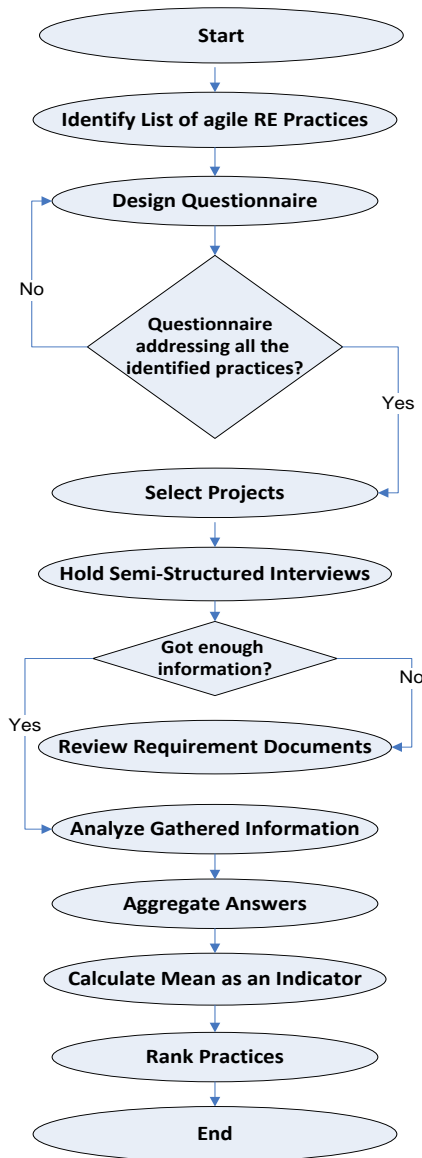


Fig. 4. Proposed Investigation Approach

VII. APPLYING INVESTIGATION APPROACH

The aim of applying the proposed approach is to measure the adherence degree (i.e. applying degree) of a set of common agile RE practices in software applications development. The list of examined agile RE practices are Face-to-Face Communication, Iterative RE, Extreme Prioritization, Constant Planning, Prototyping, Reviews & Test, and Test-Driven Development. Cases studies are applied in four different organizations that employ traditional approaches in developing their software projects (i.e., such companies is not adopting agile methods in neither their working procedures generally nor in their RE processes specifically). However, the results of our investigation show that they are actually applying agile RE practices without their awareness. These organizations are

located in Egypt's capital (i.e., Cairo). The required data is collected from the organizations through semi-structured interviews, questionnaires, in addition to reviewing requirements documents.

The first investigated project is a point of Sale banking application developed by an organization called "SEE Egypt". The investigation's result of this project shows that all of the agile RE practices are applied except Test-Driven development practice. All agile RE practices are either strongly or moderately applied. The strongly applied practices are Iterative RE, Extreme Prioritization, Constant Planning, and Reviews & Tests are, while Face-to-Face Communication and Prototyping are moderately applied practices.

The second investigated project is an application to display and deal with satellite receiver channels developed by an organization called "IG Company". The investigation's result shows that in this project all of the agile RE practices are applied except Test-Driven development and Constant Planning. All the practices are strongly applied.

The third investigated project is a website for faculty of computers and information, Cairo University called "FCI E-Community Website". This project developed by "Centre for the Study of Developing Societies – Cairo University". The investigation's result shows that all the agile RE practices are applied by different degrees. Face-to-Face Communication, Iterative RE, and Extreme Prioritization are applied strongly. Constant Planning, Prototyping and Reviews & Test are applied moderately. Test-Driven Development is weakly applied.

The fourth investigated project is an application to manage and control the employees' data developed by an organization called "Triple L Oil Service". The investigation's result shows that this organization is applying all the agile RE practices by different degrees. Face-to-Face Communication, Iterative RE, Extreme Prioritization, Reviews & Test and Prototyping are strongly applied, while Constant Planning and Test-Driven Development are moderately applied.

The fifth investigated project is an attendance system for Canadian university in Cairo which is developed by "Centre for the Study of Developing Societies – Cairo University" also.

The investigation's result shows that this all the agile RE practices are applied by different degrees. Face-to-Face Communication, Iterative RE, Extreme Prioritization, Constant Planning, Reviews & Test and Prototyping are strongly applied. Test-Driven Development is moderately applied. Prototyping is weakly applied.

By summarizing and analyzing the previous results (as shown in figure 5), it will be clear that both Iterative RE and Extreme Prioritization are the most applied agile RE practices (i.e., strongly applied) in the four organizations, while Test-Driven Development is the least applied practice (i.e., weakly applied). Face-to-Face Communication and Reviews & Test are considered to be strongly applied. Constant Planning and Prototyping are moderately applied. Figure 6 shows the detailed ranking of the applied agile RE practices.

Practice	Project 1	Project 2	Project 3	Project 4	Project 5	Adherence Degree
Face-to-Face Communication	2	3	3	3	3	2.8
Iterative RE	3	3	3	3	3	3
Extreme Prioritization	3	3	3	3	3	3
Constant Planning	3	0	2	2	3	2
Prototyping	2	3	2	3	1	2.2
Test-Driven Development	0	0	1	2	2	1
Reviews & Tests	3	3	2	3	3	2.8

0: Not Applied; 1: Weakly Applied; 2: Moderately Applied 3: Strongly Applied

Fig. 5. Results Summarization

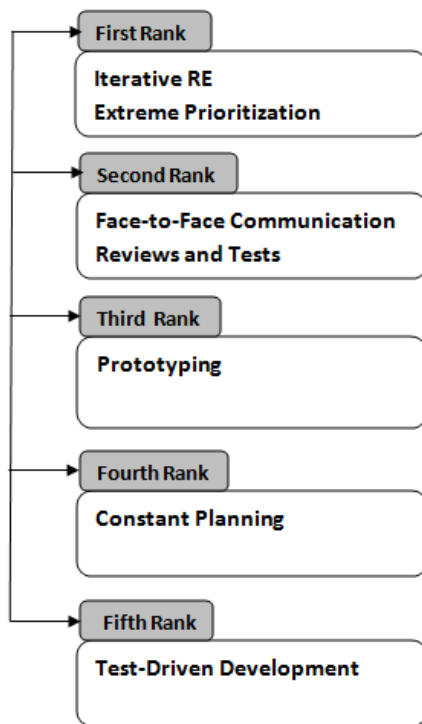


Fig. 6. Detailed Ranking of Applied Agile RE Practices

The results of the conducted case studies show that all agile RE practices are applied in the investigated projects, although, such projects are developed in organizations that are not applying agile methods in their applications development processes. The agile RE practices are applied by different degrees in each project. The most applied approaches are Iterative RE and Extreme prioritization, while, the least applied approach is Test-Driven Development. The results prove that the dynamic working environment and the unstable circumstances in which most of the organizations operate are challenging the traditional RE and enforcing such organizations to apply agile RE practices in their software development lifecycle even without their awareness.

VIII. CONCLUSION

Requirements are critical for the whole development cycle of software systems. Requirements generally define the tasks that the system should perform and the constraints posed on those tasks. Many factors cause the traditional RE to be inappropriate for software development as rapid changes in competitive threats; stakeholder preferences; development technology; and time-to-market pressures. Agile methodologies seek to address the challenges posed by dynamic environment in which most of the software development organizations operate. Agile methodologies focus on the continuous interaction between both the stakeholders and the development team.

The results of applying the proposed investigation approach show that although the investigated organizations consider themselves not applying agile methods in developing their applications, they are actually applying agile RE practices by different degrees without their awareness. The most applied practices are Iterative RE and Extreme prioritization. The least applied practice is Test-Driven Development.

IX. FUTURE WORK

There are many efforts can be done in the field of agile RE practices in the future. Briefly, the following points are expected to be focused:

- Extending the study to cover more software projects in many domains.
- Proposing an approach for evaluating the quality of applying agile RE practices using metrics.
- Evaluating the quality of applying hybrid agile methods to reveal the most used methods in conjunction.
- Using fuzzy logic in the evaluation process.

REFERENCES

- [1] A. D. Lucia, and O. Abdallah, "Requirements engineering in agile software development," *Journal of Emerging Technologies in Web Intelligence*, vol. 3, pp.212-220, 2010.
- [2] A. Sillitti, and G. Succi, "Requirements engineering for agile methods," *Engineering and Managing Software Requirements*, pp. 309-326. Berlin Heidelberg: Springer, 2005.
- [3] E. Hull, K. Jackson, and J. Dick, "Requirements engineering," London: Springer, 2005.
- [4] E. Nasr, J. McDermid and G. Bernat, "Eliciting and Specifying Requirements with Use Cases for Embedded Systems," *International Workshop on Object-Oriented Real-Time Dependable Systems*, IEEE 2002.
- [5] F. Paetsch, E. Armin, and F. Maurer, "Requirements engineering and agile software development," In *2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises*, pp. 308-308. IEEE Computer Society, 2003.
- [6] H. Elshandidy, and S. Mazen, "Agile and Traditional Requirements Engineering: A Survey," *International Journal of Scientific & Engineering Research*, vol. 9, 2013.
- [7] I. Inayat, S. S. Salim, S. Marczak, M. Daneva, and S. Shamshirband, "A systematic literature review on agile requirements engineering practices and challenges," *Computers in Human Behavior*, 2014.
- [8] K. Beck, "Extreme programming explained: Embrace change." Addison-Wesley, UK., 1999.
- [9] K. Schwaber, M. Beedle, "Agile software development with scrum," Prentice Hall PTR, Australia, 2001.

- [10] L. Cao, and R. Balasubramaniam, "Agile requirements engineering practices: An empirical study," *Software, IEEE*, vol. 25, pp. 60-67, January 2008.
- [11] P. Abrahamsson, O. Salo, J. Ronkainen, and J. Warsta, "Agile software development methods: Review and analysis," 2002.
- [12] P. Loucopoulos, and K. Vassilios, "System requirements engineering," McGraw-Hill, Inc., 1995.
- [13] S. Bose, M. Kurhekar, and J. Ghoshal, "Agile Methodology in Requirements Engineering," SETLabs Briefings Online, 2008.
- [14] V. E. Jyothi, and K. N. Rao, "Effective Implementation of Agile Practices - Ingenious and Organized Theoretical Framework", *International Journal of Advanced Computer Science and Applications*, vol. 2, March 2011.
- [15] V. Szalvay, "An introduction to agile software development," *Danube Technologies*, pp. 1-9, 2004.
- [16] V. Tripathi, and A. K. Goyal, "Agile Requirement Engineer: Roles and Responsibilities," *International Journal of Innovative Science, Engineering & Technology*, vol. 3, pp. 213-219, 2014.

A Review on Parameters Identification Methods for Asynchronous Motor

Xing Zhan, Guohui Zeng, Jin Liu, Qingzhen Wang, Sheng Ou

College of Electronic and Electrical Engineering
Shanghai university of engineering science
Shanghai, China

Abstract—The decoupling of excitation current and torque current is realized by Vector control so that the speed regulating performance of asynchronous motor is comparable with that of dc motor. The control precision is directly affected by accuracy of parameter identification in asynchronous motor. In this paper, based on the existing literatures, the existed parameters identification methods both online and offline are analyzed and compared, and the advantages and disadvantages of the various algorithms are listed in tables. Therefore, a comprehensive identification method of adjustable model which makes the least square method as adaptive method of model reference is presented. Finally, the outlook of developing direction for parameters identification of asynchronous motor are put forward to.

Keywords—Vector control; Parameters identification; MRAS; the least square method

I. INTRODUCTION

Since the vector control was presented by Felix Blaschke, the ac speed regulating performance of asynchronous motor is comparable with that of dc motor. Vector control which is through the mathematical formula and the matrix transformation to make decoupling between the excitation current and torque current of asynchronous motor [1], and the control performance of asynchronous motor is enhanced greatly by a similar dc motor control method. Now, the parameters identification methods of asynchronous motor are mainly off-line identification and online identification. Because of the basic parameter of vector control can be provided and the operation is simple, the off-line identification is used most. But in the process of the motor operation, the stator resistance (R_s) and rotor resistance (R_r), and the time constant of Motor rotor ($T_r = L_r / R_r$) are influenced by the change of environment, such as the change of temperature, air humidity, high pressure, dust, and so on. The change of the time constant of rotor is affected by the change of resistance [2], which leads to magnetic field orientation is not accurate, and a better decoupling will not be produced between the excitation current and torque current of asynchronous motor. Finally, inaccuracy and deflection are emerged, and industrial production is also affected.

Off-line identification techniques [3]: (a) on one hand, the motor speed can be made to close to the synchronous speed by

the traditional no-load experiment; On the other hand, the motor speed can be made to be zero by locked-rotor experiment. (b) The motor parameters can be identified through the data of motor structure. (c) Different voltage are injected to motor based on a converter, the motor parameters can be identified by the motor to inspire different voltage. (d) A mathematical program is provided by least-squares, and a fitting curve is gain, which is about the fitting of minimum variance sense and an experimental data of completely measuring. Then the result of identification is obtained.

The excitation component and torque component of stator current are decoupled by formula (1) and formula (2). The ac speed regulating performance of asynchronous motor is compared with dc motor speed control. asynchronous motor's operation, motor parameters are easily influenced by environment. But the off-line identification cannot solve this problem absolutely. In order to control the precision of motor, the online identification of asynchronous motor is required.

At present, there exist 4 types of online identification techniques: (a) recursive least squares [4-5]. The estimated value of objective function is corrected continuously, and the parameter is estimated step by step until the satisfied parameter value is gained. (b) The extended Kalman filter [6] is recursive estimation method [7]. The estimated value of current state is calculated through the estimated value at the state of a moment before and the observed value of current, and this method is used in linear stochastic systems [8-9]. (c) MRAS is used to identify the motor parameter, and a suitable adaptive law is found [10-11]. The output's error between the reference models without identified parameters and the adjustable models with Different voltage are injected to motor based on a converter is a more reliable method than other off-line identification method. This method has many advantages, such higher recognition efficiency, higher precision, more convenient, and so on. The electronic resistance (R_s), leakage inductance of stator and rotor (σ), rotor resistance (R_r), and the mutual inductance (L_m) could be identified through this method. The time constant of rotor ($T_r = L_r / R_r$) can be deduced by the above identified parameters. Then, the stator current of motor is decomposed into torque component (i_{τ}) and excitation component (i_{ψ}) via the mathematical model of vector control and coordinate transformation.

This work is supported by the Natural Science Foundation of Shanghai under Grant No.14ZR1418400 and the Innovation Foundation of Shanghai Education Commission under Grant No.13YZ111 and the Innovation Foundation of SUES under Grant No.E1-0903-14-01041

$$T_e = \frac{p_n L_m}{L_r} i_{st} \psi_r \quad (1)$$

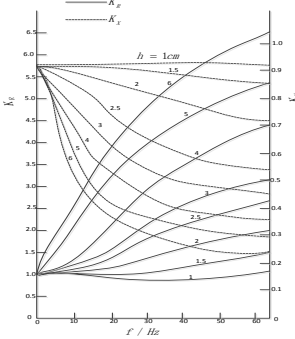
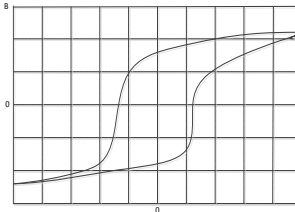
$$\psi_r = \frac{L_m}{T_r P + 1} i_{sm} \quad (2)$$

identified parameters tends to zero finally^[12]. Then the motor's parameters are identified^[13]. (d) The algorithms of artificial and intelligent identification is used to simulate natural biological systems, and totally dependent on its instinct to optimize the existence to adapt to the environment^[14].

II. THE FACTORS OF PARAMETERS

The parameters are in change for moment in the mathematical model of asynchronous motor and effected easily by factors .with changes of environmental factors and asynchronous motor, the parameters is changing. The parameters mainly contain stator resistance (R_s), rotor resistance (R_r), stator inductance (L_s), rotor inductance (L_r), and the mutual inductance between stator and rotor (L_m). The main factors of the change of the asynchronous motor running parameters are shown in table 1:

TABLE I. THE MAIN FACTORS OF THE CHANGE OF THE ASYNCHRONOUS MOTOR RUNNING PARAMETERS

Factors	Influence mechanism	Parameters are influenced	Change law
Temperature changes	The energy change into heat energy in the process of electrical energy transform into mechanical energy, the change of external environment, motor's aging degree and the degree of wear and tear.	Stator resistance of motor (R_s), rotor resistance (R_r)	$R_a = \frac{235 + t_a}{235 + t_b} R_b$ <p>(R_a) is the resistance in the temperature (t_a), (R_b) is the resistance in the temperature (t_b)</p>
Frequency changes	The skin effect is caused by frequency changes of current, which is related to the rotor's slot type of asynchronous motor.	Rotor resistance (R_r), rotor inductance (L_r)	 <p>Fig. 1. The curve is about the relationship between rotor resistance and inductance and frequency</p>
The factors of the saturated magnetic	Asynchronous motor is in the linear part of the B - H curve, while the iron core reluctance is smaller. When the magnetic is saturated, rotor resistance increases, and the inductance decreases ^[10] .	Rotor resistance (R_r), rotor inductance (L_r), the coefficient of leakage inductance (σ)	 <p>Fig. 2. B-H curve</p>
The stray loss	Eddy current loss is produced by magnetic-flux leakage, Eddy current and hysteresis are produced when Winding in the metal structure, higher-order Harmonic losses ^[9] .	stator inductance (L_s), rotor inductance (L_r), mutual inductance (L_m)	

III. THE IDENTIFICATION METHOD OF PARAMETER

A. The off-line identification method

Currently, the off-line identification method is studied in the world and advantages and disadvantages of each method are shown in table 2.

TABLE II. ADVANTAGES AND DISADVANTAGES OF THE OFF-LINE IDENTIFICATION METHOD

Identific ation method	Identification Principle	Identificati on parameters	Advantages	Disadvantages
<i>Locked-rotor method of motor</i>	Rotor winding of asynchronous motor is cut out and rotor is stuck to rotate unusually. Then parameters are calculated through the circuit principle. The equivalent circuit is shown in figure 3.	Resistance, inductance, mutual inductance.	The parameters of resistance, inductance, and mutual inductance can be calculated.	In many conditions, not only this kind of experimental conditions cannot be gotten, but also the load of system is very uneasy to be installed and removed. And the skin effect of rotor is very serious ^[11] .
<i>No-load experiment method</i>	Motor does not drag any load, and the rotor speed of motor is almost equivalent to the synchronous speed. Equivalent circuit is shown in figure 4.	Excitation reactance and the excitation resistance	Convenient is realized. In the case of motor's control precision is not high, the required parameters are measured.	The rotor circuit is ignored, and the identification precision is not high. In many occasions, the load is carried by motor. It is not convenient to no-load removed load.
<i>Auto-tuning method of motor parameters</i>	Different test signal is injected into motor to make the motor in different state. The features of frequency converter are used to perform some procedures. Finally, the purpose of identifying motor's parameters is achieved.	Stator resistance (R_s), leakage inductance (σ) rotor resistance (R_r).	Precision is higher and good reliability.	This method is suitable for being used in a control system with frequency converter, and it is not convenient to be used in the control system with no frequency converter.
<i>The least square method</i>	The experimental data in the whole stage is sampled through a complete measurement. Then the least squares curve is offline calculated and fitted, and the identified results are obtained.	Rotor resistance and inductance.	High precision.	A large amount of data and the experimental data in the whole stage are need to be measured, and measurement and t calculation are tedious.

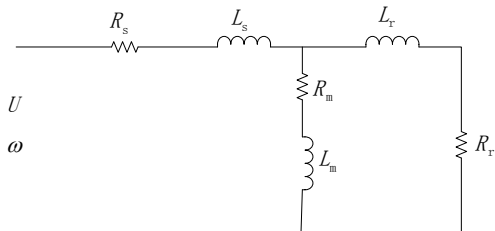


Fig. 3. The equivalent figure of motor cutting-out

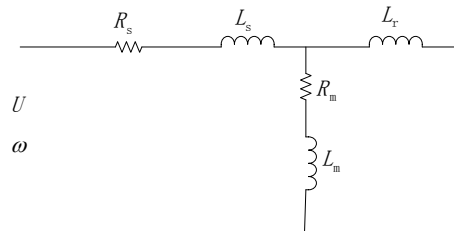


Fig. 4. The equivalent figure of motor no-load

B. Online identification method

The above off-line identification methods are the identification that in the case of motor no-load or in a moment the motor is changing with temperature and environment in the operation process. In a high-precision control system, the off-

line identification cannot meet the requirements. Thus, the parameters are required to be identified online and identified at every moment. Online identification method is shown in table 3.

TABLE III. THE ONLINE IDENTIFICATION METHOD

identification method	Identification principle	Identification parameters	Advantages	Disadvantages
<i>The extended Kalman filter^[7]</i>	The state equations of motor is used directly. The parameters which need to be identified and states are set as quantities of augmented state. The optimal estimation of the state is given solving riccati equations in online ^[8] .	Rotor resistance, mutual inductance, flux linkage	With high identified precision and more parameters. And other unpredictable state can be estimated. Itself has filtering effect and can reduce some interference ^[9] .	The algorithm is complex and good performance of processor is required
<i>MRAS</i>	The actual running of motor is used as the reference model, and the state observation equation of motor is used as the adjustable model. Motor's parameters of the adjustable model are corrected in real time by some measurable deviation	Rotor resistance, stator resistance, mutual inductance.	A small amount of calculation and high precision.	Reference model is difficult to be determined
<i>The improved least square method</i>	The experimental data in the whole stage is sampled through a complete measurement. Then, the least squares curve is calculated offline and fitted, and the identification results are obtained.	Resistance, inductance, time constant of rotor.	High precision and good reliability. Both online and offline identification.	It is Sensitive to noise of measurement and fluctuation of speed. The estimated value exists multi solutions and deviation problems owing to the singularity of structure matrix.
<i>Intelligent algorithm[15]</i>	Neural network and genetic algorithm are included in this method ^[16] . The former makes the function value of error to be minimized by learning system's Input and output. The latter is a kind of random search algorithm, which can simulates the natural evolution ^[17] .	Resistance and inductance	The precision is very high.	The calculation is so huge and the requirement for processor is relatively high.

IV. COMPREHENSIVE IDENTIFICATION SYSTEM

According to the analysis and study of existing literature, the single identification method has defects more or less.

A kind of method that both off-line identification and on-line identification are put forward by author based on existing literatures. The method is combined with model reference adaptive and improved least square method. The adjustable model is constructed by the improved least square method, and the least error is made to be minimized between the reference model and the adjustable model.

A. The principle of model reference adaptive

THE PRINCIPLE OF MODEL REFERENCE ADAPTIVE IS SHOWN IN FIGURE 5.

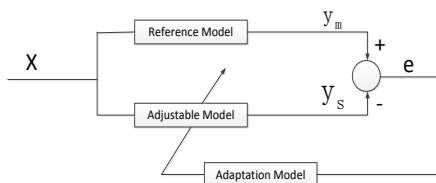


Fig. 5. Principle of model reference adaptive

This system has the same external input(X), while the (X) is input to the reference model and the adjustable model. (y_r) and (y_s) are the output of the reference model and the adjustable model. Finally, the result of $e = |y_r - y_s|$ is calculated to be minimized or zero through the adaptive adjustment. The design of adaptive control mainly includes:

1) *The optimization theory of partial parameters. The structured distance between the reference model and the adjustable model is defined, or two times performance index of state distance. The method of parameter optimization is used to determine the adjustment of parameters of the controller. Rules make the adjustable model closer to reference model to achieve the purpose of identifying parameters.*

2) *Popov super stability theory. Firstly, the model reference adaptive system should be transformed into equivalent nonlinear time-varying feedback system. Namely, the system is formed by a linear forward link and a nonlinear feedback link. A suitable law of adaptive control is gotten*

under the guarantee of meeting the two conditions. This makes the whole nonlinear system is stable, which can ensure the system error tends to zero and achieve the purpose of adaptive control.

B. B. The least square method

A linear relationship exists between the variables (y) and one dimensional variable $x = (x_1, x_2, x_3 \dots x_n)$, namely:

$$y = C_1x_1 + C_2x_2 + C_3x_3 + \dots + C_nx_n$$

The value (C) is estimated based on the observed values (y) and (x) in different time. This method requires repeated computation under the need of update data, this problem is solved by recursive least squares method. A new set of data don't need to be added and calculated again, and the amount of calculation is reduced greatly. The basic idea of recursive least square method is: the new estimate value = the old estimate value + correction term. So that $y = \theta_i x$ and $e = y_m - \theta_i x$ is its minimum.

V. CONCLUSION

Through analyzing and concluding the existing literatures, Conclusions can be gotten as follows:

1) As shown by the table 1, the rotor resistance of asynchronous motor is changed mostly by the environmental factors. As shown by the formula (1) and formula (2), the rotor flux linkage is changed directly by the change of rotor resistance. Then, the decoupling of rotor flux linkage is not sufficient. If a certain control accuracy is required by a control system, a control method which main identifies rotor resistance can be chosen.

2) Offline identification method has many kinds of classes. Being compared with the locked-rotor method of motor no-load experiment method and the least square method, auto-tuning method of motor parameters is used most widely, the technology is most mature, and the identification accuracy is more accurate and reliable.

3) The system which requires high precision of identification, low cost and reliable performance and easy to implement, the least squares method is a better method, because this method can not only realize the off-line identification but also realize the online identification.

4) As to the control system which very high accuracy of identification, intelligent algorithm can be used as the focus of future research methods. This method has the very good control precision.

With the rapid development of the economic in China, the motor's control technology with high precision is widely used in all kinds of industrial fields. The higher requirements for the identification accuracy of asynchronous motor is put forward. At present, China is still relatively backward in terms of manufacturing and control of asynchronous motor. Especially the gap from Western countries in processor manufacturing is large. How to make the motor control more accurate, which

requires higher parameter identification. Although the general industrial production requirements can be satisfied by the identification method of existing parameters, it is not enough in some high precision control system. In genetic algorithm, genetic algorithm and neural network algorithm are continuous optimization and find out the optimal solution, which needs higher requirements for the processor to be put forward. Therefore, the intelligent algorithm has a broad prospect in the future and it can be developed from two aspects. On one hand, the intelligent algorithm could be improved^[18-19]. The algorithm is simple and high precision of parameter identification could be achieved. On the other hand, the performance of processor could be improved so that the processor speed is faster and cheaper.

REFERENCES

- [1] Boshi Chen, Minxun Chen. Ac speed regulating system [M]. Beijing: Mechanical industry press, 2013
- [2] Pian Zhou, Shuyun Wang, Lijiu Wang. The influence of rotor resistance variation of vector control [J]. Electrical automation.1998 (3):18-20
- [3] Mingyu Wang, Chengyu Xian, Yaqian Hui. Induction motor vector control parameters off-line identification technology [J]. Transaction of China electrotechnical society, 2006(21):90-96
- [4] Chao Cai, Guangdong Chen. A study on parameter estimation of induction motor using least square method [J]. Hubei: journal of Wuhan institute of chemical technology, 2003, 25 (2).
- [5] Ruiming Fang, Hongguang Ma. Classification of Induction Machine Rotor Faults Based on Least Square Support Vector Machine [J]. Transaction of China electrotechnical society, 2006, 21(5).
- [6] Xiwei Zhou. Asynchronous motor parameter identification based on EKF [D]. Xian: Xian University of science and technology, 2003
- [7] Zhongbo Peng, XueFeng Han, Zixue Du. Direct Torque Control for Electric Vehicle driver Motor Based on Extended Kalman Filter.[J] Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd.
- [8] A. Lalami, R. Wamkeue, I. Kamwa, M. Saad, J.J. Beaudoin. Unscented Kalman filter for non-linear estimation of induction machine parameters, [J] IET Electric Power Applications, Received on 31st January 2012.
- [9] linBaek Kim, ByungKook Kim. Accurate States Estimation using Asynchronous Kalman Filter with Encoder Edges for TMRs,[J] International Conference on Control, Automation and Systems, Oct. 20-23, 2013 in Kimdaejung Convention Center, Gwangju, Korea.
- [10] Yin Yao, Lu Zhen. Based on model reference adaptive research of asynchronous motor vector control system [D]. Liaoning: Liaoning technical university, 2006
- [11] V.Verma, M. J. Hossain, Member, IEEE, T. Saha, Senior Member, IEEE, C. Chakraborty, Senior Member, IEEE. Performance of MRAS Based Speed Estimators for Grid Connected Doubly fed Induction Machines during Voltage Dips, [J] Power and Energy Society General Meeting, 2012 IEEE.
- [12] F. L. Mapelli, A. Bezzolato, D. Tarsitano. A Rotor Resistance MRAS Estimator for Induction Motor Traction Drive for Electrical Vehicles, [J] Electrical Machines (ICEM), 2012 XXth International Conference on, 2-5 Sept. 2012.
- [13] Ahmed S. Morsy and A. S. Abdel-khalik, Shehab Ahmed, Ahmed Massoud. Sensorless V/f Control with MRAS Speed Estimator for A Five-Phase Induction Machine under Open-Circuit Phase Faults, [J] 2013 IEEE GCC Conference and exhibition, November 17-20, Doha, Qatar.
- [14] Tan Ma, Jin Zhao. The research on intelligent control of AC drive system based on neural network [D]. Huazhong University of science & technology, 2009
- [15] Zhang Dongdong, Luo Wenguang, Chen Wenhui, Xie Rongxian. Energy-saving Control Based on Neural Network Inverse Decoupling for Asynchronous Motors, [J] Power and Energy Engineering Conference (APPEEC), 2010 Asia-Pacific.

- [16] Pham Thuong Cat, Le Hung Linh, and Minhtuan Pham. Speed Control of 3-Phase Asynchronous Motor Using Artificial Neural Network, [J] Control and Automation (ICCA), 2010 8th IEEE International Conference on.
- [17] Mu Li, Peng Liu, Jiaoming Liu. Design of stray loss test system for three-phase asynchronous motor [J]. Motor and control applications, 2008, 35(8).
- [18] Dongqi Zhu, Dewei Xu, Jianxin Jiang. Torque optimization analysis with nonlinear models for main path saturation of induction machines [J] Journal of Tsinghua university, 2001, 9 (9).
- [19] Xinzhen Wu, Yanxiang Wang. Calculation of skin effect for double-gage rotor bar of the induction machine [J].proceeding of CSEE, 2003, 23(3).

An Intelligent Natural Language Conversational System for Academic Advising

Edward M. Latorre-Navarro
Computer Science Department
University of Puerto Rico in Arecibo
Arecibo, Puerto Rico

John G. Harris
Electrical and Computer Engineering Department
University of Florida
Gainesville, Florida, U.S.A.

Abstract—Academic advisors assist students in academic, professional, social and personal matters. Successful advising increases student retention, improves graduation rates and helps students meet educational goals. This work presents an advising system that assists advisors in multiple tasks using natural language. This system features a conversational agent as the user interface, an academic advising knowledge base with a method to allow the users to contribute to it, an expert system for academic planning, and a web design structure for the implementation platform. The system is operational for several hundred students from a university department. The system performed well, obtaining close to 80%, on the traditional language processing measures of precision, recall, accuracy and F1 score. Assessment from the constituencies showed positive and assuring reviews. This work provides an assessment and technological solution to the academic advising field, i.e., the first-known advising multi-task conversational system with adaptive measures for improvement. The evaluation in a real-world scenario shows its viability, and initiated the development of a corpus for academic advising, valuable for the academic and language processing research communities.

Keywords—Natural Language Processing; Dialog System; Conversational Agent; Academic Advising; Advising System; Engineering Education; E-learning; Human Computer Interaction

I. INTRODUCTION

Higher education institutions employ academic advisors to assist students in academic, professional, social and personal matters [1]. Today, academic advising is also a peer reviewed research area given the many important implications of a successful advising system such as student retention, graduation rates and student educational goals including academic engagement and performance, and career planning [1]-[2]. Therefore, it is essential for academic programs to offer students an effective advising experience, which requires advisors to innovate at the speed of their students, who each year tend to have higher expectations from their education institutions and a stronger synergy with digital technology. Thus we see the innovation trend in advising has been towards the use of communication technologies such as email, instant messaging, social networking, course-management systems, podcasts, mobile applications, online videos and blogs [3]-[5]. A quick survey of university websites shows that many have introduced the concept of eAdvising, that is, utilizing electronic means, usually web-based, to offer advising to students [6]-[8].

In 2008, Leonard detailed the profound effect of technology use by advisors and referenced the idea of an

interactive advising expert system as a possible future trend, but few institutions had shown interest in developing such a system [3]. A fully automated system is a better solution for the innovation trend in advising and addresses the concern of advisor to student ratio, in an economically challenged environment [1]. Such a system lessens the burden of academic advisors from several mundane tasks and frees up more of their time for the deeper aspects of advising, such as career planning or managing extraordinary personal situations.

There are several research publications describing advising expert systems for helping students with straightforward repetitive tasks such as choosing majors, adhering to an appropriate curriculum sequence or accessing degree audits [9]-[13]. This work is inspired by the belief that, following current technological trends, new interactive advising systems should also include a natural language interface to allow students to communicate as freely and openly as with their actual advisors. Such an innovative system could be much more attractive to students as it would allow them to easily ask a wider range of questions than those in previous expert systems and obtain immediate responses to these, instead of waiting for peers or advisors to read and reply, as with current eAdvising methods.

The main objective of this work was to construct and deploy a real-world academic advising system that allows the students to communicate freely in natural language. The system was designed to serve students of the Department of Electrical and Computer Engineering in the University of Florida (ECE-UF). This task allows for training and testing of the algorithms developed in a well-defined, domain-specific, conversational question answering application, where there are no available corpora for machine training. This work also introduces a methodology to allow the users to manage the system scale up process.

To the authors' knowledge, two publications exist that present advising systems combined with natural language processing (NLP) techniques for communication [14]-[15]. The first system was limited to only yes/no type questions and a few phrases to manage state transitions [14]. The other system features an ontology-based information retrieval engine to guide students in searching for answers after entering keywords [15]. The system developed for this research work allows unrestricted natural language communication utilizing state of the art NLP techniques.

More information on related work and the fundamental requirements of this system are available in a previous publication [16]. That work introduces the system along with its dialog manager, justifies the design components and presents preliminary results. Section 2 briefly describes the academic advising task, the user base and the advising system developed. Section 3 describes the system dialog manager and task managing components. Section 4 explains the academic planning system, including the components for communication between the students and their advisors. Section 5 describes the field tests and the analysis of the results. Section 6 contains the conclusion and future work. Finally, the appendix includes the list of academic advising topics covered in the system, a selection of user dialog from the experiments and screenshots of the web interface.

II. THE ACADEMIC ADVISING SYSTEM

A. Academic Advising

Advising tasks are identified as prescriptive, providing expert advice, and developmental, where the advisor engages in a mutual learning process with the student, in order to help the student's problem solving, decision making and evaluations skills [17]. Other studies support the idea of educative advising, where advisors are the teachers of the philosophy of the curriculum and the principles of how students learn [18]. This work provides advisors with a tool to streamline many prescriptive tasks, allowing further developmental and educative tasks during their limited face-to-face time.

In ECE-UF, students visit their academic advisor mostly by request of the advisor. These meetings must occur at least once per academic semester to evaluate the student's current academic progress and enrollment for the following term. During these meetings, most topics involve queries with accessible answers, i.e., prescriptive advising. Many of these answers are obtained directly from facts or through logic analysis, and thus may be solved algorithmically. This system is designed for these tasks, but also includes answers to some developmental advising topics. See the appendixes for the list of topics and user dialog examples.

B. Albert, the Natural Language Academic Advising System

Albert, the natural language academic advising system, provides students with an academic advising service that reflects a human interaction experience through an online text application [16]. The system does not require student training or additional human resources from the academic departments. This system enhances the academic advising experience by offering students a service that is available at any time. Albert includes multiple advising services accessible from any device with web access. Albert respects the privacy of its users, and

encourages the students to become independent and take responsibility for making decisions. As a courteous advisor employed by an academic department, the system output reflects an advisor who is polite, maintains a positive affective state with its users and simulates a personality trait intended to sympathize with the students of ECE-UF.

Albert includes knowledge about the academic programs and policies, answers to a wide range of academic frequently asked questions (FAQ), it offers recommendations for the development of a successful academic plan and referrals to other academic services. The target users are students from the ECE-UF two undergraduate degrees, Bachelor of Science in Electrical Engineering (BSEE) and Bachelor of Science in Computer Engineering – Hardware emphasis (BSCEE).

Albert contains the course scheduling information for all ECE-UF courses. A Python script reads the information from the UF Registrar's webpage and a cron job updates the knowledgebase (KB) daily. The information is stored indefinitely and is available for access when queried by date. This process assures the information is always up to date without dependency of human maintenance. All the other information in Albert is hand scripted in the KB of the system.

The main script of Albert is written in the Python programming language, version 2.7.5. This script controls all the functions of Albert and communication with each module including the web interface, the gateway interface for web communications, the user login routine, the dialog manager, the expert system for academic planning and the database. Fig. 1 shows a model of the Albert system. The dialog manager includes the natural language understanding and generation systems, and the task manager.

This website is hosted in a desktop computer at ECE-UF facilities and accessible via the Internet address <http://advising.ece.ufl.edu>. The computer has an Intel Pentium 4 processor, 1.8 GB of RAM and is running the operating system (OS) Red Hat Enterprise Linux 6.4.

The website contains scripts written in Python, PHP, JavaScript, HTML and CSS programming languages, in addition to Unix scripts to manage the daily tasks. Communication is through socket technology, which is widely used in web-based software. The website was designed for simplicity and speed, with a load time of approximately one second on contemporary versions of the popular web browsers. Users who access Albert using the Google Chrome web browser can send messages using speech recognition software. The website includes an open-source speech recognition API that with the required hardware can interpret speech data [19].

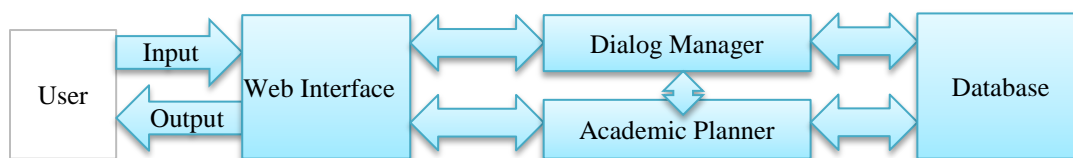


Fig. 1. A model of the main components of Albert



Fig. 2. A screenshot of the website for Albert in its initial state

Fig. 2 shows a screenshot of Albert in its initial state. The mid-upper left area shows the forms for logging into the system. Below these forms, the main window for communicating with Albert shows the instructions for connecting with the system. On the right of the main window, an independent web frame shows examples of FAQs the system contains. A screenshot showing this right-side frame is available in Appendix C. The website meets the accessibility compliances of the Americans with Disabilities Act and the Workforce Rehabilitation Act of 1973.

To protect the information of the students, the system requires users to register an account using an anonymous username and password combination. This data is stored encrypted in the Albert web server using the standard secured hash algorithm SHA-512 – 64 bits. The data are case sensitive; furthermore, the system does not allow repeated usernames entered in different letter case. These security features allow students to share their usernames with their advisors, yet keep their data secured through the password. Anonymous accounts also encourage students to communicate freely without repercussions, which is the best way to obtain sincere feedback about the system and the academic services. The account also stores the user information for recurrent advising sessions.

III. THE ACADEMIC ADVISING DIALOG MANAGER

The heart of Albert is the NLP system that drives all the input, output and states of the system, i.e., the advising Dialog Manager (DM). Given the advising task requirements, Albert's DM is built around the ChatScript (CS) scripting language [16]. This work uses CS version 2.0. The design of the DM allows a straightforward method for redrafting and distributing the system for other academic programs, by using variables in the input patterns for the proper nouns and the contents of the well-defined knowledge base. Correspondingly, the advising FAQ corpus input patterns allow for deploying to other academic programs by adding the data to the KB, without reworking the existing input templates.

The DM controls communication between CS and Albert's other functions written in Python and PHP. For example, to manage the unique technical terms and proper nouns, a

supplementary spell-checker was constructed in Python using methods surveyed from the literature [20]-[22]. When CS finds a candidate for the spell-checker, it sends the term to the Python spell checker, which returns the possible corrections to CS to determine the system response.

Regarding computational performance, a query response through the website takes approximately one second when tested from computers connected through a local area network.

A. Natural Language Generation

Language generation in the DM consists of templates containing text, pointers, variables and other control functions. To create the KB, as no data corpus is available, the output information was obtained through university documents and interviews with the academic advisors. The answers are constructed using logic functions, random generators, control structures and queries. Queries include course-scheduling information from the Registrar's Office webpage and student academic information saved with the academic planning process (APP). The KB has the information organized as a list of triples containing a question, its answer and the topic to which the answer was classified.

The system also contains a trivial amount of grammar rules, mostly for handling unrecognized input and extending off-topic dialog. For example, when responding to unrecognized questions, the system may either change the verb tense or rearrange the parts-of-speech (POS) to let the user know the answer to that question is not available.

B. Natural Language Understanding

Input template design involves identifying the keywords, POS tags, noun-phrase chunks and lexical relations of each input statement, then selecting the features that define the keywords of the template and respective topic. Fig. 3 shows an algorithm for the input *who will teach a specified course*. In this example, the topic *course schedule* includes as keywords, the list of all course names, all professor names and the words *professor* and *teach*. Alternatively, the user could also ask, *who is the professor of C++*, or if the request is within the context of the previous input, *who is teaching it*.

```
Algorithm: Respond to Who teaches X course?  
Input string S: Who teaches C++ next semester?  
Desired output: C++ is taught next semester by Name  
or C++ is not offered next semester  
If S contains a keyword of the topic course schedule  
If S matches with a Who-Teaches pattern  
If S contains a Term-Phrase keyword  
Calculate the Term value  
Else  
Use currently stored Term value  
If the course C++ exists in the schedule of the Term  
Find the corresponding data element Instructor  
Return C++ is taught by Instructor in Term  
Else  
Return C++ is not offered in Term
```

Fig. 3. Example of a procedure to match an input requesting who teaches a specified course during the next semester

These examples require additional templates either mapped to the algorithm of the template defined above or with a method to determine the context of the previous input, followed by the same mapping. The system identifies the context using features such as the current and previous input keywords, the keywords for the topics matched, the tense of any verb and the state of the variables representing potentially missing keywords. Additional details about language processing in Albert are available in the previous work [16].

For reference resolution and elliptical questions, in addition to the method described above for recognizing context, the system uses the CS feature of rejoinders. Rejoinders are input templates, which follow parent templates that elicit some expected user response. Rejoinders also allowed some input templates at the end of topics to assume certain keywords were implied. The entity recognition problem is managed by defining in CS case-insensitive concepts with pre-classified POS tags. In addition, the system has concepts defined for all the technical terms, neologisms, slang and significant LUs not available in the CS or Word Net dictionaries.

The system also has measures to deal with nonlinguistic ambiguity that the users inadvertently convey. In some cases, the best solution was to return the most likely answers, in other cases, the best solution was to request more information from the user. For instance, when a user asks, *who teaches Circuits in the next term*, the meaning of next term depends on the current date and during the spring semester, it could refer to either the summer or the fall term. For this case, the system will decide on a value for the term variable, determine the response and return the response including a method to quickly obtain the response for another term value. As evident in Fig. 3, once the user states an academic term, the system stores this value to use as current default.

When writing the input templates, the key tradeoff is between over-fitting and not generalizing well, thus increasing missed inputs, or under-fitting and causing false positives. In this work, the precision of the template is inversely proportional to the rate of occurrence of the template. That is, the responses that users most seek have a lower accuracy and higher coverage. In contrast, the precision of the template is proportional to the intricacy of the response, i.e., very specific answers have templates with higher accuracy.

When an input statement does not match with any template, the system will respond with an estimated match or request a new entry. For evaluation purposes, the system classifies these responses generated as not answered correctly. For the tests described in this work, Albert had approximately 415 input templates, for over 200 unique responses.

C. Task Manager

The task manager (TM) represents all the functions Albert executes to complement the dialog task. These functions include user account management, database management, input validation, spell checking, automatic updates, a scaling-up routine, statistical data collection and the APP. The APP is discussed in the next section. The TM is built with Python.

When the DM makes a spelling correction, the system alerts the user that a correction was made and encourages them

to verify the new input. When the DM cannot make a definitive correction, the system will give the closest answer and a short list of alternative responses with shortcuts for answers.

Albert has two main procedures for automatic updates; one procedure updates the schedule of courses, the other updates the CS scripts daily. As all template-based systems are limited by the amount of patterns for input matching, ideally the system will include methods to allow scaling-up with minimal involvement from the developers. For this reason, the design of Albert has measures in place to allow such improvements to the system. This work includes the design of one scaling-up routine, specifically, to allow users to suggest unofficial names for courses, such as the nicknames, abbreviations and acronyms that the community commonly uses.

Users have two methods to submit course names to Albert. The first method is through a direct request, i.e., they implicitly state that they want to submit a course name. The second method occurs when the system recognizes a request for course information, but the name of the course is not recognized and no spelling correction was obtained. The system obtains from the user the official name of the implied course and the recommended course moniker. This 2-tuple is automatically sent to the ECE-UF advisor database, which the advisors can access online through an independent webpage developed in this work for the task.

The second phase of the scaling-up procedure involves the ECE-UF advisors using the online system to accept or reject the user proposed course name. If the name is rejected, the process ends. If the name is accepted, the 2-tuple is added to a table in CS, which was designed in this work for this purpose. With the automatic daily updates, this data pair is available for users by the next calendar day.

To evaluate Albert, following the academic advising guidelines of the National Academic Advising Association [23] and the Council for the Advancement of Standards in Higher Education [24], assessment includes direct and indirect evaluation, qualitative and quantitative methodologies, and data collected from students and other constituencies. For qualitative analysis, the measures include collecting feedback from the students and assessment reports from the ECE-UF academic advisors.

For quantitative analysis, data from the user log files are used to measure the information of the input-output messages, login events, message timestamps and suggestions each user made through the scaling-up process. In addition, the system classifies its responses in three categories, a positive response, a negative response and off-topic responses. The system does not provide an automatic estimate of false positive (FP) outcomes, that is, input statements incorrectly matched to a determinate response, or false negative (FN) outcomes, i.e., input that should have matched. Therefore, the analysis of the FP and FN outcomes is done through an estimation of the data. The results from the estimation allow computing the standard statistical evaluation metrics for similar NLP systems, namely, precision, recall, accuracy and the F1 measure [20], [25].

Albert also includes a questionnaire for students who complete the APP.

- Is this process helpful for your academic planning?
Very helpful 5 4 3 2 1 Not at all helpful
- Is this system easy to use?
Very easy 5 4 3 2 1 Not at all easy
- What is your opinion on the natural language chat application?
I like it a lot 5 4 3 2 1 I do not like it at all

Any comments, complaints, suggestions?

Fig. 4. Questionnaire required to submit the APP information

To submit the APP information, users must answer a questionnaire of three Likert items, as shown in Fig. 4. The figure also shows an optional write-in text area where students can leave feedback. As the system is anonymous, these results and comments are not sent directly to the ECE-UF advisors.

IV. THE ACADEMIC PLANNING PROCESS

The APP is an expert system designed to offer students guidance and recommendations when preparing their course plans for each term. Students can enter their academic record in the APP and receive recommendations on how to develop their academic plan up to graduation, based on courses completed, course prerequisites and all academic rules. Since Albert knows the schedule of ECE courses for each semester, the APP helps students create their course plan for the next semester and send it electronically to their advisor for review.

The algorithms of the system were developed in Python, while the user interface was built with PHP. The user interface of the APP runs inside the Albert webpage as an independent frame, allowing users to work on both systems simultaneously, analogous to a student filling up a form in the advisor's office. Students will complete a course plan and submit it to a database, which advisors can access via an independent website. This process is part of the objective of allowing advisors to integrate Albert into their daily advising practice. Fig. 5 shows a screenshot of Albert, where the web frame on the right shows the initial webpage for APP. The course information is accessible by scrolling down on the frame.

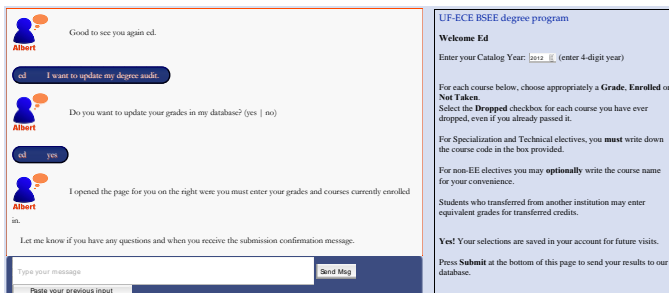


Fig. 5. A screenshot of Albert with the right side showing the first page of the APP opened inside the frame. The left side, the main dialog window, shows the conversational exchange that took place to initiate the APP

Students can initialize the APP by explicitly requesting it, with expressions similar to *I want to update my degree audit* and *I want to enter my grades*, as evident in Fig. 5. Upon initialization, the APP presents a template with the student's academic curriculum, similar to how it appears in the UF catalog. Using drop-down menus and text-boxes, students can enter the following academic information in this template; all fields in the template are validated upon submission.

- Catalog year (admission into the academic program)
- Grades in completed courses
- Courses currently enrolled in
- Courses dropped

Ideally, students would access their academic information from the university's digital records and upload the data to Albert. However, at the time, it was not possible to access the university's data and not practical to develop an interpreter for print scans of the data. In any case, freshmen had no records while others only have to enter the data once for their account, and having students examine their grades is a favorable self-assessment exercise for writing the course plan, albeit an exercise most students do not dedicate the effort to complete.

After submitting their records, Albert will invite students to prepare a course plan for the next term. For this process, Albert opens the second webpage of the APP, which contains the courses remaining to complete the degree requirements, the courses the student can take, any course the student can repeat, a checkmark for graduation candidates and the questionnaire in Fig. 4. Appendix C contains a screenshot with this APP frame. Students can repeat at any time, any step of this process, as advisors are not automatically notified of student submissions.

When a student is prepared to discuss the academic plan with the advisor, he will provide the username with Albert, to allow the advisor to obtain the plan from the online database. Students do not need to share their passwords. Fig. 6 shows this webpage, after a successful search of the user *ed*. The webpage provides advisors with a text pad area to save notes about this user.

Regarding computational performance, the response time of the APP is also within the approximately one second delay between events the rest of Albert offers.

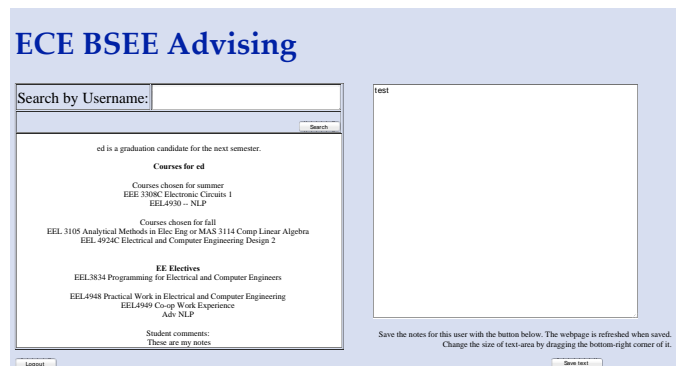


Fig. 6. The advisor database website, with an example student report

V. EXPERIMENTS AND ASSESSMENT

The experiments were designed for students of the BSEE and BSCEE programs, though the APP was only available to BSEE students. There were approximately 950 students in these programs, with approximately 500 in the BSEE program. The tests took place between October 4 and November 27, 2013. The tests began when the ECE-UF advisors informed the students of the availability of Albert and provided the web address for the system. The address was not available in any other medium. During this period, the advising staff was unexpectedly short-handed, so students were encouraged to utilize Albert to get their academic plan approved in time for registration period, except first-year students, who were encouraged to personally visit an advisor. The students did not have workshops on how to use the system. A video tutorial was available via a link in the website; see Fig. 1. The tutorial's website registered about 90 unique cookies during the period.

Results refer to users as unique accounts created. Identifiable log files were removed, including those from faculty and advisors. Users who made less than three statements were also removed. The remaining user files were included in the results, even when the user never meant to converse about academic advising. The system compiled data from 387 users. The data showed that 53% of these users made less than ten entries. Many of these used the system explicitly for the APP process. Registering and completing the APP process required a minimum of four statements. The average login per user was two, as 78% of users had at most two.

From the user total, 292 completed the first part of the APP. From the catalog year data, as expected, most students were in the mid years of academic progress for the four-year program. For the second part of the APP, Albert collected survey results from 224 students. Fig. 7 shows the results from the survey. The results reflect mostly positive reviews, as a majority of students gave values of three and four for how much they liked it and how easy it was. The rating of helpfulness is smeared over two, three and four. Conversational systems with similar surveys obtained averages within the low threes, to four and a half [26]-[27]. However, these systems have a reduced input domain, as the system is who drives the dialog, and users know beforehand what to expect from the system.

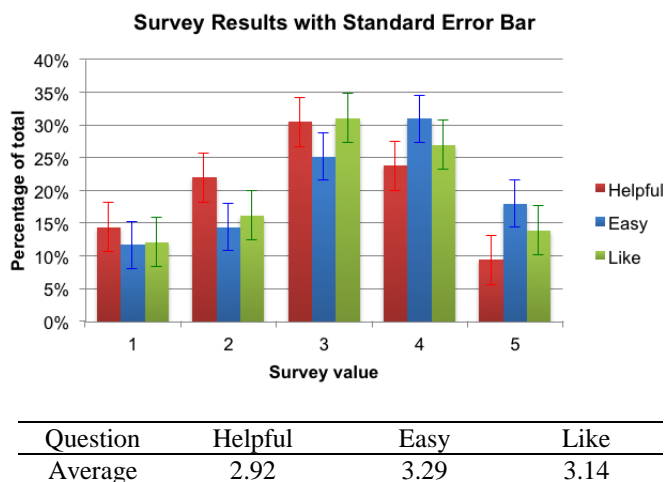


Fig. 7. Survey results with a standard error bar for each question

To support the survey results, the comment section was added to the survey on October 12, 2013. Of the 191 users who completed the APP after this date, 61 wrote comments, of which eleven were positive feedback, 45 were negatives remarks and five were technical suggestions with neutral sentiment. The positive remarks were general compliments and appreciating the course plan system. Of the negative comments, 47% criticized the system as not appropriate for substituting a human advisor, 42% criticized having to manually enter their academic record, while the remaining 11% reported technical difficulties and unique situations.

The disapproval of the method for entering the grades was understandable and expected. As described previously, given the privacy concerns, the method was a quick resolution. Once academic departments manage the system, they will have the resources to access the official data upon student request.

To address the comments about substituting a human advisor, the main solution is to inform the students about the objectives for Albert, as the advisor integrates the system into the regular tasks. A basis throughout the design of Albert is that to appreciate the value of this service, students must understand that the objective is to assist and not replace. The fact that almost half of the negative comments concern this statement validates its significance. Unfortunately, the disturbance in the advising services hampered this guideline. Altogether, 89% of the negative remarks comprise two realistic provisions that are straightforward to implement. Appendix B shows dialog extracted from the user log files.

The ECE advisor, who processed the student submissions from Albert, completed a review of the system and the course submission process. In general, the advisor was very encouraging of the system and service provided, with recommendations in line with student reviews. For the scale up process, six users completed submissions for course names, three of which made meaningful contributions.

A measure to evaluate the NLP of the system was developed, where user input is classified as the following.

- Literal match (LM) are input statements that exactly match a FAQ listed on the webpage FAQ examples.
- Partial match (PM) are statements that have partially matching templates.
- Outcome negative (ON) includes false negatives (FN), i.e., statements not recognized and true negatives (TN), i.e., statements outside the design scope.
- Outcome positive (OP) includes correct responses or true positives (TP), and false positives (FP).

The LM statements are the 71 FAQs listed on the webpage. These include the examples for initiating the APP, the help command and an example from each topic in Albert. The LM statements do not have any uncertainty for recognition; therefore, these are subtracted from the total input to evaluate the system error. As the APP was a main feature of Albert, eliminating these commands increases the estimated error. As a subsequent improvement, the webpage includes fewer examples, to encourage users to utilize original expressions.

The PM statements are to help the user obtain the information of interest, return incomplete answers and for statements that are outside the scope of the system, to which the response is a statement related to that topic and lets the user know that more information on that topic is not available. Although these templates were successful in their NLP design, given the user did not directly receive the desired response, these are not classified as outcome positive.

Results are available from 366 users between October 7 and November 27, 2013. Table I shows the results from these users and the amount of input statements under each classification. All non-LM statements are, accordingly, Original statements.

The results in Table I show that about 60% of the students copied an instruction exactly as written, which for all purposes is akin to making a selection from a menu. While this result suggests that the interface could benefit from menu selections, the objective of this experiment is to encourage unrestricted expressions. Having approximately 80% of the students initiate the APP process and 60% using example input shows the student preference of using the quickest possible method to achieve a goal.

Nevertheless, to serve as an educational tool, the previous data showed students benefit from a display of the FAQ. A solution is to include a display of topics with less example statements, while extending the NLP routines that allow users to access data by navigating through topics. While this approach is not within the scope of the vision for Albert, the data shows that including both approaches would increase the user-base by allowing users to explore the capabilities of the system through a more familiar experience.

Table I also shows that less than 15% of all input was not recognized, however, these statements came from 55% of the users. This result is expected from a system with a restricted domain that accepts all type of input. Any user who decides to test the boundaries of the system will contribute to this result. Fig. 8 shows the distribution of the ON classified statements.

To reduce the ON responses, false negatives were continuously identified and updated in the system. As the data collection increased, the number of false negatives decreased. By the fifth week of the almost eight-week period, the amount of false negatives per user had dropped to almost zero. Approximately 25% of the users accessed the system during this culminating period. During these last three weeks, no measures were taken to update the system.

To determine the number of FP and FN statements, it is necessary to manually evaluate the OP and ON statements. For this evaluation, a FP outcome is a response that is not relevant to the input statement. This definition of FP error conditions the system's capabilities of responding to the statement with respect to the available information.

TABLE I. RESULTS OF THE INPUT STATEMENT CLASSIFICATION

	Total	LM	Original	PM	ON	OP
Users	366	60.5%	-	11.9%	55.0%	99.7%
Input	4952	12.5%	4332	1.7%	14.7%	83.5%

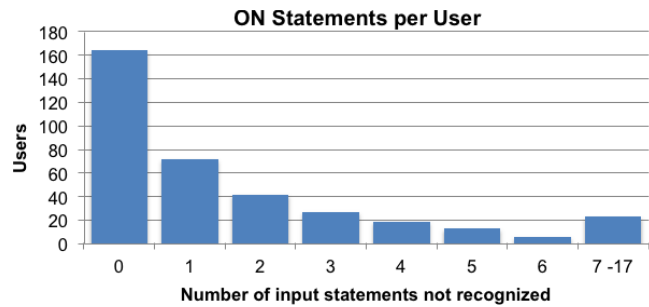


Fig. 8. Distribution of ON input statements

For example, if a user asks, *Who is my financial advisor*, the system response, *For financial information please see the following website*, is not classified FP as the answer is relevant and it is the best response available in the system. If for the same input the system replied *Your BSEE advisor is Mrs. Chillingworth*, the response is classified as FP, given the BSEE advisor is not relevant to the user question.

A FN outcome occurs when the input statement has an available answer, yet the system did not respond correctly. Following the same example as above, if a user asks, *Who is the person in charge of financial advising*, the system response, *I do not understand that question*, is classified FN, as this question should have been responded as stated previously. The FN outcomes include cases caused by technical difficulties in any area of the system; however, with respect to overall system performance, they are incorrect responses. TN outcomes occur when the input statement is outside the design scope.

We estimated FP and FN by selecting random samples from the OP and ON statements. To select the samples, a uniform random number is assigned to each of the 3618 OP statements and to each of the 641 ON statements. Each statement is tested for the binary outcome FP or not FP, and FN or not FN respectively. By treating each statement as an independent random variable, the outcomes of the tests follow a Bernoulli distribution. In this distribution, the maximum variance, $p * (1 - p)$, occurs when the mean p is equal to one half, thus the variance is equal to one fourth. For a given sample size n , by the central limit theorem, the distribution is closest to the standard normal distribution when p is near one half. Under this scenario, a conservative estimate of the sample size n needed to estimate p with a confidence level $1 - \alpha$ and margin of error e is given in (1), where the $[]$ operator represents rounding to the next integer and Z_α is the estimated standard score for a given two-sided confidence level [28].

$$n = [Z_\alpha / (4 \square)] \tag{1}$$

For a maximum margin of error equal to 10% and a confidence interval of 95%, (1) returns a minimum size of 97 samples. Table II summarizes the result for each test.

TABLE II. RESULTS FROM THE OUTCOME ERROR ESTIMATION

Outcome	Total Statements	Confidence Interval	Error Margin	Sample Size	Error
Positive	3618	95%	10%	97	15.46%
Negative	641	95%	10%	97	24.74%

The tests returned 15 false positive statements, for 15.46% of the OP samples and 24 false negatives for 24.74% of the ON samples. Using these results, it is possible to determine the statistical measures precision, recall, accuracy and the F1 measure [20]. Table III shows the result of each measure.

The results show the system performance metrics are all close to 80%. These results are estimated minimums; given the error is an estimated maximum. Recent studies for comparable e-learning systems show Albert offers a high-level performance for the complexity of the task [29]–[30]. Regarding the main NLP tasks in Albert, specifically keyword extraction, question answering and natural language interfaces, recently published systems that require data corpora for training, obtain results that show the performance of Albert is competitive [31]–[35]. Another performance measure is a study of human accuracy and response time, though such a study is not available. While Albert cannot yet compete in accuracy with a human, its instant response time is hard to beat. In any case, the results show Albert offers a competitive performance with much promise for advancement as data collection continues.

The results for precision and recall are in line with the design goal of providing students with high precision answers and minimizing false positives. At the same time, the results show that over 75% of the queries to the system obtained a reply that is effective versus an input-not-recognized type answer. This result is valuable considering users did not have any training on how to use the system.

Overall, ECE-UF students and personnel were appreciative of the service provided and supportive for future developments. The results offer advisors valuable assessment for the areas that most concern students.

The data shows that, as is customary in electronic text communications and online search engines, students prefer to ask questions using the minimum amount of words possible, i.e., entering isolated keywords, instead of through a Standard English sentence. Therefore, Albert should include methods to allow this user preference, while concurrently inspiring students to use complete expressions. Facilitating speech recognition will also expedite this design.

Initially, students were reluctant to follow a new process for the advising chores. Indeed, more effort is required in marketing the service as a practical option versus visiting a human advisor, who could be a short walk away. However, during periods when the waiting time is long or when the advisor is not available, the predisposition to experiment with an option is very high.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

The fundamental objective of this work is to provide students with an automated online advising experience that is as close as possible to traditional human interaction.

TABLE III. SYSTEM NLP PERFORMANCE ESTIMATION RESULTS

Precision	Recall	Accuracy	F ₁ measure
≥ 84.54%	≥ 77.36 %	≥ 79.90%	≥ 0.81

This work presents the design, development, deployment and evaluation of an intelligent natural language conversational system for academic advising. The online system includes an advising dialog engine, an academic planning system that integrates the students with the advisors, and a method to allow the users to contribute to the system's KB.

This work contributes real-world solutions for the academic community, through a unique combination of software applications and advanced NLP techniques. With the experimental data collected, this work has originated a KB of academic advising FAQs that will serve to build a corpus on the topic, to allow the development of systems for other academic programs by mostly replacing answers.

This system is currently in operation for students of the BSEE and BSCEE degree programs of ECE-UF. Experiments with the students showed Albert is a practical and valuable application for academic advising, based on positive reviews from all constituents and system performance analysis. Within a confidence interval of 95% and a margin of error of 10%, Albert showed an estimated precision of 85%, a recall of 77%, an accuracy of 80% and a F1 measure of 0.81. Overall, users were supportive and excited about future developments.

B. Future Work

Albert continued to serve the ECE-UF department during the year 2014. During Spring 2014, the system featured a twofold process for the APP, in which students could simultaneously prepare a course plan for summer and fall.

The immediate development is adding a computerized method for submitting the academic record, preferably through the university's online system. Obtaining the full academic record from the students will increase the services the system can implement for students and advisors.

Additional future work includes an escalation mechanism to forward selected student conversations to the advisors, adding academic information from other academic departments and increasing the KB. With an extensive KB, the system can include statistical learning techniques to the NLP system.

VII. APPENDIXES

A. Topics included in Albert

- Academic curriculum and rules
- Course schedules
- Internships
- Graduate programs and minors
- Financial assistance
- Faculty and staff
- The academic calendar
- Student organizations
- University campus
- Information about Albert

B. User Dialog from the Experiments

The following are excerpts from the log files of three users. User statements are in bold font weight and their syntactic errors are preserved for authenticity. Italic type represents actions. Albert's answers are curtailed with an ellipsis.

User 1 – F9788

What courses do you recommend I take? ==> I will tell you which courses you could take next semester ... Do you want to enter your grades into the database? (yes | no)

y ==> *The APP is launched at this time.*

confirm ==> *The first part of the APP is completed.* Do you want to choose your courses for the next semester and submit the list to your advisor?

y ==> *The second part of the APP is launched at this time.*

what courses do you recommend I take? ==> Assuming your info is up to date, these are the courses you could take next semester.

 EEL3472 Electromagnetic Fields ...

User 2 – Willwrdrn

Welcome back willwrdrn

I want to drop a course ==> So you want to drop a course? Are you in upper division or lower division? ...

upper ==> Students are given TWO drops after completing 60 credit hours. Any unused lower division drop ...

what is the deadline for dropping a course ==> I'll tell you this, in rare cases it is possible to drop a course after the ...

am I able to drop a course via ISIS? ==> So you want to drop a course? Are you in upper division or lower division? ...

upper ==> Students are given TWO drops after ...

so how do I drop a course? ==> You drop classes via ISIS. If the add/drop period has ended you can still do it, but ...

alright! thank you! That's all the information I need today ==> Well, I am learning every day. What can I do for you?

logout ==> Before you go, please share any feedback ...

Perhaps it was just that my queries were better and fewer this time, but the system seems to have improved. Good job! I hope it gets to access to more information soon so that it can answer more questions instead of some variation of I don't know ==> All right, thanks!
 Later, Gator!

User 3 – Novae

show my schedule ==> You want me to show you the courses you chose for next semester? OK. These are your ...

programming classes ==> Electrical Engineering students must take one programming course in Java, C or C++ ...

prerequisite for eel 3701c ==> The prerequisite for EEL 3701C is a Programming Course.

is eel3834 being taught next semester? ==> The course EEL 3834 is not offered in Spring 2014.

what classes can i take ==> Assuming your info is up to date, these are the courses you could take next semester ...

circuits 2 ==> I recognize CIRCUITS 2 is a course in our department, but could you be more specific about ...

Who will teach EEE3308C? ==> For Spring 2014 we have
 EEE 3308C by Yoon, Yong Kyu

eel3701c any sections available ==> For Spring 2014 we have
 EEL 3701C by Gugel, Karl S

Is it ok if I take less than 12 credit hours? ==> UF policy states that the minimum full-time load for undergraduate ...

How many critical tracking courses are required for my major? ==> You need to pass 8 critical tracking courses ...

C. Screenshots of the Website

Ask Albert, our Natural Language Advisor, the same questions you would ask your advisor, including

I want to choose my courses for the next semester.

What courses do you recommend I take?

When is Circuits 2 taught next semester?

Where is EEL 3112 given next spring?

What will Dr. Gugel teach?

Who will teach EEE3308C?

Start your course selection process with

I want to update my degree audit.

or

I want to enter my grades.

Teach Albert your preferred name for a course:

I want to teach you a course name.

Type **Help** at any time for more assistance.

Press here for more example questions

Fig. 9. Main FAQ webpage

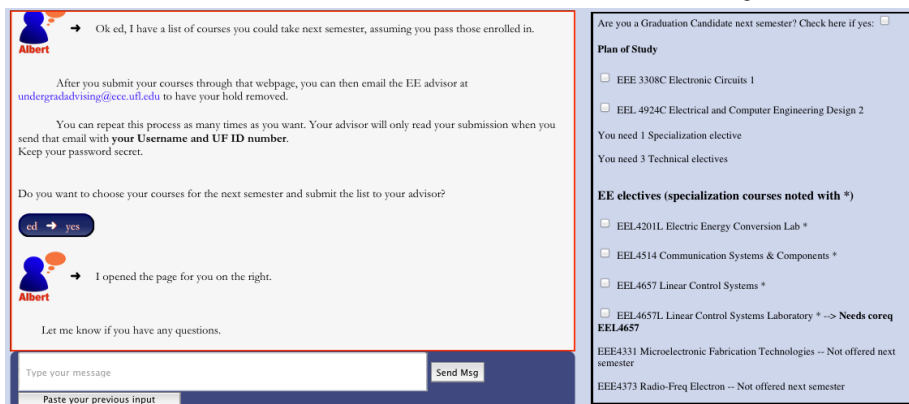


Fig. 10. Main webpage for Albert showing an example of the second part of the APP

ACKNOWLEDGMENT

We appreciate the support from participating ECE-UF students, advisors and personnel of the student services office.

REFERENCES

- [1] V. N. Gordon, W. R. Habley, T. J. Grites and National Academic Advising Association. *Academic Advising: A Comprehensive Handbook*, 2nd ed. San Francisco: Jossey-Bass, 2008.
- [2] K. Soria, "Advising satisfaction: implications for first-year students' sense of belonging and student retention", in *The Mentor: An Academic Advising J.*, Oct 2012. [Online]. Available: <http://dus.psu.edu/mentor/2012/10/advising-satisfaction/> Retrieved Feb 2013.
- [3] M. J. Leonard, "Advising delivery: using technology", in *Academic Advising: A Comprehensive Handbook*, V. Gordon, W. R. Habley, T. J. Grites and National Academic Advising Association, Eds. San Francisco: Jossey-Bass, pp. 292-306, 2008.
- [4] L. Waldner, D. McDaniel, T. Esteves and T. Anderson, "The eQuad: a next-generation eAdvising tool to build community and retain students", in *The Mentor: An Academic Advising J.*, Oct 2012. [Online]. Available: <http://dus.psu.edu/mentor/2012/10/quad-eadvising-tool-build-community-retain-students/> Retrieved Feb 2013.
- [5] National Academic Advising Association, *Advising Technology Innovation Awards*. [Online]. Available: <http://www.nacada.ksu.edu/Events-Programs/Awards/Association-Awards/Technology/Technology-Winners.aspx> Retrieved Feb 2013.
- [6] H. S. Hart, R. B. Hussey, M. J. Leonard, J. Levin and S. M. Winck, "eLion: Penn State's comprehensive web-based academic advising system", in *The Mentor: An Academic Advising J.* [Online]. Available: <http://dus.psu.edu/mentor/bookstore/elion-advising-system/> Retrieved Feb 2013.
- [7] University Advising Center, The University of Texas at Arlington, Arlington, TX. [Online]. Available: <http://www.uta.edu/universitycollege/current/academic-planning/uac/index.php> Retrieved Feb 2013.
- [8] Academic Advising and Career Center, eAdvising, University of Michigan-Flint, Flint, Michigan. [Online]. Available: <http://www.umflint.edu/advising/eadvising.htm> Retrieved Feb 2013.
- [9] R. M. Siegfried, A. M. Wittenstein and T. Sharma, "An automated advising system for course selection and scheduling", *J. Comput. Sci. Coll.*, vol. 18, no. 3, pp.17-25, Feb 2003.
- [10] F. Albaloooshi and S. Shatnawi, "HE-Advisor: A multidisciplinary web-based higher education advisory system", *Global J. Computer Sci. and Tech.*, vol. 10, no. 7, pp. 37-49, Sept. 2010.
- [11] A. N. Nambiar and A. K. Dutta, "Expert system for student advising using JESS", *Int. Conf. Educational and Inform. Tech.*, vol. 1, pp. VI-31 – VI-315, Sept. 2010.
- [12] T. Feghali, I. Zbib and S. Hallal, "A web-based decision support tool for academic advising", *J. Educational Technology & Society*, vol. 14, no. 1, pp. 82–94, 2011.
- [13] E. Onyeka, D. Olawande and A. Charles, "CAES: A model of an RBR-CBR course advisory expert system," *Int. Conf. Inform. Soc.*, pp.37-42, June 2010.
- [14] B. C. McMahan and R. A. Bates, "An automatic dialog system for student advising", in *J. Undergraduate Research*, Minnesota State University, Mankato, 2010.
- [15] C. M. Leung, E. Y. M. Tsang, F. S. S. Lam and D. P. C. Wai, "Intelligent counseling system: a 24 x 7 academic advisor," *EDUCAUSE Quart.* 33, no. 4, 2010. [Online]. Available: <http://www.educause.edu/edUCAUSE+Quarterly/edUCAUSEQuarterlymagazineVolum/intelligentCounselingSystema24/219101> Retrieved Feb 2013.
- [16] E. Latorre and J. Harris, "A natural language conversational system for online academic advising", *Florida Artificial Intell. Research Soc. Conf.*, North America, May. 2014. Available: <http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS14/paper/view/7842> Retrieved Dec 2014.
- [17] D. C. Appleby, "Advising as Teaching and Learning", in *Academic Advising: A Comprehensive Handbook*, V. Gordon, W. R. Habley, T. J. Grites and National Academic Advising Association, Eds. San Francisco: Jossey-Bass, pp. 85-102, 2008.
- [18] P. L. Hagen and P. Jordan, "Theoretical foundations of academic advising", in *Academic Advising: A Comprehensive Handbook*, V. Gordon, W. R. Habley, T. J. Grites and National Academic Advising Association, Eds. San Francisco: Jossey-Bass, pp. 17-35, 2008.
- [19] Contributors to the Web Speech API Specification, Web Speech API, Speech API Community Group, 2012. [Online]. Available: <https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html> Retrieved Feb 2013.
- [20] D. Jurasky and J. H. Martin, *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd Ed, New Jersey: Pearson Education, 2008.
- [21] S. Bird, E. Klein and E. Loper, *Natural Language Processing with Python*, O'Reilly Media, 2009. [Online]. Available: <http://nltk.org/book/> Retrieved Feb 2013.
- [22] P. Norvig. *How to Write a Spelling Corrector*. [Online]. Available: <http://norvig.com/spell-correct.html> Retrieved Feb 2013.
- [23] National Academic Advising Association, Kansas State University, Manhattan, KS. <https://www.nacada.ksu.edu>
- [24] Council for the Advancement of Standards in Higher Education, "The Role of Academic Advancing Programs", Washington DC, 2011.
- [25] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley Longman Publishing, Boston, MA, 1999.
- [26] S. Gandhe, N. Whitman, D. Traum and R. Artstein, "An integrated authoring tool for tactical questioning dialog systems", *Proc. 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialog Systems*, Assoc. for the Advancement of Artificial Intell., 2009.
- [27] D. Mori, R. Berta, A. De Gloria, V. Fiore and L. Magnani, "An easy to author dialog management system for serious games", *J. Comput. Cult. Herit.*, vol. 6, no. 2, pp. 10:1-10:15, May 2013.
- [28] J. A. Gubner. *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, Cambridge, United Kingdom, 2006.
- [29] A. Olney, M. Louwerse, E. Matthews, J. Marineau, H. Hite-Mitchell, and A. Graesser, "Utterance classification in AutoTutor", *HLT-NAACL Building Educational Applcat. using NLP, Assoc for Computational Linguistics*, vol. 2, 1-8, USA, 2003.
- [30] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, S. Van Vuuren, T. Weston, J. Zheng and L. Becker, "My science tutor: a conversational multimedia virtual tutor for elementary school science", *ACM Trans Speech and Language Process.*, TSLP, vol. 7, no. 4, 2011.
- [31] D. de Castro Reis, F. Goldstein and F. Quintao, "Extracting unambiguous keywords from microposts using web and query logs data", *2nd Workshop Making Sense of Microposts*, 21st Int. Conf. World Wide Web, 2012.
- [32] E. Sneiders, "Automated FAQ answering with question-specific knowledge representation for web self-service", *Proc 2nd Int Conf Human Syst Interaction*, IEEE, May 21-23, Italy, pp. 298-305, 2009
- [33] M. Minock, "C-Phrase: a system for building robust natural language interfaces to databases", *Data & Knowledge Eng.*, vol. 69 no. 3, pp. 290-302, 2010.
- [34] M. Olvera-Lobo and J. Gutiérrez-Artacho, "Open-vs. restricted-domain QA systems in the biomedical field.", *J. Inf. Sci.*, vol. 37, no. 2, pp. 152-162, 2011.
- [35] V. Lopez, C. Unger, P. Cimiano and E. Motta, "Evaluating question answering over linked data", *Web Semantics: Sci., Services and Agents on the World Wide Web*, vol. 21, pp. 3-13, 2013.

A Comparative Study of Thresholding Algorithms on Breast Area and Fibroglandular Tissue

Shofwatul 'Uyun^{1,a}

¹Department of Informatics, Faculty of Science and Technology, Sunan Kalijaga State Islamic University, Yogyakarta, Indonesia

Sri Hartati^{2,b}, Agus Harjoko^{2,c}, Lina Choridah^{3,d}

²Department of Computer Science and Electronics,
²Faculty of Mathematics and Natural Sciences,
³Department of Radiology, Faculty of Medicine,
^{2,3}Gadjah Mada University, Yogyakarta, Indonesia

Abstract—One of the independently risk factors of breast cancer is mammographic density reflecting the composition of the fibroglandular tissue in breast area. Tumor in the mammogram is precisely complicated to detect as it is covered by the density (the masking effect). The determination of mammographic density may be implemented by calculating percentage of mammographic density (quantitative and objective approaches). Thereby, the use of a proper thresholding algorithm is highly required in order to obtain the fibroglandular tissue area and breast area. The mammograms used in the research were derived from Oncology Clinic, Yogyakarta that had been verified by Radiologists using semi-automatic thresholding. This research was aimed to compare the performance of the thresholding algorithm using three parameters, namely: PME, RAE and MHD. Zack Algorithm had the best performance to obtain the breast area with PME, RAE and MHD of about 0.33%, 0.71% and 0.01 respectively. Meanwhile, there were two algorithms having good performance to obtain the fibroglandular tissue area, i.e. multilevel thresholding and maximum entropy with the value for PME (13.34%; 11:27%), RAE (53.34%; 51.26%) and MHD (1:47; 33.92) respectively. The obtained results suggest that zack algorithm is perfectly suited for getting breast area than multilevel thresholding and maximum entropy for getting fibroglandular tissue. It is one of the components to determine risk factors of breast cancer based on percentage of breast density.

Keywords—Thresholding Algorithm; Breast Area; fibroglandular Area

I. INTRODUCTION

One of the preventive measures to decrease the number of breast cancer patients is by having routine screenings. The mammographic density is one of the parts of BI-RADS assessment proposed by the American College of Radiology (ACR) in 2004 modified from Wolfe Standards and widely used by Radiologists. One of the approaches used to assess mammographic density is quantitative and objective approach, by calculating the percentage of mammographic density by means of comparing relative amount of fibroglandular tissue and breast area [1] [2] [3] [4] and [5]. Women will have a greater risk than men if their fibroglandular tissue area is higher compared to their fat tissue in the breast area [3]. To obtain areas, (fibroglandular tissue area and fat tissue in the breast area), it is necessary to conduct segmentation process automatically by employing thresholding method. The use of proper thresholding method will be able to separate discriminate the fat tissue in the breast area and its background

and/or to separate the fibroglandular tissue and fat tissue based on the threshold value obtained. After obtaining these two areas, the ratio value can be calculated, between the fibroglandular tissue and breast area, indicating the risk factors of breast cancer. The result of threshold value can be performed either automatically or semi-automatically.

Several previous researches have used a semi-automated thresholding in mammogram image, including: [2] [5] [6] and [7]. Meanwhile, several previous researches only focused on the use of automated thresholding to obtain fibroglandular tissue area or breast area. The automated thresholding methods that have been used to obtain fibroglandular tissue include: Gaussian mixture modeling by [8] and minimum-cross entropy by [9], while the automatic thresholding methods that have been used to get breast areas include: row by row method thresholding (RRT) and average row threshold (ART) by [10] and by employing the threshold value of 18 by [11]. [4] had proposed a calculation model of breast cancer risks by computing the percentage of mammographic density. This model could be applied as a reference to help decrease breast cancer risks. In the research, [5] it did not only use the risk factors of mammographic density but also the use of other risk factors, such as estradiol level and polymorphism ESR1 as a predictor of estrogenic factors related to breast cancer in the population of Javanese people in Indonesia. The calculation model of the percentage of mammographic density was conducted by the semi-automatic thresholding method and was named GAMA DEJAVU. Meanwhile, [6] semi-automated thresholding was also employed to determine breast cancer risk factors into four risks (BI-RADS standard), by involving three Radiologists for statistically extracted rules (mean, kurtosis and skewness). Other researches which also employed semi-automated thresholding method were [2] and [7]. The objective of the use of the semi-automated thresholding method was to calculate the mammographic density based on BI-RADS on mammogram using craniocaudal view which had been previously determined on the basis of Tabar parenchymal pattern by Radiologists.

The use of RRT and ART methods by [10] has been implemented on 50 mammogram images from the public database DDSM for normal mammogram and breast cancer. The extraction results of both methods look similar. However, the performance of the ART method is better compared to RRT method for it's capability to extract breast area by eliminating the background perfectly. In addition, the limits of the breast

area of the ART methods look smoother, thus the output is more proper. On the other hand, the RRT method generates a larger breast extraction compared to breast area. Meanwhile, [11] used a threshold value of 18 to separate the breast area from its background. The result obtained from threshold 18 is the best compared to the previous two methods in the case that the periphery of the breast is highly smooth. However, the use of the threshold value of 18 has a weakness for its static nature. It means that no matter what the histogram condition of mammogram is, the threshold value used is still 18. Thus, when applied to the mammogram image possessing very little or much difference histogram, the threshold value of 18 is not the best threshold value.

On the other hand, the use of several automated thresholding methods to obtain fibroglandular tissue areas, such as Gaussian mixture modeling by [8], is aimed at conducting mammogram image segmentation by using mediolateral oblique view into several areas or sections anatomically. The mammogram is segmented into five components, namely: background, uncompressed fat, fat, dense tissue and muscles. Meanwhile, the minimum cross entropy by [9] is used to obtain the fibroglandular tissue area by separating the fat tissue from the breast area. [4] has developed the computational model in determining the breast cancer risk factors based on the percentage of mammographic density. The use of Zack algorithm to obtain the breast area and multilevel thresholding to obtain fibroglandular tissue area in the proposed model has better accuracy, sensitivity and specificity if compared to the use of maximum Zack algorithm and maximum entropy. The assessment of algorithm performance for the new thresholding was performed simultaneously to obtain breast cancer risk factors. Thereby, this research would be focused more on comparing the performance of several automated thresholding methods if employed to obtain both objects.

II. MATERIAL AND METHOD

This research used mammograms taken from patients who had mammography check-up in Oncology Clinic, Kotabaru, Yogyakarta, with craniocaudal views. Those images were the digitalization from analogue images into digital images with bmp extension in various sizes. They had been classified by Radiologists into four risk factor categories in accordance with BI-RADS standards.

A. Pre-processing

In the pre-processing stage, there was only one process conducted to simplify the segmentation process. The process was the conversion of RGB images into to gray images. Subsequently, the gray image from the stage results would undergo segmentation process by using several automated thresholding algorithms with two different objectives, i.e. to obtain the breast area and to obtain fibroglandular tissue area.

B. Segmentation

The segmentation process was performed by using five automated thresholding methods, namely: Zack algorithm, Otsu, multilevel thresholding, maximum entropy and minimum entropy. Those five algorithms generated threshold value which was automatically implemented on the mammogram

images with the aim to separate the breast area from its background and to separate the fibroglandular tissue from the breast area. *Firstly*, Zack algorithm or triangle thresholding is algorithm to be used to determine the generated threshold value based on the gray intensity histogram ($h[x]$) out of some component of the image parts associated with a line. In broad sense, the algorithm is consisted of several procedures, namely: finding the min and max value of the degree of grayness, finding the farthest periphery and describing the connecting lines [12]. *Secondly*, the Otsu thresholding is a searching method of an optimal threshold value obtained by using discriminating criteria to maximize the distribution result of the two classes on the grayness level. This method was done to minimize the total weights of some variants in the class of the background and foreground pixels to obtain the optimal threshold [13]. *Thirdly*, the multilevel thresholding is a recursive algorithm based on the Otsu method introduced by [11]. It is considered effective in computing to find many threshold levels in the images by using table look-up. The working of this method is by modifying the class variance which is previously calculated and stored in the look-up table to reduce the computation complexity of cumulative probability and the mean of each class [14]. *Fourthly and fifthly*, the maximum and minimum thresholding entropy is a thresholding algorithm based on the entropy distribution from the degree of gray image. The maximum entropy obtained based on the maximization of the entropy value of the two classes is foreground and background [15]. Meanwhile, the search process of threshold value in minimal entropy is based on the minimizing of entropy value between the two classes.

C. The Analysis of Segmentation Method Performance

The performance comparison of several thresholding algorithms in the segmentation process was assessed based on the value of three parameters, namely: PME, RAE and MHD [16] and [17]. The use of those three parameters was aimed to compare the quality of several mammogram images as the results of segmentation process generated based on the threshold value from the thresholding algorithm. The images from segmentation results were compared to the reference images which had been verified by Radiologists using semi-automated thresholding.

1) Percentage Misclassification Error (PME)

PME is a picture of correlation between segmentation results image and Radiology observations result reflecting the percentage between some mistaken pixel background as it is considered as the pixels of the objects or vice versa. The formula for PME is shown in Equation 1.

$$PME = 1 - \frac{|B_O \cap B_T| + |F_O \cap F_T|}{B_O + F_O} * 100\% \quad (1)$$

B_O is the number of pixels on the background of Radiology observation results, F_O shows the number of pixels on the object of Radiology observation results, B_T represents the number pixels on the background of the segmentation result images generated by thresholding method, and F_T shows the number of pixels on the object from the images of segmentation results produced by the thresholding method.

2) Relative Foreground Area Error (RAE)

RAE is a parameter for measuring the number of difference among thresholding result images on reference images in which the Radiology observation result. The formula for RAE is defined in Equation (2) and (3)

$$RAE = \frac{A_O - A_T}{A_O}, \text{ Jika } A_T < A_O \quad (2)$$

$$RAE = \frac{A_T - A_O}{A_T}, \text{ Jika } A_T \geq A_O \quad \square (3) \square$$

in which A_O is the object area of the reference images, and A_T is the object area of binary image which is the result of the use of thresholding method.

3) Modified Hausdorff Distance (MHD)

MHD is a method used to measure the distortion of the object form resulted from the thresholding process from the reference images object. The MHD formula is shown in Equation (4) and (5).

$$MHD(F_O, F_T) \max(d_{MHD}(F_O, F_T), d_{MHD}(F_T, F_O)) \quad (4)$$

Where,

$$d_{MHD}(F_O, F_T) = \frac{1}{|F_O|} \sum_{f_O \in F_O} \min_{f_T \in F_T} ||f_O - f_T|| \quad (5)$$

F_O dan F_T represent the number of pixels on the object area derived from reference images and the images resulted from thresholding process.

III. RESULTS AND DISCUSSION

The final process in this research is the analysis to determine the performance of each automated thresholding algorithm that is used to obtain the breast area and the fibroglandular tissue area. The use of the five thresholding algorithms tested on five mammogram images as the samples is to the extent of 1 mammogram to 5 mammograms shown in Figure 1. (a) to 1 (e). The histogram of the five mammogram images is shown in Figure 2. (a) to 2. (e), which means that the mammogram image in Figure 1. (a) has the form of a histogram shown in Figure 2 (a), so as for the other four mammogram image types. The observation results on the form of histogram of the five mammogram images show that the histogram resulted has various forms. The histogram forms are not consistent in bimodal or nearly-bimodal forms, but there are several unimodal forms or in multimodal forms.

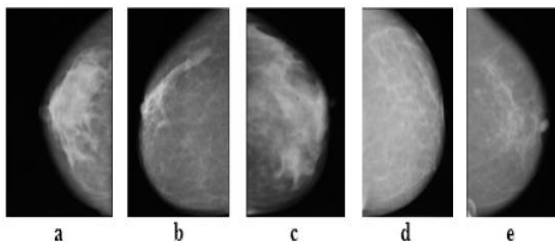


Fig. 1. Examples of mammogram : (a) mammogram 1, (b) mammogram 2, (c) mammogram 3, (d) mammogram 4, (e) mammogram 5

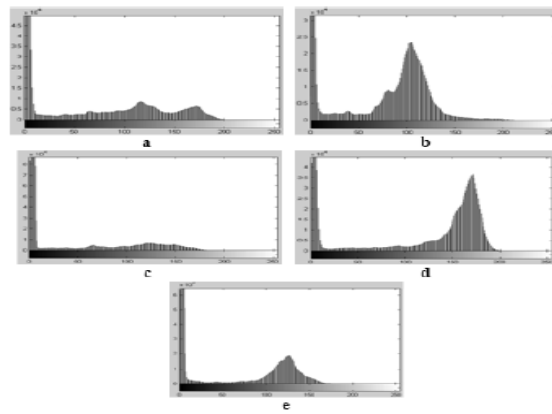


Fig. 2. Histogram mammogram : (a) mammogram 1, (b) mammogram 2, (c) mammogram 3, (d) mammogram 4, (e) mammogram 5

Subsequently, the threshold value was sought for those five algorithms described in the previous sub-chapter. The threshold value obtained for each mammogram image by thresholding algorithm is shown in Table 1. For example, for mammogram 1, it has a threshold value of 13 for Zack algorithm, 70 for Otsu, 141 for multilevel thresholding, 128 for maximum entropy and 50 for minimum entropy. The reference images made as the comparator are the reference binary images resulted from segmentation by using semi-automated thresholding conducted by Radiologists. There are two threshold values used to obtain fibroglandular tissue area and breast area. The complete result of those five mammograms is shown in table 2. For example, to obtain breast area and fibroglandular area on mammogram 1, the threshold value used is 13 and 122.

The performance evaluation results for breast images using those five thresholding algorithms are shown in Table 3. The first parameter, PME based on a formula (1) reflects the percentage between several mistaken background pixels considered as the pixels of the object or vice versa. For the second parameter, RAE is based on formula (2) and (3) functions to measure the difference between the images resulted from thresholding algorithm and each of their reference images. In the third parameter, MHD is aimed to measure the distortion of the object forms resulted from the use of the five thresholding algorithms based on the objects of the reference images. The smaller the value for the three parameters indicates its better performance. It means that threshold values resulted from automated thresholding have similar values to the threshold values resulted from Radiology observations using semi-automated thresholding. Likewise, the computation process for the three parameters is to obtain a complete fibroglandular tissue area.

TABLE I. THRESHOLDING VALUE RESULTING FROM THRESHOLDING ALGORITHM

Mammo Gram	Thresholding Algorithm				
	Zack	Otsu	Multi level	Max Entropy	Min Entropy
1	13	70	141	128	50
2	12	58	131	141	51

Mammo Gram	Thresholding Algorithm				
	Zack	Otsu	Multi level	Max Entropy	Min Entropy
3	11	63	130	82	46
4	14	85	154	146	35
5	12	64	125	179	54

TABLE II. THRESHOLD VALUE DETERMINED BY RADIOLOGISTS

Mammogram	Semi-Automatic Thresholding	
	Breast area	Fibroglandular area
1	13	122
2	11	122
3	15	82
4	21	177
5	17	137

The complete results of the threshold value obtained for each mammogram with the five thresholding algorithms are shown in Table 4. Subsequently, the computation results of the performance of the five thresholding algorithms are shown in Table (3) and (4). A computation of the mean value was done and the results are shown in Table (5).

The smaller the value indicates the smaller the difference, meaning that the images resulted from the segmentation by using thresholding algorithm is close to the images resulted from segmentation by using semi-automated thresholding by Radiologists.

TABLE III. THE PERFORMANCE EVALUATION OF THRESHOLDING METHOD TO GET THE BREAST AREA

Mammogram	Thresholding Algorithm				
	Zack	Otsu	Multi level	Max Ent	Min Ent
<i>PME</i>					
1	0.00%	61.82%	86.00%	81.58%	57.75%
2	0.17%	6.96%	65.79%	67.69%	6.04%
3	0.42%	6.02%	23.15%	10.24%	3.76%
4	0.40%	5.05%	24.33%	17.67%	0.82%
5	0.68%	3.76%	27.16%	44.89%	3.06%
<i>RAE</i>					
1	0.00%	61.87%	86.07%	81.64%	57.79%
2	0.24%	9.80%	92.56%	95.22%	8.50%
3	1.23%	17.84%	68.66%	30.37%	11.16%
4	0.60%	7.60%	36.58%	26.56%	1.23%
5	1.50%	8.37%	60.43%	99.89%	6.82%

<i>MHD</i>					
1	0.00	1.62	6.18	4.45	1.37
2	0.00	0.11	12.43	19.93	0.09
3	0.01	0.22	0.02	0.44	0.13
4	0.01	0.08	0.01	0.36	0.01
5	0.02	0.09	1.53	906.28	0.07

TABLE IV. THE PERFORMANCE EVALUATION OF THRESHOLDING ALGORITHM TO OBTAIN FIBROGLANDULAR AREA

Mammogram	Thresholding Algorithm				
	Zack	Otsu	Multi level	Max Ent	Min Ent
<i>PME</i>					
1	79.77%	17.95%	6.23%	1.8%	22.03%
2	60.66%	53.87%	4.96%	6.86%	54.79%
3	10.66%	4.22%	12.91%	0%	6.48%
4	57.69%	52.24%	32.96%	39.62%	56.46%
5	37.49%	33.05%	9.65%	8.08%	33.74%
<i>RAE</i>					
1	79.83%	47.12%	30.92%	8.96%	52.22%
2	85.54%	84.01%	48.38%	66.87%	84.24%
3	31.23%	15.25%	54.99%	0%	21.63%
4	86.21%	84.99%	78.12%	81.11%	85.95%
5	82.18%	80.25%	54.26%	99.39%	80.58%
<i>MHD</i>					
1	3.96	0.89	0.45	0.1	1.09
2	5.92	5.25	0.94	2.02	5.34
3	0.45	0.18	1.22	0.0	0.28
4	6.25	5.66	3.57	4.29	6.12
5	4.61	4.06	1.19	163.18	4.15

The computation results for the mean value of the performance of the thresholding algorithm are shown in Table 5. For example, Zack algorithm has the smallest value for all of the three parameters compared to other four algorithms, with respective value for PME, RAE and MHD by 0.33%; 0.71% and 0.01. It indicates that Zack algorithm has the best performance to obtain the breast area. Meanwhile, to obtain the fibroglandular tissue area, there are two algorithms having nearly identical performance, i.e. multilevel thresholding and maximum entropy. The values for parameter PME, RAE and MHD for multilevel thresholding respectively are 13.34%; 53.34% and 1.47, while for the maximum entropy is 11.27%; 51.26% and 33.92.

TABLE V. AVERAGE PERFORMANCE OF THE THRESHOLDING ALGORITHM

Algorithm	Breast area		
	PME	RAE	MHD
Zack	0.3%	0.7%	0.01
Otsu	16.7%	21.1%	0.42
Multi level	45.3%	68.8%	4.58
Max Entropy	44.4%	66.7%	18.63
Min Entropy	14.3%	17.1%	0.33
Algorithm	Fibroglandular area		
	PME	RAE	MHD
Zack	49.3%	73.0%	4.24
Otsu	32.3%	62.3%	3.21
Multi level	13.3%	53.3	1.47
Max Entropy	11.3%	51.3%	33.9
Min Entropy	34.7%	64.9%	3.4

IV. CONCLUSION

The comparison results of the thresholding algorithm performance are designated for two different purposes, i.e. to obtain the areas of breast and fibroglandular. By the virtue of the comparison results of the thresholding algorithm performance by using the three parameters of PME, RAE and MHD, it shows that Zack algorithm has the best performance to obtain the breast area. Meanwhile, to obtain fibroglandular tissue area, there are two thresholding algorithms having the best performance, i.e. multilevel thresholding and maximum entropy. The obtained results suggest that zack algorithm is perfectly suited for getting breast area than multilevel thresholding and maximum entropy for getting fibroglandular tissue. Further research needs to be conducted to improve the performance of the thresholding algorithm in obtaining fibroglandular tissue area using such as fuzzy c-partition entropy or some methods of intelligent system.

REFERENCES

[1] S. Uyun, Dissertation : *Computation Model on the Pattern and the Percentage of Mammographic Density for Determining the Risk Level of Breast Cancer*. Department of Computer Science and Electronics Faculty of Mathematics and Natural Sciences Universitas Gadjah Mada Yogyakarta, Indonesia, 2014.

[2] N. Ng. K. H. Jamal, L. M. Looi, D. McLean, A. Zulfiqar, S. P. Tan and S. Ranganathan, Quantitative assessment of breast density from digitized mammograms into Tabar's patterns. *Physics in medicine and biology*,

Vol. 51, No. 22, 2006, pp. 5843-5857.

[3] M. J. Yaffe, Measurement of mammographic density. *Breast Cancer Res*, Vol. 10, No. 3, 2008, pp 209-219.

[4] S. Uyun, S. Hartati, A. Harjoko, Subanar and L. Choridah, Comparison between Automatic and Semiautomatic Thresholding Method for Mammographic Density Classification. *Advanced Materials Research*, Vol. 896, 2014, pp. 672-675.

[5] L. Choridah, Disertasi S3 : *Mammographic Density (Threshold method, estradiol level and polimorfisme estrogen reseptor) as a Predictor of Breast Cancer*, Department of Radiology, Faculty of Medicine Universitas Gadjah Mada Yogyakarta, Indonesia. 2013.

[6] M. Langarizadeh and R. Mahmud, Breast Density Classification Using Histogram-Based Features. *Iranian Journal of Medical Informatics*, Vol. 1, No. 1, 2012, pp. 1-5.

[7] N. Ng. K. H. Jamal, S. Ranganathan and L. K. Tan, Comparison of Computerised Assessment of Breast Density with Subjective BI-RADS Classification and Tabar's Pattern from Two-View CR Mammography. In *World Congress on Medical Physics and Biomedical Engineering 2006* (pp. 1405-1408). Springer Berlin Heidelberg.

[8] R. J. Ferrari, P. R. Rangayyan, R. A. Borges and A. F. Frere, Segmentation of the fibro-glandular disc in mammograms using Gaussian mixture modelling. *Medical and Biological Engineering and Computing*, Vol. 42, No. 3, 2004, pp. 378-387

[9] C. Olsén and A. Mukhdoomi, Automatic segmentation of fibroglandular tissue. In *Image Analysis*, 2007, pp. 679-688, Springer Berlin Heidelberg.

[10] A. Abubaker, R. S. Qahwaji, M. J. Aqel, M. H. Saleh, Average Row Thresholding Method for Mammogram Segmentation. In *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the* (pp. 3288-3291). IEEE.

[11] J. Nagi, S. Abdul Kareem, F. Nagi and S. K. Ahmed, Automated breast profile segmentation for ROI detection using digital mammograms. In *Biomedical Engineering and Sciences (IECBES), 2010*, pp. 87-92, IEEE.

[12] B. C. Ko, J. W. Gim & J. Y. Nam, J. Y. Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake. *Micron*, Vol. 42, No. 7, 2011, pp. 695-705.

[13] J. H. Xue & Y. J. Zhang, Ridler and Calvard's, Kittler and Illingworth's and Otsu's methods for image thresholding. *Pattern Recognition Letters*, Vol. 33, No. 6, 2012, pp. 793-797.

[14] P. S. Liao, T. S. Chen and P. C. Chung, A fast algorithm for multilevel thresholding, *Journal of Information Science and Engineering*, Vol. 17, No. 5, 2001, pp. 713-727.

[15] K. Tang, X. Yuan, T. Sun, J. Yang & S. Gao. An improved scheme for minimum cross entropy threshold selection based on genetic algorithm. *Knowledge-Based Systems*, Vol. 24, No. 8, 2011, pp. 1131-1138.

[16] A. Z. Arifin, A. F. Heddyanna and H. Studiawan, *Image thresholding using ultrafuzziness optimization based on type II fuzzy sets*. In *Instrumentation, Communications, Information Technology, and Biomedical Engineering(ICICI-BME), 2009 International Conference*, pp. 1-6, IEEE.

[17] K. C. Singh, L. M. Satapathy, B. Dash and S. K. Routray, Comparative Study on Thresholding. *International Journal of Instrumentation Control & Automation (IJICA)*, Vol. 1, No. 1, 2011, pp. 73-77.

Using Heavy Clique Base Coarsening to Enhance Virtual Network Embedding

Ashraf A. Shahin^{1,2}

¹College of Computer and Information Sciences,
Al Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh, Kingdom of Saudi Arabia

²Department of Computer and Information Sciences, Institute of Statistical Studies & Research,
Cairo University,
Cairo, Egypt

Abstract—Network virtualization allows cloud infrastructure providers to accommodate multiple virtual networks on a single physical network. However, mapping multiple virtual network resources to physical network components, called virtual network embedding (VNE), is known to be non-deterministic polynomial-time hard (NP-hard). Effective virtual network embedding increases the revenue by increasing the number of accepted virtual networks. In this paper, we propose virtual network embedding algorithm, which improves virtual network embedding by coarsening virtual networks. *Heavy Clique* matching technique is used to coarsen virtual networks. Then, the coarsened virtual networks are enhanced by using a refined *Kernighan-Lin* algorithm. The performance of the proposed algorithm is evaluated and compared with existing algorithms using extensive simulations, which show that the proposed algorithm improves virtual network embedding by increasing the acceptance ratio and the revenue.

Keywords—cloud computing; network virtualization; resource allocation; substrate network fragmentation; virtual network embedding; virtual network coarsening

I. INTRODUCTION

In cloud computing data centers, virtualization is employed to accommodate multiple virtual networks (VNs) on a single substrate network (SN), and multiple virtual servers on a single physical server [1]. Consolidating multiple virtual servers from the same virtual network to a single physical server coarsens virtual network down to a few physical servers. Coarsening VN reduces the cost of embedding by eliminating the cost of embedding virtual links between virtual nodes on the same substrate node. Although, effective VN coarsening can improve the utilization of SN's resources and increase the acceptance ratio of VNs and the revenue of infrastructure providers, most of current virtual network embedding algorithms do not take into account VN coarsening [2, 3, 4, 5, 6, 7].

In this paper, we propose virtual network embedding algorithm, which coarsens virtual networks using *Heavy Clique* matching technique. Then, the coarsened virtual networks are enhanced by using a refined *Kernighan-Lin* algorithm. The performance of the proposed algorithm is evaluated and compared with existing algorithms using extensive simulations, which show that the proposed algorithm

improves virtual network embedding by increasing the acceptance ratio and the revenue.

The rest of this paper is organized as follows. Section 2 gives a short overview of related work. Section 3 presents the VN embedding model and problem formulation. Section 4 describes the proposed algorithm. Section 5 evaluates the proposed VN embedding algorithm. Finally, we conclude in section 6.

II. RELATED WORK

In the last few years, many algorithms have been proposed for efficient VNE. VN embedding problem is NP-hard, and finding optimal solution can only be found for small problem instances [8]. Therefore, several heuristic algorithms have been proposed to find a good solution [5, 6, 7, 9]. Some algorithms have been proposed to find exact VNE solutions to be used as optimal bound for the heuristic based VNE solutions [4, 10].

Zhu and Ammar [11] proposed two VN embedding algorithms. In the first algorithm, allocated substrate resources are fixed throughout the VN lifetime. The performance of the first algorithm is improved by using heuristics and adaptive optimization. In the second algorithm, allocated substrate resources are reconfigured to increase the utilization of the underlying substrate resources. However, the proposed algorithms deal only with VNRs that are previously known and do not deal with VNRs that dynamically arrive over time.

In [12], Lischka and Karl proposed online VNE algorithm, which maps nodes and links during the same stage. The proposed algorithm maps VN to a sub-physical network that is similar to the topology of the VN and achieves previously defined constraints (e.g. CPU capacity, link bandwidth). During nodes mapping process, virtual nodes are sorted in descending order based on its required CPU and mapped sequentially to substrate nodes without allowing coexisting multiple virtual nodes from the same VN on one substrate node. To minimize the mapping cost, virtual links are mapped to substrate paths with minimal hops by incrementally increasing the maximum hop limit. However, the computational complexity of the proposed algorithm is high due to multiple operations. In [13], Di et al. improved performance and complexity of the proposed algorithm in [12] by considering the cost of mapping links during the process of

sorting virtual nodes and choosing the maximal hop limit. Fischer et al. [3] modified the algorithm proposed in [12] to consider energy efficiency during nodes and links mapping. Fischer et al. allowed mapping several virtual nodes of the same virtual network to the same substrate node. Although, they take into account the energy efficiency during consolidating virtual nodes, they did not consider the mapping cost.

In [10], Cheng et al. proposed two-stage VN embedding algorithm, called RW-MaxMatch, which ranks nodes using topology-aware node ranking technique to reflect the topological structure of the VNs and the SN. However, RW-MaxMatch algorithm maps nodes without considering its relation to the link mapping, which leads to high consumption of the underlying SN's resources. This is due to mapping neighboring virtual nodes widely separated in the SN.

In [10], Cheng et al. improved the coordination between nodes and links mapping in the RW-MaxMatch algorithm by proposing RW-BFS algorithm. RW-BFS algorithm is a backtracking one-stage VN embedding algorithm, which maps nodes and links at the same stage. In [14, 15], Zhang et al. proposed two VN embedding models: an integer linear programming model and a mixed integer-programming model. To solve these models, Zhang et al. proposed an enhanced version of the MaxMatch algorithm, called RW-PSO algorithm, based on particle swarm optimization. RW-PSO algorithm reduces the time complexity of the link mapping stage by using shortest path algorithm and greedy k-shortest paths algorithm.

To improve the coordination between nodes mapping stage and links mapping stage, Chowdhury et al. [16, 17] formulated the VNE problem as a mixed integer program (MIP), which is NP-hard. To obtain polynomial-time solvable algorithms, they relaxed the integer program to linear program, and proposed two VNE algorithms: D-ViNE (deterministic VNE algorithm) and R-ViNE (randomized VNE algorithm). Nogueira et al. [18] proposed heuristic-based VN embedding algorithm to deal with the heterogeneity of VNs and SN, in both links and nodes. The proposed algorithm is one stage VNE algorithm.

Some of existing works proposed VN embedding algorithms to embed VNRs in distributed cloud computing environments [19, 20, 21, 22]. Houidi et al. [23] proposed exact and heuristics VN embedding algorithms, which split virtual network requests using max-flow min-cut algorithms and linear programming techniques. Leivadeas et al. [24] proposed VN embedding algorithm based on linear programming.

The proposed algorithm partitions VNRs using partitioning approach based on Iterated Local Search. Houidi et al. [25] proposed distributed VN embedding algorithm, which is performed by agent-based substrate nodes. The authors proposed VN embedding protocol to allow communication between the agent-based substrate nodes. However, the proposed algorithm deals only with the offline VN embedding problem.

III. VIRTUAL NETWORK EMBEDDING MODEL AND PROBLEM FORMULATION

Substrate network (SN): We model the substrate network as a weighted undirected graph $G_s = (N_s, L_s)$, where N_s is the set of substrate nodes and L_s is the set of substrate links. Each substrate node $n_s \in N_s$ is weighted by the CPU capacity, and each substrate link $l_s \in L_s$ is weighted by the bandwidth capacity. Fig. 1(b) shows a simple SN example, where the available CPU resources are represented by numbers in rectangles and the available bandwidths are represented by numbers over the links.

Virtual network (VN): virtual network VN_i is modeled as a weighted undirected graph $G_{v_i} = (N_{v_i}, L_{v_i})$, where N_{v_i} is the set of virtual nodes and L_{v_i} is the set of virtual links. Virtual nodes and virtual links are weighted by the required CPU and bandwidth, respectively. Fig. 1(a) shows an example of VN with required CPU and bandwidth.

Virtual network requests (VNR): the i^{th} VN request vnr_i in the set of all VN requests VNR is modeled as $(G_{v_i}, t_{a_i}, t_{l_i})$, where G_{v_i} is the required VN to be embedded, t_{a_i} is the arrival time, and t_{l_i} is the lifetime. When vnr_i arrives, substrate nodes' CPU and substrate links' bandwidth are allocated to achieve the vnr_i . If the substrate network does not have enough resources to achieve vnr_i , vnr_i is rejected. At the end of vnr_i lifetime, all allocated resources to vnr_i are released.

Virtual Network Embedding (VNE): embedding VN_i on SN is defined as a map $M: G_{v_i} \rightarrow (N'_s, P'_s)$, where $N'_s \subseteq N_s$, and $P'_s \subseteq P_s$, where P_s is the set of all loop free substrate paths in G_s . Embedding VN_i can be decomposed into node and link mapping as follows:

$$\text{Node mapping: } M_N: N_{v_i} \rightarrow N'_s$$

$$\text{Link mapping: } M_L: L_{v_i} \rightarrow P'_s$$

For example, mapping of the VN in Fig. 1(a) on SN in Fig. 1(b) can be decomposed into:

$$\text{Node mapping: } \{a \rightarrow B, b \rightarrow A, c \rightarrow C\}$$

$$\text{Link mapping: } \{(a, b) \rightarrow \{(B, A)\}, (b, c) \rightarrow \{(A, D), (D, C)\}, (c, a) \rightarrow \{(C, B)\}\}$$

Virtual Network Embedding Revenue: as in [8, 10, 14], the revenue of embedding vnr_i at time t is defined as the sum of all required substrate CPU and substrate bandwidth by vnr_i at time t .

$$R(vnr_i, t) = Life(vnr_i, t) \cdot \left(\sum_{n_{v_i} \in N_{v_i}} CPU(n_{v_i}) + \sum_{l_{v_i} \in L_{v_i}} BW(l_{v_i}) \right)$$

Where $CPU(n_{v_i})$ is the required CPU for the virtual node n_{v_i} , $BW(l_{v_i})$ is the required bandwidth for the virtual link l_{v_i} , and $Life(vnr_i, t) = 1$ if vnr_i is in its lifetime and substrate resources are allocated to it, otherwise $Life(G_{v_i}, t) = 0$.

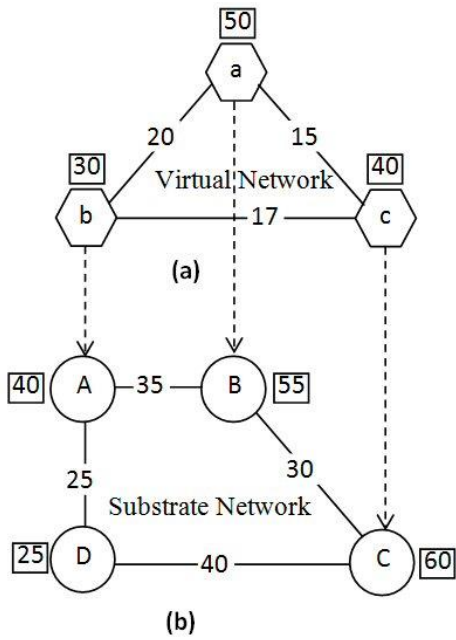


Fig. 1. Example 1 of VNE

Virtual Network Embedding Cost: as in [8, 10, 14], the cost of embedding vnr_i at time t is defined as the sum of all allocated substrate CPU and substrate bandwidth to vnr_i at time t .

$$Cost(vnr_i, t) = Life(vnr_i, t) \cdot \left(\sum_{n_{v_i} \in N_{v_i}} CPU(n_{v_i}) + \sum_{l_{v_i} \in L_{v_i}} BW(l_{v_i}) \cdot Length(M_{L_{v_i}}(l_{v_i})) \right) \quad (1)$$

Where $Length(M_{L_{v_i}}(l_{v_i}))$ is the length of the substrate path that the virtual link l_{v_i} is mapped to.

Objectives: the main objectives are to increase the revenue and decrease the cost of embedding virtual networks in the long run. To evaluate the achievement of these objectives, we use the following metrics:

- *The long-term average revenue*, which is defined by

$$\lim_{T \rightarrow \infty} \left(\frac{\sum_{t=0}^T \sum_{i=1}^I R(vnr_i, t)}{T} \right) \quad (2)$$

Where $I = \|VNR\|$, and T is the total time.

- *The VNR acceptance ratio*, which is defined by

$$\frac{\|VNR_s\|}{\|VNR\|}$$

Where VNR_s is the set of all accepted virtual network requests.

- *The long term R/Cost ratio*, which is defined by

$$\lim_{T \rightarrow \infty} \left(\frac{\sum_{t=0}^T \sum_{i=1}^I R(vnr_i, t)}{\sum_{t=0}^T \sum_{i=1}^I Cost(vnr_i, t)} \right) \quad (4)$$

IV. THE PROPOSED ALGORITHM

In this section, we describe the motivation behind the proposed algorithm and describe the details of the proposed algorithm, which is called HCM-VNE algorithm.

A. Motivation

VN embedding cost (defined by equation 1) depends on allocated substrate CPU and allocated substrate bandwidth. VN embedding cost can be reduced by minimizing these resources. However, minimizing allocated substrate CPU may violate service level agreement and reduce the quality of the service provided to the customers. Allocated substrate bandwidth can be reduced by increasing the number of virtual links between virtual nodes that are mapped to the same substrate node. VN embedding cost is reduced by eliminating the cost of embedding such virtual links. However, finding VN embedding solution with maximum number of eliminated virtual links is not easy task. For example, to map VN in Fig. 2(a) to SN in Fig. 2(b), Fig. 2 shows the mapping solution with the maximum number of eliminated virtual links among other solutions. This solution can be reached by finding sub-VNs that are close to be clique and map each sub-VN to one substrate node. This example motivates us to propose HCM-VNE algorithm, which coarsens VNs using heavy clique matching technique before mapping it.

B. The HCM-VNE algorithm

Algorithm 1 shows the steps of the proposed HCM-VNE algorithm. In line 1, CPU_{max} , which is the upper bound of the coarsened node CPU, is set to the maximum available CPU in SN. In line 2, the upper bound of the total coarsened node bandwidth, BW_{max} , is set to the maximum available bandwidth in SN. VNs are coarsened using *Coarsening()* function and coarsened VNs are optimized using *Optimize()* function. *Coarsening()* function and *Optimize()* function will be described later on. The HCM-VNE algorithm constructs breadth-first searching tree for the graph of the coarsened VN. The root node of the constructed tree is the coarsened virtual node with the largest resources (sum of CPU and BW). Nodes in each level in the created breadth-first searching tree are sorted in descending order based on their resources. Finally, in line 8, the HCM-VNE algorithm embeds coarsened VN on SN using *Embed()* function.

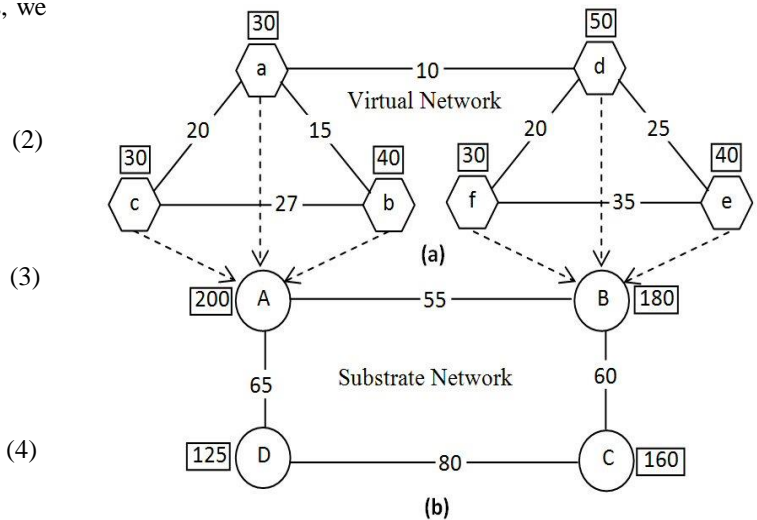


Fig. 2. Example 2 of VNE

ALGORITHM 1: The details of the HCM-VNE algorithm

INPUTS:

$G_v = (N_v, L_v)$: VN to be embed
 $G_s = (N_s, L_s)$: SN to embed on
 Max_hops : maximum allowed substrate path length
 $Max_backtrack$: upper bound of nodes re-mapping operation

OUTPUTS:

$M(G_v)$: map VN nodes and links to SN's resources
 S_VNE : VN embedding success flag

Begin

```

1:  $CPU_{max} = \text{Max}_{n_s \in N_s} (CPU(n_s))$ 
2:  $BW_{max} = \text{Max}_{n_s \in N_s} (\sum_{l_s \in L'_s} BW(l_s))$ ,
   where  $L'_s \subseteq L_s$  and  $n_s$  is incident in each  $l_s \in L'_s$ 
3:  $G_c = \text{Coarsening}(G_v, CPU_{max}, BW_{max})$ 
4:  $G_c = \text{Optimize}(G_c, CPU_{max}, BW_{max})$ 
5: Build breadth-first searching tree of  $G_c$  from coarsened virtual node with
   largest resources.
6: Sort all nodes in each level in the created breadth-first tree in descending
   order according to their required resources.
7: backtrack_count=0
8: if Embed( $G_{c_{root}}$ ,  $G_s$ ,  $M(G_v)$ ) then
9:    $S\_VNE = true$ 
10:  return
11: else
12:    $S\_VNE = false$ 
13:  return
14: end if

```

End

C. Coarsening() function

Virtual networks are coarsened using heavy clique matching technique. A clique in undirected graph is a fully connected subgraph. The cost of embedding VNs is reduced by embedding each sub-VN that is close to clique on one substrate node.

To determine how close sub-VN $G'_v = (N'_v, L'_v)$ is to a clique, we define link density $L_{density}(G'_v)$ as

$$L_{density}(G'_v) = 2\|L'_v\| / (\|N'_v\|(\|N'_v\| - 1))$$

If the sub-VN G'_v is clique (or fully connected), the number of edges is equal to $(\|N'_v\|(\|N'_v\| - 1))/2$ and the link density $L_{density}(G'_v)$ goes to one. $L_{density}(G'_v)$ is small if the sub-VN G'_v is far from being clique.

Algorithm 2 shows the details of the *Coarsening()* function. Coarsening process is iterative and starts with an initial coarsening graph $G_c = (N_c, L_c)$, which is created and initialized by creating coarsened node for each virtual node and coarsened link for each virtual link. Each coarsened node $n_{c_i} \in N_c$ can be considered as a sub-VN $G_{v_i} = (N_{v_i}, L_{v_i})$, where $N_{v_i} \subseteq N_v$ (at this time each N_{v_i} contains only one virtual node), and $L_{v_i} \subseteq L_v$, such that each virtual link $l_{v_i} \in L_{v_i}$ connects two virtual nodes in N_{v_i} . Each coarsened link $l_{c_i} \in L_c$ between two coarsened nodes is a set of virtual links connect virtual nodes in these coarsened nodes. Each virtual node exists in exactly one coarsened node, and each virtual link exists in exactly one coarsened node or one coarsened link. For example, VN in Fig. 2(a) can be coarsened as in Fig. 3.

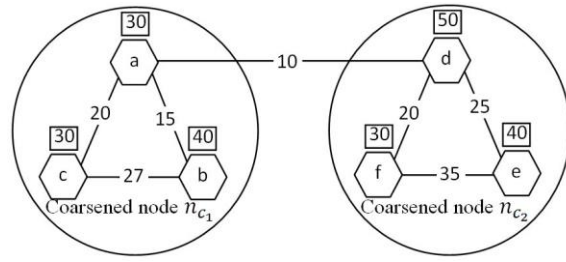


Fig. 3. Coarser VN for the VN in Fig. 2(a)

The graph of the coarsened VN in Fig. 3 is

$G_c = (\{n_{c_1}, n_{c_2}\}, \{l_{v_1}\})$, where

$n_{c_1} = (\{a, b, c\}, \{(a, b), (b, c), (c, a)\})$,

$n_{c_2} = (\{d, e, f\}, \{(d, e), (e, f), (f, d)\})$, and

$l_{v_1} = \{(a, d)\}$

In *Coarsening()* function, coarsened nodes are visited in a sequential way, and each unmatched coarsened node n_{c_i} is matched with its unmatched neighbor n_{c_j} such that the new coarsened node created by combining n_{c_i} and n_{c_j} achieves the CPU and BW constraints and its $L_{density}$ is the largest among all possible coarsened nodes created by combining n_{c_i} with other unmatched neighbors. If such neighbor exists, we add coarsened node n_{c_i} with its neighbor n_{c_j} to the matching list $M_{List} = \{(n_{c_i}, n_{c_j}) \mid n_{c_i}, n_{c_j} \in N_c \text{ and } n_{c_i}, n_{c_j} \text{ are matched}\}$. At the end of each iteration, coarser graph G_c is updated by combining each pair in M_{List} to a new coarsened node. If M_{List} is empty the *Coarsening()* function terminates.

D. Optimize() function

Coarsening() function coarsens VN in Fig. 2(a) as in Fig. 3. However, *Coarsening()* function combines coarsened nodes only based on link density and does not consider the required bandwidth for each virtual link, which sometimes increases the cost of VN embedding. For example, if the virtual link (a, d) in Fig. 3 has bandwidth equal to 50, coarser VN can be improved by moving the virtual node a from the coarsened node n_{c_1} to the coarsened node n_{c_2} . Fig. 4 shows the optimized coarsened VN.

To optimize coarsened VN, we used a refined Kernighan-Lin (KL) algorithm. In 1970, Kernighan-Lin (KL) algorithm was proposed by Kernighan and Lin for graph partitioning problem. Kernighan-Lin (KL) algorithm partitions graph into two parts with equal sizes and with minimal number of cutting edges. It starts with an initial bipartition of the graph and searches for two subsets of vertices from each part of the graph, such that they have the same number of vertices and swapping them improves the cost of the partition. Kernighan-Lin algorithm swaps the selected subsets and repeats the entire process until no such subsets found [26]. However, standard Kernighan-Lin algorithm deals only with typical graph partitioning problem, so it is not directly applicable to optimize coarsened VNs, which may be partitioned to more than two partitions with different sizes.

ALGORITHM 2: The details of the *Coarsening()* function

INPUTS:

G_v : VN graph to be coarsened
 CPU_{max} : the upper bound of the coarsened node CPU
 BW_{max} : the upper bound of the total coarsened node BW

OUTPUTS:

G_c : coarsened VN graph

Begin

```

1: Create and initialize coarsening graph  $G_c = (N_c, L_c)$ 
2: Create new matching list  $M_{List}$ 
3: while(true)
4:   for each unmatched coarsened node  $n_{c_i} \in N_c$ 
5:     Find unmatched neighbor  $n_{c_j} \in N_c$  such that
         $CPU(n_{c_i} \cup n_{c_j}) \leq CPU_{max}$ ,
         $BW(n_{c_i} \cup n_{c_j}) \leq BW_{max}$ , and
         $L_{density}(n_{c_i} \cup n_{c_j}) = \text{Max}_{n_{c_k} \in N'_c} (L_{density}(n_{c_i} \cup n_{c_k}))$ ,
        where  $N'_c$  is the set of all neighbors of the node  $n_{c_i}$ 
6:     Add  $(n_{c_i}, n_{c_j})$  to  $M_{List}$ 
7:   end for
8:   if  $M_{List} == \emptyset$  then
9:     break
10:  else
11:    Update  $G_c$  by combining each pair in  $M_{List}$ 
12:     $M_{List} = \emptyset$ 
13:  end if
14: end while

```

End

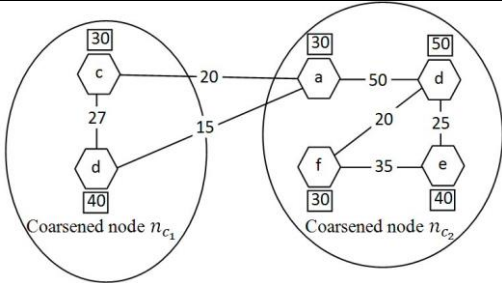


Fig. 4. Optimized coarser VN for the coarser VN in Fig. 3

To optimize coarsened VN, we redefined Kernighan-Lin (KL) algorithm as shown in algorithm 4. *Optimize()* function starts with the partition performed by the *Coarsening()* function and moves boundary virtual nodes between coarsened nodes to improve edge-cut, such that this movement does not violate the CPU and BW constraints. Virtual node is called boundary node, if it is connected to virtual nodes outside its coarsened node. For example, in Fig. 3, virtual node *a* is a boundary virtual node for the coarsened node n_{c_1} , because it has virtual link to the virtual node *d*, which is not in the coarsened node n_{c_1} .

If moving the selected boundary virtual node to the target coarsened node violates the CPU or BW constraints, we try to find one or more boundary nodes in the target coarsened node to be swapped with the selected boundary virtual node. If no such boundary virtual nodes found, we postpone this movement and recheck it again in the next iteration. The whole process is repeated until no movements are performed.

ALGORITHM 3: The details of the *Optimize()* function

INPUTS:

$G_c = (N_c, L_c)$: coarsened VN to be optimized
 CPU_{max} : upper bound of the coarsened node CPU
 BW_{max} : upper bound of the Total coarsened node BW

OUTPUTS:

G_c : optimized coarsened VN

Begin

```

1: Terminate=false
2: while (NOT Terminate)
3:   Terminate=true
4:   for each  $n_{c_i} \in N_c$ 
5:     for each boundary virtual node  $n_v \in n_{c_i}$ 
6:       if  $\exists n_{c_j} \in N_c$ , such that  $\sum_{l_v \in l'_v} BW(l_v) > \sum_{l_v \in l''_v} BW(l_v)$ ,
           where
            $l'_v$  is the set of all virtual links between  $n_v$  and virtual nodes in  $n_{c_j}$ ,
           and
            $l''_v$  is the set of all virtual links between  $n_v$  and virtual nodes in  $n_{c_i}$ 
7:       then
8:         if moving  $n_v$  from  $n_{c_i}$  to  $n_{c_j}$  does not violate CPU and BW
           constraints.
9:         then
10:          Move  $n_v$  from  $n_{c_i}$  to  $n_{c_j}$ 
11:          Terminate=false
12:        else
13:          Find set of boundary virtual nodes  $n'_v$  in the coarsened node
            $n_{c_j}$ , such that swapping  $n_v$  and  $n'_v$  improves bandwidth and
           does not violate CPU and BW constraints.
14:          if such node found swap them
15:            Terminate=false
16:          end if
17:        end if
18:      end if
19:    end for
20:  end for
21: end while

```

End

E. Embed() function

The *Embed()* function embeds coarsened VN on SN as described in algorithm 4. In the *Embed()* function, candidate substrate node list for each coarsened virtual node is built by collecting all substrate nodes that have available CPU capacity at least as large as the coarsened virtual node CPU and have a loop free substrate path to each substrate node contains one of the previously mapped neighbors. Each substrate path should satisfy the constraint of the maximum substrate path length, and have available bandwidth greater than or equal the bandwidth of the coarsened virtual link between the coarsened virtual node and its previously mapped neighbor.

Candidate substrate nodes for each coarsened virtual node are collected by creating a breadth-first search tree from each substrate node contains one of the previously mapped neighbors, and finding the common substrate nodes between the created trees. In the constructed trees, substrate nodes should satisfy the CPU constraints for coarsened virtual node, and substrate paths should satisfy the connectivity constraints to connect the coarsened virtual node with its neighbors. By this way, all candidate substrate nodes in the candidate

substrate node list satisfy all constraints (CPU and connectivity constraints).

Substrate nodes in the candidate substrate node list are sorted in ascending order according to the total cost of embedding coarsened virtual links from the coarsened virtual node to all previously embedded neighbors. If the coarsened virtual node is a root node, the candidate substrate node list is a set of all substrate nodes that have enough resources to embed the coarsened virtual node. The candidate substrate nodes for the root are sorted in descending order according to the total available resources.

Coarsened virtual node is sequentially mapped to substrate nodes in its candidate substrate node list. If there is no appropriate substrate node in its candidate substrate node list, we backtrack to the previously mapped node, re-map it to the next candidate substrate node, and continue to the next node. In line 3, mappings of the coarsened virtual node and its coarsened virtual links are added to $M(G_v)$ by using the function $Add()$. To map coarsened node n_{c_i} to substrate node n_s , the function $Add()$ adds maps from each virtual node in n_{c_i} to the substrate node n_s . All virtual links in the coarsened node n_{c_i} are mapped to substrate paths with length zero from the substrate node n_s to itself. For each coarsened link from n_{c_i} to one of the previously mapped coarsened nodes, the function $Add()$ adds maps for all virtual links in these coarsened links. Virtual links are mapped to shortest loop free substrate paths, which are specified by breadth-first search manner. In line 6, $Delete()$ function is used to perform the backtracking process.

ALGORITHM 4: The details of $Embed()$ Function

INPUTS:

n_{c_i} : current coarsened virtual node to be embedded
 G_s : substrate network to embed on
 $M(G_v)$: map of the previously mapped nodes and links

OUTPUTS:

$M(G_v)$: updated map
 S_{VNE} : VN embedding success flag

Begin

```
1: Build candidate substrate node list  $C_i$  for  $n_{c_i}$ 
2: for each  $n_s$  in  $C_i$ 
3:   Add( $(n_{c_i}, n_s), M(G_v)$ )
4:   if Embed( $n_{c_{i+1}}, G_{s_i}, M(G_v)$ ) then return true
5:   else
6:     Delete( $(n_{c_i}, n_s), M(G_v)$ )
7:   end if
8: if backtrack_count > Max_backtrack then return false
9: end for
10: backtrack_count ++
11: return false
End
```

V. PERFORMANCE EVALUATION

We evaluated the proposed HCM-VNE algorithm by comparing its performance with some of existing algorithms.

First, we implemented three algorithms: HCM-VNE, RW-MaxMatch [15], and RW-BFS [10]. Second, we generated SN topology and 3000 VN topologies to be used as inputs to the implemented algorithms. Finally, we compared the results from the implemented algorithms. In the following sub-sections, we describe the evaluation environment settings and discuss the results of the simulations.

A. Evaluation environment settings

In our evaluation, the substrate network topology is configured to have 200 nodes with 1000 links. Substrate network is generated using Waxman generator. Bandwidths of the substrate links are real numbers uniformly distributed between 50 and 100 with average 75. We have selected two server configurations: HP ProLiant ML110 G4 (Intel Xeon 3040, 2 cores X 1860 MHz, 4 GB), and HP ProLiant ML110 G5 (Intel Xeon 3075, 2 cores X 2660 MHz, 4 GB). Each substrate node is randomly assigned one of these server configurations.

Virtual network topologies are generated using Waxman generator with average connectivity 50%. Number of virtual nodes in each VN is variant from 2 to 20. Each virtual node is randomly assigned one of the following CPU: 2500 MIPS, 2000 MIPS, 1000 MIPS, and 500 MIPS, which are correspond to the CPU of Amazon EC2 instance types. Bandwidths of the virtual links are real numbers uniformly distributed between 1 and 50. VN's arrival times are generated randomly with arrival rate 10 VNs per 100 time units. The lifetimes of the VNRs are generated randomly between 300 and 700 time units with average 500 time units. 3000 VN topologies are generated and stored in brite format. For each algorithm, we run the simulation for 30000 time units with the previously generated VNRs¹. For all algorithms, we set the maximum allowed hops (Max_hops) to 2, and the upper bound of remapping process (Max_backtrack) to $3n$, where n is the number of nodes in each VNR.

B. Evaluation results

Three metrics have been used to evaluate the performance of the proposed algorithms: *the long-term average revenue*, which is defined by Equation (2), *the VNR acceptance ratio*, which is defined by Equation (3), and *the long-term R/Cost ratio*, which is defined by Equation (4). Fig. 5 shows the simulation results using the VNR acceptance ratio to compare the different VNE algorithms. It can be seen that the proposed algorithm that coarsened VNs using heavy clique matching increases the acceptance ratio compared with other algorithms. For example, at time unit 30000, in Fig. 5, the VNR acceptance ratio for the RW-BFS and RW-MaxMatch are 20 and 16 percent, while the VNR acceptance ratio for the HCM-VNE is 53 percent. In other words, the proposed algorithm can embed more VNs on the same SN at the same time. Consequently, the proposed algorithm increases the long-term average revenue compared with other algorithms, as shown in figure 6.

¹The generated SN topology, generated VNRs topologies, and outputs are available online at (https://drive.google.com/folderview?id=0BxEBmTQ0WG5RcnBYLVZhdW42bjg&usp=drive_web)

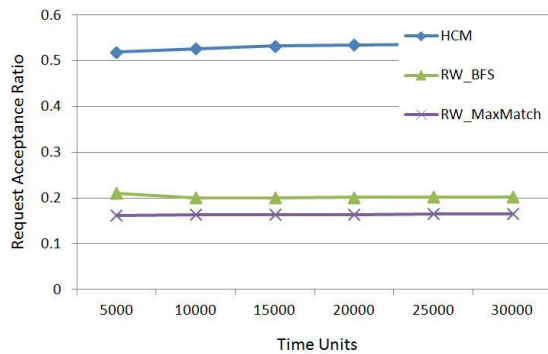


Fig. 5. The VNR acceptance ratio comparison

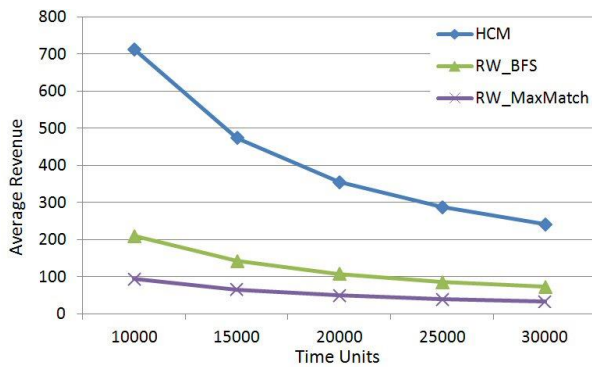


Fig. 6. The long-term average revenue comparison

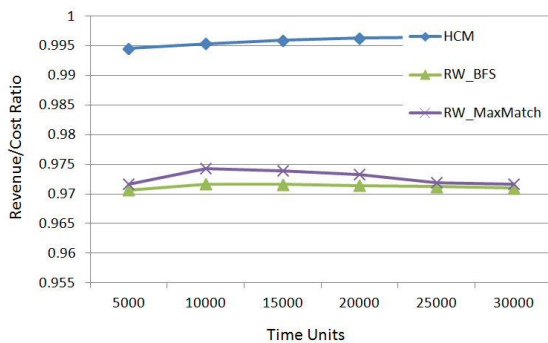


Fig. 7. The long-term Revenue/Cost ratio comparison

For example, at time unit 30000, the average revenue for the RW-BFS and RW-MaxMatch are 72 and 33, while the average revenue for the HCM-VNE is 240. As shown in Fig. 7, the long-term Revenue/Cost ratio of all algorithms are nearly the same, but the proposed algorithm performs slightly better than other algorithms.

VI. CONCLUSION

In this paper, we proposed virtual network embedding algorithm, which coarsens virtual networks using heavy clique matching and optimizes the coarser virtual networks by applying a refined Kernighan-Lin (KL) algorithm. The proposed algorithm coarsens sub-virtual networks that are close to clique and embeds each sub-virtual network to substrate node. The cost of embedding virtual networks is reduced by eliminating the cost of embedding virtual links between virtual nodes on the same substrate node. Performance

of the proposed algorithm has been evaluated and compared with some of the existing algorithms using extensive simulations. Extensive simulation experiments show that the proposed algorithm increases the acceptance ratio and the revenue. For the future work, we plan to investigate other coarsening techniques (e.g. Random Matching and Light Edge Matching) to find the best coarsening technique, which increases the acceptance ratio and the revenue while decreasing the embedding cost.

REFERENCES

- [1] I. Fajjari, N. Aitsaadi, and G. Pujolle, "Cloud networking: An overview of virtual network embedding strategies," in Global Information Infrastructure Symposium, 2013, Oct 2013, pp. 1–7.
- [2] S. Su, Z. Zhang, X. Cheng, Y. Wang, Y. Luo, and J. Wang, "Energy-aware virtual network embedding through consolidation," in 2012 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), March 2012, pp. 127–132.
- [3] A. Fischer, M. Beck, and H. de Meer, "An approach to energy-efficient virtual network embeddings," in 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013), May 2013, pp. 1142–1147.
- [4] J. Botero, X. Hesselbach, M. Duelli, D. Schlosser, A. Fischer, and H. de Meer, "Energy efficient virtual network embedding," Communications Letters, IEEE, vol. 16, no. 5, pp. 756–759, May 2012.
- [5] Z. Zhang, S. Su, X. Niu, J. Ma, X. Cheng, and K. Shuang, "Minimizing electricity cost in geographical virtual network embedding," in 2012 IEEE Global Communications Conference (GLOBECOM), Dec 2012, pp. 2609–2614.
- [6] G. Sun, V. Anand, D. Liao, C. Lu, X. Zhang, and N.-H. Bao, "Power-efficient provisioning for online virtual network requests in cloud-based data centers," IEEE Systems Journal, vol. PP, no. 99, pp. 1–15, 2013.
- [7] C. Ghribi, M. Hadji, and D. Zeglache, "Energy efficient vm scheduling for cloud data centers: Exact allocation and migration algorithms," in 2013 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), May 2013, pp. 671–678.
- [8] A. Fischer, J. Botero, M. Till Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey," Communications Surveys Tutorials, IEEE, vol. 15, no. 4, pp. 1888–1906, Fourth 2013.
- [9] [9] A. Beloglazov and R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," Concurrency and Computation: Practice & Experience, vol. 24, no. 13, pp. 1397–1420, Sep. 2012.
- [10] X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology-aware node ranking," SIGCOMM Comput. Commun. Rev., vol. 41, no. 2, pp. 38–47, Apr. 2011.
- [11] Y. Zhu and M. Ammar, "Algorithms for assigning substrate network resources to virtual network components," in INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings, April 2006, pp. 1–12.
- [12] J. Lischka and H. Karl, "A virtual network mapping algorithm based on subgraph isomorphism detection," in Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures, ser. VISA '09. New York, NY, USA: ACM, 2009, pp. 81–88.
- [13] H. Di, L. Li, V. Anand, H. Yu, and G. Sun, "Cost efficient virtual infrastructure mapping using subgraph isomorphism," in Communications and Photonics Conference and Exhibition (ACP), 2010 Asia, Dec 2010, pp. 533–534.
- [14] Z. Zhang, X. Cheng, S. Su, Y. Wang, K. Shuang, and Y. Luo, "A unified enhanced particle swarm optimization-based virtual network embedding algorithm," Int. J. Communication Systems, vol. 26, no. 8, pp. 1054–1073, 2013.
- [15] X. Cheng, S. Su, Z. Zhang, K. Shuang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology awareness and optimization," Computer Networks, vol. 56, no. 6, pp. 1797 – 1813, 2012.

- [16] N. Chowdhury, M. Rahman, and R. Boutaba, "Virtual network embedding with coordinated node and link mapping," in INFOCOM 2009, IEEE, April 2009, pp. 783–791.
- [17] M. Chowdhury, M. Rahman, and R. Boutaba, "Vineyard: Virtual network embedding algorithms with coordinated node and link mapping," *IEEE/ACM Transactions on Networking*, vol. 20, no. 1, pp. 206–219, Feb 2012.
- [18] J. Nogueira, M. Melo, J. Carapinha, and S. Sargento, "Virtual network mapping into heterogeneous substrate networks," in *Computers and Communications (ISCC)*, 2011 IEEE Symposium on, June 2011, pp. 438–444.
- [19] F. Samuel, M. Chowdhury, and R. Boutaba, "Polyvine: policy-based virtual network embedding across multiple domains," *Journal of Internet Services and Applications*, vol. 4, no. 1, 2013.
- [20] I. Houidi, W. Louati, and D. Zeghlache, "A distributed and autonomic virtual network mapping framework," in *Autonomic and Autonomous Systems*, 2008. ICAS 2008. Fourth International Conference on, March 2008, pp. 241–247.
- [21] Y. Xin, I. Baldine, A. Mandal, C. Heermann, J. Chase, and A. Yumerefendi, "Embedding virtual topologies in networked clouds," in *Proceedings of the 6th International Conference on Future Internet Technologies*, ser. CFI '11. New York, NY, USA: ACM, 2011, pp. 26–29.
- [22] B. Lv, Z. Wang, T. Huang, J. Chen, and Y. Liu, "Virtual resource organization and virtual network embedding across multiple domains," in *Multimedia Information Networking and Security (MINES)*, 2010 International Conference on, Nov 2010, pp. 725–728.
- [23] I. Houidi, W. Louati, W. B. Ameer, and D. Zeghlache, "Virtual network provisioning across multiple substrate networks," *Computer Networks*, vol. 55, no. 4, pp. 1011 – 1023, 2011, special Issue on Architectures and Protocols for the Future Internet.
- [24] A. Leivadeas, C. Papagianni, and S. Papavassiliou, "Efficient resource mapping framework over networked clouds via iterated local search-based request partitioning," *Parallel and Distributed Systems*, IEEE Transactions on, vol. 24, no. 6, pp. 1077–1086, June 2013.
- [25] I. Houidi, W. Louati, and D. Zeghlache, "A distributed virtual network mapping algorithm," in *Communications*, 2008. ICC '08. IEEE International Conference on, May 2008, pp. 5634–5640.
- [26] Y. Weihong, Y. Yuehui, and T. Guozhen, "Recursive kernighan-lin algorithm (rkl) scheme for cooperative road-side units in vehicular networks," in *Parallel Computational Fluid Dynamics*, ser. Communications in Computer and Information Science, K. Li, Z. Xiao, Y. Wang, J. Du, and K. Li, Eds. Springer Berlin Heidelberg, 2014, vol. 405, pp. 321–331.

Ontology Based SMS Controller for Smart Phones

Mohammed A. Balubaid¹, Umar Manzoor²

¹Industrial Engineering Department, Engineering Faculty

²Faculty of Computing and Information Technology,
King Abdulaziz University,
Jeddah, Saudi Arabia

Bassam Zafar², Abdullah Qureshi³, Numairul Ghani³

²Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia.

³National University of Computer and Emerging Sciences,
Islamabad, Pakistan

Abstract—Text analysis includes lexical analysis of the text and has been widely studied and used in diverse applications. In the last decade, researchers have proposed many efficient solutions to analyze / classify large text dataset, however, analysis / classification of short text is still a challenge because 1) the data is very sparse 2) It contains noise words and 3) It is difficult to understand the syntactical structure of the text. Short Messaging Service (SMS) is a text messaging service for mobile/smart phone and this service is frequently used by all mobile users. Because of the popularity of SMS service, marketing companies nowadays are also using this service for direct marketing also known as SMS marketing. In this paper, we have proposed Ontology based SMS Controller which analyze the text message and classify it using ontology as legitimate or spam. The proposed system has been tested on different scenarios and experimental results shows that the proposed solution is effective both in terms of efficiency and time.

Keywords—Short Text Classification; SMS Spam; Text Analysis; Ontology based SMS Spam; Text Analysis and Ontology

I. INTRODUCTION

Mobile phones were initially developed to make and receive calls while being mobile using radio link. Later on, services like text messaging (SMS), multimedia messaging (MMS) were added in the mobile phone devices. In the last two decades, mobile phones have evolved and have become smarter / intelligent devices commonly known as smart phones [1, 2]. Smart phones are built on mobile computing platform and usually have advanced computing abilities / connectivity as compared to the simple mobile phones [8]. Initially smart phones were developed with the integration of mobile phone and personal digital assistant (PDA) functions. These smart phones include number of exciting features like touch screen, mobile web browser (i.e. access websites on mobile phone) and WiFi (i.e. access internet using wireless connection) support.

In year 2000, high resolution touch screen smart phone named Ericsson R380 was released which has its own Operating System. This was first ever smart phone with its own OS, the Operating System used was Symbian OS. In 2005, Google entered into the mobile market with the help of an open source operating system for smart phones called Android. In 2007, Apple [9] introduced a smart phone named iPhone [10] which made big change in the history of smart phones development; Apple development their own Mobile Operating System named as IOS for iPhone and this OS is not open source. Therefore, Android operating system is supported by most of the smart phones companies (such as HTC, Samsung, Sony Ericson).

Along with the launch of iPhone, Apple introduced AppStore (Application Store) where 3rd party applications were hosted for distribution (i.e. single platform distribution). Before, Apple AppStore, smart phone applications distribution were largely dependent on third-party sources that developed the application(s) such as GetJar, Handmark, Handango, PocketGear, etc. Application development for Android OS is greatly increasing as compared to IOS because 1) development toolkit is free, 2) Android is open-source, therefore it's easy to integrate applications and 3) Android software suite allows easy integration with Google applications such as Maps, Calendar, web browser etc. Android based smart phones are giving great competition to iPhone.

Text analysis includes lexical analysis of the text and has been widely studied and used in diverse applications. In the last decade, researchers have proposed many efficient solutions to analyze / classify large text dataset, however, analysis / classification of short text is still a challenge because 1) the data is very sparse 2) It contains noise words and 3) It is difficult to understand the syntactical structure of the text [21, 22, 25, 28]. The concept of Short Messaging Service (SMS) was developed in the Franco-German GSM cooperation in 1984 by Bernard Ghillebaert and Friedhelm Hillebrand [11]. SMS is a text messaging service on the phone, web or mobile system and mostly used data application is SMS text messaging. SMS nowadays is also used for direct marketing also known as SMS marketing.

In the last few years, many SMS managers have been developed for managing the SMS on smart phones and the most of them focuses on Spam filtering, Scheduled SMS and automatic-Reply generation. Few popular android applications are 1) Anti SMS Spam: It is a spam filtering application and spams all incoming SMS from unknown numbers when Spam filtering is turned on. 2) Schedule SMS: It is scheduled SMS application and gives time, date, recipient number and text (SMS content) option to the user. The application sends the SMS to the recipient on specified time and date specified by the user. 3) SMS Auto Reply: It is an Auto Reply SMS application which sends an automated reply to all the incoming texts when auto reply is turned on. The content (text) of auto reply is selected / configured by the user.

Ontologies have been widely used for knowledge representation / sharing and have been used in diverse areas [23, 24, 26, 27]. Ontology based SMS Controller is an Android Based Application developed on Android Jelly Beans 4.1, the proposed solution is all in one SMS manager and includes some previous features with advancements as well as some

new and exciting Features like ontology based SMS spam detection, Group chat etc. The default android messaging application gives few options to user such as send message / receive message / save message etc. whereas the proposed application provides some additional features in addition the default features. The major features of the proposed application are:

- Automated text replies to messages when a profile is activated
- Scheduled SMS sent on specific dates and events
- Group chat including multiple users like we do in different messengers
- Content based Spam filtering

The above features make the proposed application unique as these features are missing in the existing applications. The remainder of this paper is organized as follows. In Section 2, we present brief overview of related work, this section is followed by the discussion of the Ontology based SMS Controller architecture including the SMS text analysis and classification method. In Section 4, the simulation and experimental analysis of proposed solution is presented. Finally, the conclusion is drawn in Section 5.

II. RELATED WORK

With the evolution in Smartphone era, leading IT companies and researchers have proposed many efficient applications for the same. In this section, we will review few related applications developed for managing SMS on android platform.

1) *Anti-Spam SMS and Private Box by Droid Mate [12]* is an android based anti-spam application with a private box. Its spam feature helps filter unwanted messages from any sender. Key features of this application are: a) Can block SMS from unknown numbers, and b) User can create a block list and can add existing contact or new numbers in the block list.

2) *Handcent SMS by Handcent Market [13]* is an android based SMS scheduling application with the following key features: a) It helps schedule SMS/MMS messages at specific times or at regular intervals e.g. daily and b) It supports blacklist (i.e. deletes incoming SMS / MMS from number in blacklist which also helps block spam messages).

3) *SMS Scheduling applications: The best three scheduled SMS android applications are as follows: a) Schedule SMS Wishes enables the user to schedule SMS on the contact list (i.e. the user can select the date / time on which the SMS needs to be send to the selected contact or can use the repeat option to send the SMS regularly [daily / weekly] to the selected contacts) b) Google Voice SMS tool enables the user to schedule (i.e. daily, weekly, etc) the SMS with Google voice c) Scheduled Message application enables the user to schedule SMS or email at any given date / time.*

4) *SMS auto reply by Kirill kruchinkin [14]* is an auto reply application which sends reply to each incoming SMS when the auto reply option is enabled. The user has to

configure the reply to be send when the auto reply option mode is turn on.

5) *Intelligent auto reply by John Tsau [15]* is a rule based SMS application that has Auto Reply and Auto Forwarding features. It automatically replies to the SMS and missed calls according to the rules set by the user.

6) *GO SMS PRO by GO Dev Team [16]* supports the features of scheduled SMS and Group Texting.

The previous applications include most of the exciting features but they have the following limitations:

- What if a user does not want to spam all the SMS from unknown numbers?
- What if a user wants to spam the SMS from some unusual numbers only?
- What if a user wants his application to auto reply to a certain Group?
- What if a user wants to send different replies to different group of recipients?
- What if a user wants to have all these features in one application?

Ontology based SMS Controller has Solutions to all these questions. It auto reply to a certain Group and sends different replies to different group of recipients. It gives solution for detecting spam SMS using content analysis. Above all these features are all integrated in one application so that a user can easily manage all the features from one application. Plus it includes new features like Group chat and Auto scheduled SMS by synchronizing the events in the Calendar. The key features of the proposed application include:

- Auto-Reply Modes
 - Profile Based,
 - Group Based,
 - Group-Profile Based,
- Auto-Messaging Modes
 - Event mode
 - Birthday mode
 - Scheduled Texts
- Group Chat
- Content based SMS Spam detection
- Auto-Message Report.

III. SYSTEM ARCHITECTURE

Ontology based SMS Controller is an Android Based Application developed on Android Jelly Beans4.1 as shown in figure 1 and has the following four modules:

- SMS Spam
- Group Chat
- Auto Reply
- Event-Based Messages

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:



Fig. 1. System Architecture of Ontology based SMS Controller

A. SMS Spam

For each incoming SMS, the Ontology based SMS Controller interrupts the SMS and executes the SMS Spam classification algorithm to verify the spam messages as shown in figure 2. The SMS Spam classification algorithm is comprised of three steps: (I) Pre-processing (II) Content Analysis (III) Spam Classification.

- **Pre-processing:** In the first step, for each incoming SMS, the Ontology based SMS Controller validates the sender number with the spam blacklist numbers, if the number is found in the blacklist numbers, the SMS is send to SPAM folder without further processing. The Ontology based SMS Controller also provides user the option to SPAM all SMS messages from unknown numbers or specific numbers or weird numbers. If this option is selected by the user, all SMS belonging to these categories will be send to SPAM folder without further processing.

In Step 2, all standard stop-list / stemmer words like (“is”, “the”, “on”, “and”, “in”, “with”, “for”, “by”...) are eliminated from the SMS Text. In Step 3, homogeneous words like { (“chat”, “chatting”, “chatted”), (“Advertize”, “Advertizing”, “Advertized”) } are all substituted by the single word “chat” and “Advertize” respectively. Also, multiple entries for each word are eliminated from the SMS text.

- **Content Analysis:** This module uses the filtered SMS text from the previous step which contains n keywords where each keyword can express n possible meanings. In order to assign proper meaning to each keyword, every keyword is compared with every other keyword and most related sense (i.e. semantically related) is selected. To calculate the most related sense the

shortest path (i.e. minimum number of nodes present in the path connecting the keywords is used); WordNet [18] is used for this purpose. In the next step, Concept set is generated which contains either the original keywords or Lowest Super Ordinate (LSO) for each pair of keywords, the selection depends on the parameter h , for more details please see [19].

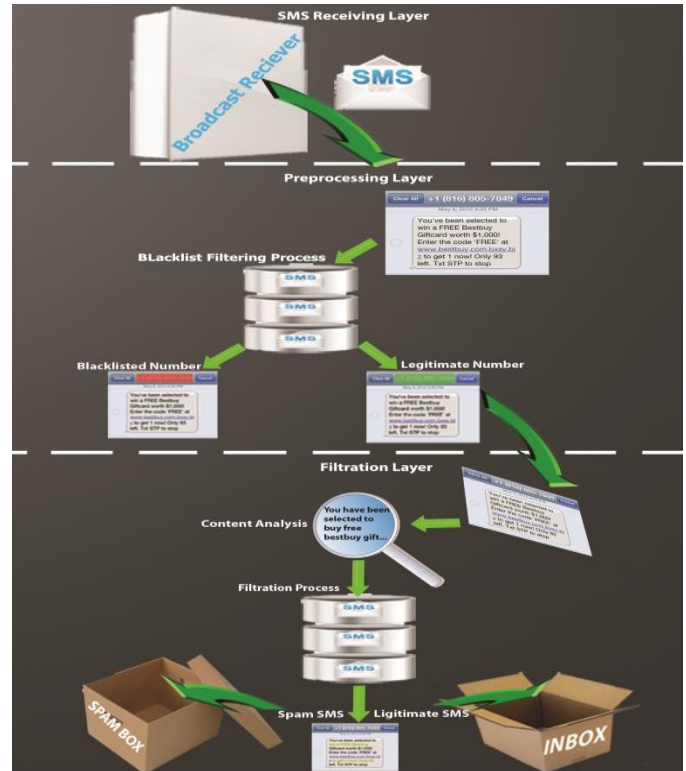


Fig. 2. Ontology based SMS Controller Spam Module

Spam concepts which include spam keywords, their synonym and hypernym are loaded from the ontology and stored in spam set. Each concept is compared with the spam concepts one by one and matches are stored in a separate resultant set with labels O (Original Keyword), S (Keyword Synonym) or H (Keyword Hypernym). The labels are assigned on the basis of comparison; if the concept is matched with spam keyword, O is assigned. Similarly if the concept is match with Spam keyword synonym or hypernym, S or H is assigned respectively. Each concept in the resultant set is assigned a score based on the assigned label (i.e. O=1, S=0.50 and H=0.25). The collective spam score of the resultant set is calculated by adding the all individual concept scores.

- **Classification and Ontology Enhancement:** This module uses the resultant set extracted from the previous step, to classify the SMS as spam or not. Furthermore, if the SMS is classified as spam, the ontology will be enhanced by adding new concepts (i.e. Original keywords, synonym and hypernym). In order to classify the SMS as spam, the Collective Spam score (CS) calculated in the previous set is used, if condition (where is configurable) is satisfied, the SMS is classified as Spam and forwarded to Spam folder. Once the SMS is classified as Spam, Keywords Synonyms

and Hypernyms are extracted from WordNet and new concepts (i.e. Keywords, Synonyms and Hypernyms) are added to the ontology knowledge base.

B. Group Chat

The prerequisite of using this feature is that all participating users should have Ontology based SMS Controller installed on the smart phone. One of the application user has to start the Group chat by sending “join group chat” invitation to the others. Invitation is sent through SMS message and for this purpose a special SMS is sent to the invitee which Ontology based SMS Controller interprets and asks the user to join the group chat. The user can accept or reject the request, if the user accepts, the details of new user is send to all the active members of the group chat and chat window is loaded on the new user’s Smartphone.

Similarly, if any active user during the group chat closes the application, the details of disconnected user is send to all members of the group chat. Each group chat is assigned unique chat code to the same and each SMS message send / received from chat window contains this unique chat code, which makes it easy to identify; to which chat this message belongs. Each member sets a nick at the start of chat, and these are displayed on the chat window instead of the numbers.

Each incoming SMS is interrupted by Ontology based SMS Controller and it validates the type of the SMS message (i.e. Invitation, Chat Message, or Normal Message) and perform actions accordingly. If the SMS is chat message, it forwards the same to the corresponding chat window and delete it from the inbox. If the SMS is invitation SMS, it displays the invitation to the user and wait for the response. Based on user response it either opens the chat windows or sends rejection message. When a user joins or leaves the chat, all other members are informed and the list of chat members is updated accordingly.

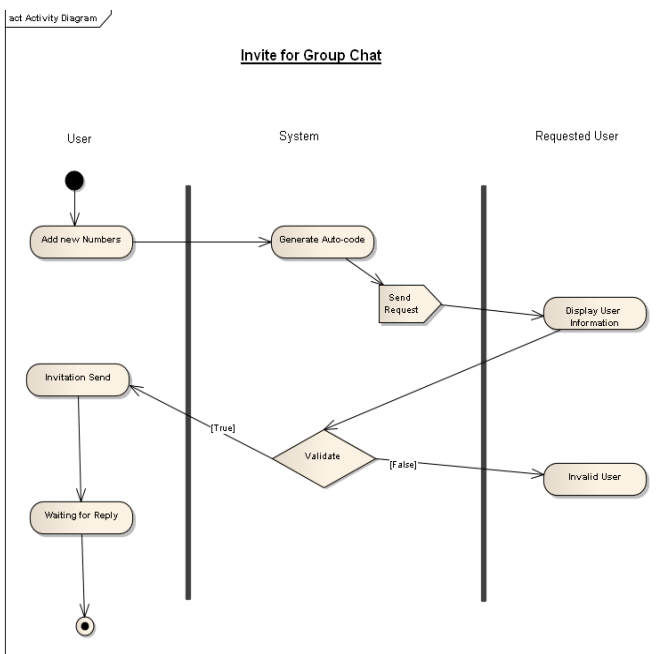


Fig. 3. Group Chat Flow

C. Auto Reply

Ontology based SMS Controller sends auto-reply according to the user-defined profiles; the user is responsible to create auto-reply groups and reply messages for each group. If the auto-reply mode is on and no auto-reply profile is activated, a default auto-reply message is sent otherwise the user defined auto-reply message according to the profile is send. User can create groups and add numbers in these groups from contact list.

D. Event-Based Messages

Ontology based SMS Controller automatically synchronizes itself with the calendar and generates automatic SMS based on Events. User is responsible to define event(s) by setting date / time of the event(s) and the message to be send. User can add group(s) to an event by selecting from list of available groups. Message defined against the event is send to all members of the group(s) associated with the event. Birthday event is predefined in the application, which picks the birthdays of the contacts (if available) and create birthday event for each of them. User can define the birthday message for the birthday event, if no message is defined; default birthday wish is send automatically on respective birthdays.

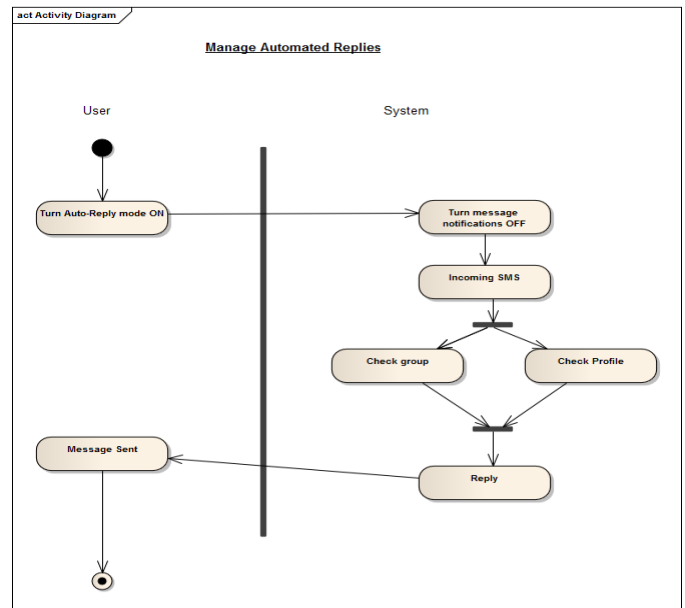


Fig. 4. Auto-Reply flow

IV. SIMULATION AND EXPERIMENTAL RESULTS

The Application is developed on Android version 4.1 (Jelly Beans) and is compatible with all the next versions of Android (min SDK 8). Android Virtual Device (AVD) manager is installed with Android Software Development kit which allows the programmer to create an AVD for specific version of Android. The simulator used for testing and debugging of the proposed application is AVD 4.1.

Initially for the experimentation, we built the ontology concepts using one hundred known spam messages; afterward we tested the proposed solution on large number of SMS, figure 5 shows spam detection percentage over number of SMS, as shown in figure 5 the proposed solution spam

detection percentage over number of SMS increases as ontology knowledgebase is enhanced (i.e. new spam concepts are updated in the ontology).

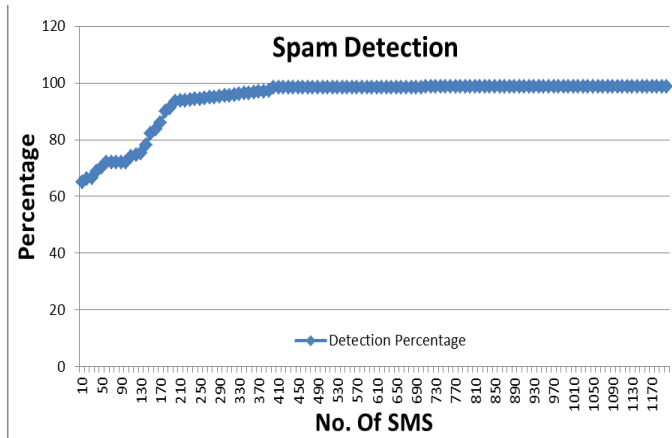


Fig. 5. Spam Detection vs No. of SMS

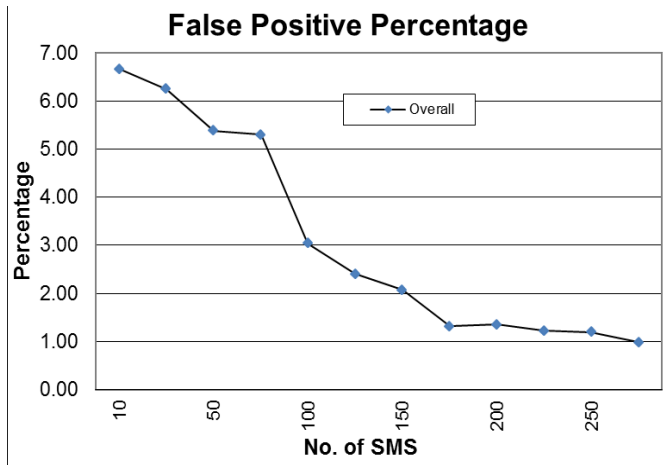


Fig. 6. False Positive Percentage vs No. of SMS

Figure 6 shows false positive percentage over number of SMS, as shown in figure 6 the proposed solution false positive percentage over number of SMS decreases as the system receives feedback from user over wrongly predicted SMS(s) which helps in updating the ontology knowledgebase by removing concepts related to wrongly predicted SMS(s).

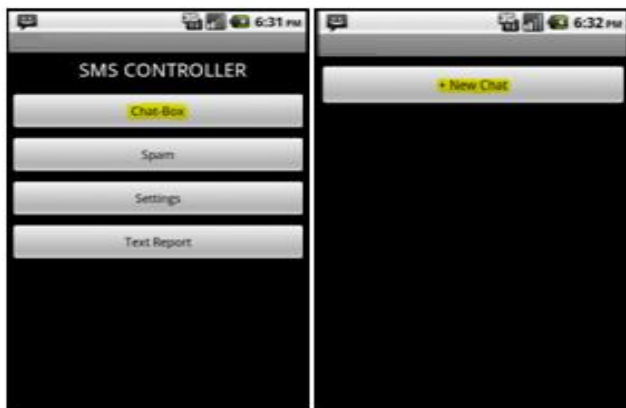


Fig. 7. (a) Main View (b) New Chat Window

Figure 7(a) shows the main view of the application, if user press Chat-Box new window opens as shown in figure 7(b). If user press New Chat button, new chat is started.

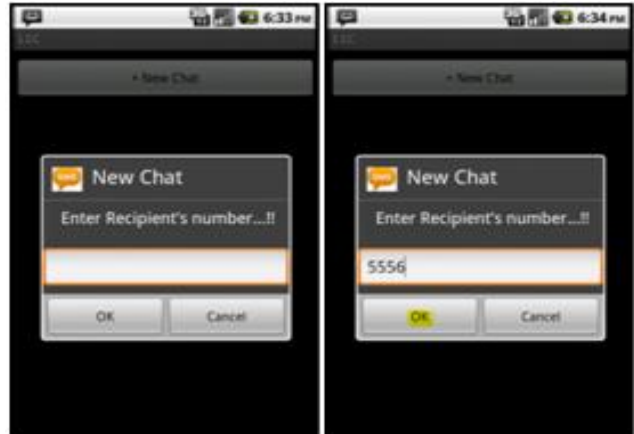


Fig. 8. New Chat

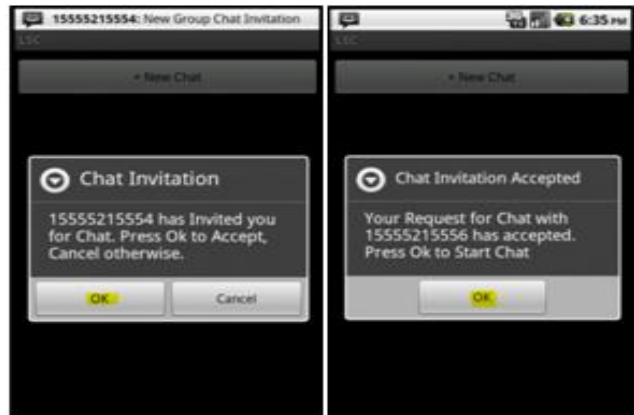


Fig. 9. (a) Chat Invitation (b) Invitation Accepted

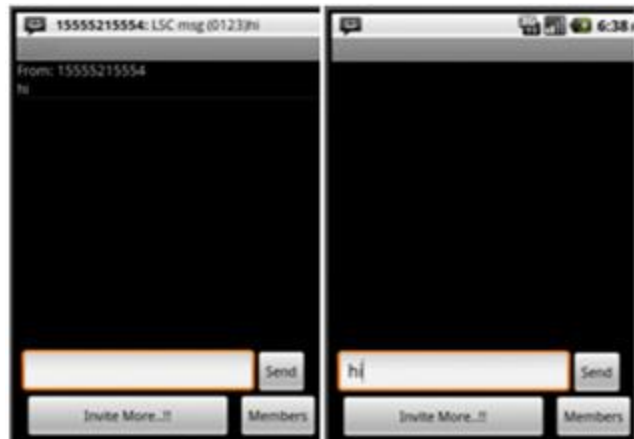


Fig. 10. Group Chat Window

Figure 8 shows new chat window where the user needs to enter the Recipient's number and press ok to invite the same to group chat.

Figure 9(a) shows the invitation for group chat message which appears on the recipient's device, if the user accepts the

invitation figure 9(b) is shown on the sender device. After handshaking the group chat is started and chat window appears on each participating device (i.e. Sender & Receivers) as shown in figure 10. User can write the text in text Box and use the send button to send the message. The message is sent to all recipients by using normal SMS and is shown on each participating device. New members can be added at any time using the invite more feature.



Fig. 11. (a) Group title (b) Group contacts

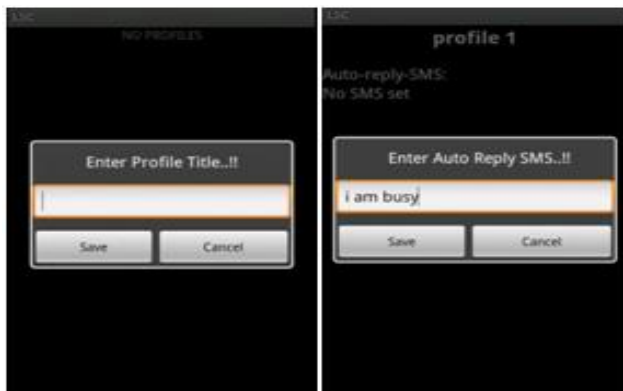


Fig. 12. (a) Profile title (b) Auto-reply SMS

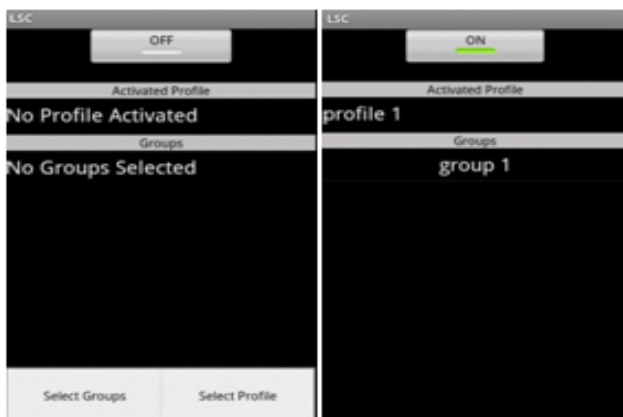


Fig. 13. Auto-reply

The user can create new groups, just he needs to give the title of the group and add contacts in the group as shown in figure 11(a) and (b) respectively. Contacts can be added to group by going to the desired group and selecting “add contact” from menu. Similarly user can create new profile(s), each profile contains profile title and auto-reply SMS. The user

needs to enter the title of the profile and auto-reply SMS as shown in figure 12(a) and (b) respectively. Auto-Reply SMS can be added by going to the desired profile and selecting “Select auto-reply SMS” from menu.



Fig. 14. Spam Options

Auto-reply mode can be activated by clicking the toggle (on/off) button in figure 13. Left window of figure 13 shows the view where auto-reply mode is off, no group and profile selected. Right window of figure 13 is showing the view where mode is on, one group is selected and profile is activated.

Spam Filter can be turned on by clicking toggle button on the spam window. Spam folder contains all the spam messages. Spam List contains all the numbers which are marked as spam. Spam Settings allows user to spam messages from weird numbers or unknown numbers by checking the checkboxes. A group can be added and marked as spam as well as shown in figure 14.

Similarly new events can be added by the user specifying event title, date / time, profile / SMS and group(s) / numbers associated with the event. The application sends the specified SMS to all the members associated with the event on the event date / time.

V. CONCLUSION

With the evolution in Smartphone era, leading IT companies and researchers have proposed many efficient applications; one of them is SMS Manager which helps to manage the SMS on smart phones. In this paper, we proposed Ontology based SMS Controller which is all in one SMS manager and includes new features like content based SMS Detection, Group chat etc. SMS Spam classification algorithm of Ontology based SMS Controller analyses the text of SMS and uses ontology to classify it as Spam or legitimate. The proposed algorithm has been tested on large number of test cases; the experimental results are satisfactory and supports the implementation of the solution.

REFERENCES

- [1] Yung Fu Chang, C.S. Chen, Hao Zhou, "Smart phone for mobile commerce", Computer Standards & Interfaces, Volume 31, Issue 4, June 2009, Pages 740-747.

- [2] Yung-Fu Chang, C.S. Chen, "Smart phone – the choice of client platform for mobile commerce", *Computer Standards & Interfaces*, Volume 27, Issue 4, April 2005, Pages 329-336.
- [3] Lorena Otero-Cerdeira, Francisco J. Rodríguez-Martínez, Alma Gómez-Rodríguez, "Ontology matching: A literature review", *Expert Systems with Applications*, Volume 42, Issue 2, 1 February 2015, Pages 949-971.
- [4] Carla Faria, Ivo Serra, Rosario Girardi, "A domain-independent process for automatic ontology population from text", *Science of Computer Programming*, Volume 95, Part 1, 1 December 2014, Pages 26-43.
- [5] Umar Manzoor, Samia Nefti, Yacine Rezgui "Categorization of malicious behaviors using ontology-based cognitive agents", *Data & Knowledge Engineering*, Volume 85, May 2013, Pages 40-56.
- [6] Francesco Rea, Samia Nefti-Meziani, Umar Manzoor, Steve Davis "Ontology enhancing process for a situated and curiosity-driven robot", *Robotics and Autonomous Systems*, Volume 62, Issue 12, December 2014, Pages 1837-1847.
- [7] Mohamed Yehia Dahab, Hesham A. Hassan, Ahmed Rafea "TextOntoEx: Automatic ontology construction from natural English text" *Expert Systems with Applications*, Volume 34, Issue 2, February 2008, Pages 1474-1480.
- [8] Ming-Hsiung Hsiao, Liang-Chun Chen "Smart phone demand: An empirical study on the relationships between phone handset, Internet access and mobile services" *Telematics and Informatics*, Volume 32, Issue 1, February 2015, Pages 158-168.
- [9] Apple Inc, (2012), <http://www.apple.com/> (Access Date: 12-06-2014)
- [10] Apple iPhone, (2012), <http://www.apple.com/iphone/> (Access Date: 12-06-2014)
- [11] Short Messaging Service (SMS), 2012, <http://en.wikipedia.org/wiki/SMS> (Access Date: 02-02-2014)
- [12] Droid Mate, "Anti Spam SMS and Private Box", <https://market.android.com/details?id=org.baole.app.antismsspam&hl=en> (Access Date: 15-01-2014)
- [13] Handcent Market - Handcent SMS, <https://market.android.com/details?id=com.handcent.nextsms&hl=en> (Access Date: 15-01-2014)
- [14] Kirill kruchinkin, "SMS AutoReply", <http://www.appbrain.com/app/sms-autoreply/auto.msg> (Access Date: 15-01-2014)
- [15] John Tsau, "Intelligent AutoReply", <https://market.android.com/details?id=com.jtsau.autoReply>, (Access Date: 16-01-2014)
- [16] GO Dev Team - GO SMS PRO, <https://market.android.com/details?id=com.jb.gosms&hl=en>, (Access Date: 15-01-2014)
- [17] Inna Novalija, Dunja Mladenčić, Luka Bradeško "OntoPlus: Text-driven ontology extension using ontology content, structure and co-occurrence information", *Knowledge-Based Systems*, Volume 24, Issue 8, December 2011, Pages 1261-1276.
- [18] WordNet (2014), <http://wordnet.princeton.edu/>
- [19] Samia Nefti, M. Oussalah, Yacine Rezgui "A modified fuzzy clustering for documents retrieval: application to document categorization", *Journal of the Operational Research Society*, Volume 60, Number 3, pp. 384-394, March 2009.
- [20] Manzoor, U.; Khan, M.; Qureshi, A.; ul Ghani, N., "Luxus SMS controller for android based smart phones," *International Conference on Information Society (i-Society)*, pp. 315-320, 25-28 June 2012.
- [21] Kwanho Kim, Beom-suk Chung, Yerim Choi, Seungjun Lee, Jae-Yoon Jung, Jonghun Park "Language independent semantic kernels for short-text classification", *Expert Systems with Applications*, Volume 41, Issue 2, 1 February 2014, Pages 735-743.
- [22] Duc-Thuan Vo, Cheol-Young Ock "Learning to classify short text from scientific documents using topic models with various types of knowledge", *Expert Systems with Applications*, Volume 42, Issue 3, 15 February 2015, Pages 1684-1698.
- [23] Umar Manzoor, Samia Nefti, Yacine Rezgui "Autonomous Malicious Activity Inspector – AMAI" *Natural Language Processing and Information Systems*, *Lecture Notes in Computer Science* Volume 6177, 2010, pp 204-215.
- [24] Umar Manzoor, Samia Nefti "iDetect: Content Based Monitoring of Complex Networks using Mobile Agents", *Applied Soft Computing*, Volume 12, Issue 5, May 2012, Pages 1607–1619.
- [25] Hao-jin TANG, Dan-feng YAN, Yuan TIAN "Semantic dictionary based method for short text classification", *The Journal of China Universities of Posts and Telecommunications*, Volume 20, Supplement 1, August 2013, Pages 15-19.
- [26] Umar Manzoor, Samia Nefti "Autonomous agents: Smart network installer and tester (SNIT)", *Expert Systems with Applications*, Volume 38, Issue 1, January 2011, Pages 884–893.
- [27] Umar Manzoor, Bassam Zafar "Multi-Agent Modeling Toolkit – MAMT", *Simulation Modelling Practice and Theory*, Volume 49, December 2014, Pages 215–227.
- [28] Lili Yang, Chunping Li, Qiang Ding, Li Li "Combining Lexical and Semantic Features for Short Text Classification", *Procedia Computer Science*, Volume 22, 2013, Pages 78-86.

Android Platform Malware Analysis

Rubayyi Alghamdi

Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

Khalid Alfalqi

Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

Mofareh Waqdan

Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

Abstract—Mobile devices have evolved from simple devices, which are used for a phone call and SMS messages to smartphone devices that can run third party applications. Nowadays, malicious software, which is also known as malware, imposes a larger threat to these mobile devices. Recently, many news items were posted about the increase of the Android malware. There were a lot of Android applications pulled from the Android Market because they contained malware. The vulnerabilities of those Applications or Android operating systems are being exploited by the attackers who got the capability of penetrating into the mobile systems without user authorization causing compromise the confidentiality, integrity and availability of the applications and the user. This paper, it gave an update to the work done in the project.

Moreover, this paper focuses on the Android Operating System and aim to detect existing Android malware. It has a dataset that contained 104 malware samples. This Paper chooses several malware from the dataset and attempting to analyze them to understand their installation methods and activation. In addition, it evaluates the most popular existing anti-virus software to see if these 104 malware could be detected.

Keywords—*Smartphone Security; Malware Analysis; Android Malware; Static Analysis; Dynamic Analysis; SDK; VAD*

I. INTRODUCTION

Several years ago, smartphone and tablet have become more common. They provide services such as social networking, banking, etc. Also, they are equipped with many features like Wi-Fi and GPS, make video calls and many more things. With all these features, there also comes a need for security for the mobile phones. This paper is focusing on the Android Operating System.

Android, which is open source operating system, will be more popular. Currently there are over 50 mobile phone companies are manufacturing smartphones with Android operating system. Increasing the number of the Android devices causes concern in term of user security. McAfee Labs report showed that in the first quarter in 2012, there is a large increase in mobile malware, and the increase was targeted almost only at the Android platform [1]. Figure 1 shows that there were 10 billion application downloaded by the end of the 2011. These rapid increases in applications download make Android to be the most targets for malware.

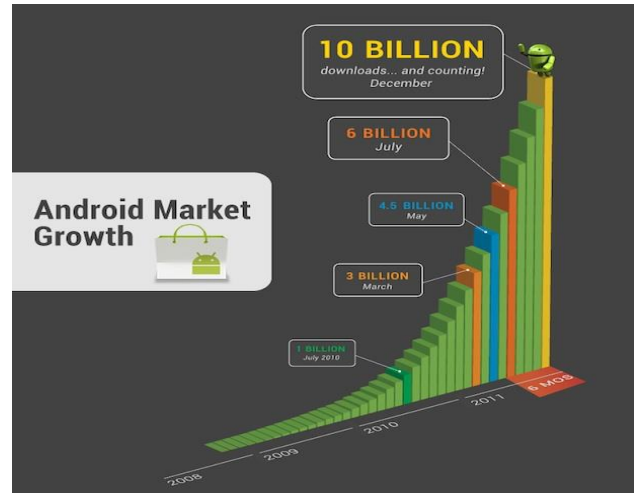


Fig. 1. Android Market Growth

In this paper, we are learning how a malware can target the Android phones and how it could be installed and activated in the device by performing a malware analysis using static and dynamic tools to understand the malware operations and functionalities. To achieve these tasks it is required to understand the Android architecture and its security model.

The rest of this report provide a description of the project and is organized as follows: Section II presents an overview of Android architecture Section III describe Android security model. After that, Section IV describes Android application Section V describes malware analysis followed by analysis result in Section VI. Section VII shows the detection results with four mobile anti-virus software. Section VIII discusses one way for future improvement. Lastly, it concludes the paper in Section IX.

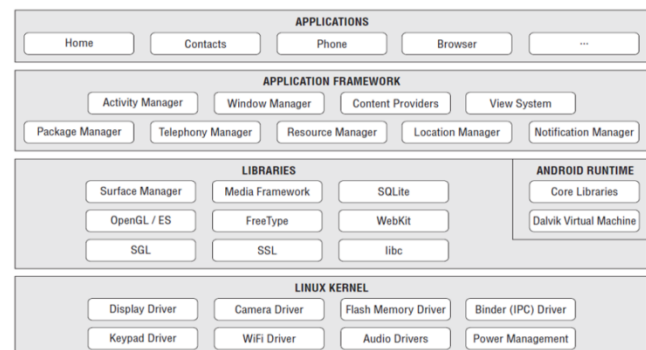


Fig. 2. Android Architecture [3]

II. OVERVIEW OF ANDROID ARCHITECTURE

Android is open source software for mobile development developed by Google. The Android architecture as shown Figure 2 can be divided into five layers. The first layer from the bottom is the kernel, which is based on the Linux 2.6 kernel. It is used as hardware abstraction layer. The reason Google are using Linux is because it provides memory management, process management, security model, networking, a lot of core operating system infrastructure that are robust and have been proven over time. The next level up is a native or basic library, which is written in C and C++. The next level is the Android runtime. The main component in the Android runtime is the Dalvik virtual machine. It was designed specifically for Android to meet the need of running in an embedded environment where you have limited battery, limited memory, limited CPU. The Dalvik virtual machine runs something called DEX files. Those files are bytes codes that are results of converting at java .classes and .jar files. Those files when are converted to .dex files become much more efficient bytes code that can run very well on small processors. They use memory very efficiency. The next level up from that are the core libraries. They are written in java programming language. It contains all of the collection classes, utilities, I/O, etc. The upper level is the application framework. This is also writing in java programming language. It provides abstractions of the underlying native libraries and Dalvik capabilities to application. Each Android applications run on its Dalvik virtual machine [4].

III. ANDROID SECURITY MODLE

The whole idea behind mobile platform is the fact that the user can run a lot and a lot of different applications on the device. The user might be installing and downloading a banking application that can be doing some sensitive data. On other hand, the user might be installing a game application right next to previous application and running on the same device. The user obviously does not want the game application to be able to access the sensitive data that banking application is operation on. So to achieve this Android platform makes sure that any application is isolated from each other. Basically when the user download and install an application, it will be given a unique UID. In addition, each application will run on separate process on separate virtual machine. Therefore, application cannot read other application private data [4].

As it was mentioned on Section II, Android was built on the top of the Linux, so the Linux file permission are applied. Permission allows the user to protect his/her sensitive data that are stored on the device. Also, it protects access to content provider, which basically is a database in the device. Permissions are requested by an application at install time and they are granted or denied once at the install time which requires the user approval [4].

IV. ANDROID APPLICATION

Android application has an extension file .apk which is stand for Android package. It is basically an archive file contains all the necessary files and folder in an application.

Each application is divided to four main components namely Activities, Service, Broadcast Receivers and Content provider [4].

- Activity is essentially just a piece of User Interface (UI). So any visual screens that allow user to see and interact with in an Android application. It can consist of views such as Button View, Text view Table view, etc.
- Intent Recivier which is a way for which an application to register some code that will not be running until it's triggered by some external event. Developer can write some code through XML and register it to be running when something happens, e.g. network connectivity is established at a certain time, or when the phone is ring.
- Service is a task that does not have any user interfaces. It is a component running in the background. For example, when lunched a music player application, the first screen is an activity. But as soon as selecting a song to play and move to other application, the service keeps running in the background.
- Contant Provider is data storage, which allows the applications to share the data with other application.

Each application contains a manifest file named Androidmanifest.xml. This file declares applications components, specifies the application requirements, and contains the permissions. These permissions will be shown to the user, when he would be installing the application. Figure 3 is an example of the Androidmanifestfile.xml.

```
root@bt:~/Sample Malware # ./aapt d xmltree RU.apk AndroidManifest.xml
N: android=http://schemas.android.com/apk/res/android
E: manifest (line=2)
  A: package="org.me.androidapplication1" (Raw: "org.me.androidapplication1")
  E: application (line=4)
    A: android:icon(0x01010002)=@0x7f020000
    E: activity (line=5)
      A: android:label(0x01010001)="Movie Player" (Raw: "Movie Player")
      A: android:name(0x01010003)="".MoviePlayer" (Raw: ".MoviePlayer")
      E: intent-filter (line=6)
        E: action (line=7)
          A: android:name(0x01010003)="android.intent.action.MAIN" (Raw:
"android.intent.action.MAIN")
        E: category (line=8)
          A: android:name(0x01010003)="android.intent.category.LAUNCHER" (Raw:
"android.intept.category.LAUNCHER")
      E: uses-permission (line=12)
        A: android:name(0x01010003)="android.permission.SEND_SMS" (Raw:
"android.permission.SEND_SMS")
```

Fig. 3. Example of Androidmanifest.xml

In the above example, it is clear that the application is trying to access the Send SMS feature of the phone, which is stated as the permission Android.permission.SEND_SMS.

The Android Manifest file also helps a user in determining whether an application is a legitimate one or it is a malicious one. For example, a game application does not need permissions such as SEND_SMSM, READ_CONTACTS. In this case, It should be known that if the application is a legitimate or not.

A. MALWARE INFECTION METHODS

There are several methods that the Android devices could be infected with malware. The following are four different methods which malware can be installed on the phone [1,11]:

1) Repackaging legitimate application

This is one of the most common methods used by the attackers. They may locate and download legitimate popular application from the market, disassemble it, add malicious code and then re-assemble and submit the new apps to the official or alternative Android market. Users could be vulnerable by being enticed to download and install these infected applications. It was found that 86.0% repackaged legitimate application including malicious payloads after analyzing more than 1,200 Android malware samples [1].

2) Exploiting Android's application bug

There could be a bug in the application itself. The attacker may use this vulnerability to compromise the phone and install the malware on the device.

3) Fake applications

It was also discovered that there are fake applications created to include malware which allows attacker to access your mobile device. Attackers upload on the market fake applications that seems are legitimate to users but they are malware by themselves. For example, Spyeeye's fake security tool was found in the market which is a malware.

4) Remote Install

The malware could be installed in the user phone remotely. If the attacker could compromise users' credentials and pass them in the market, then in this case, the malware will be installed into the device without the user knowledge. This application will contain malicious codes that allow attacker to access personal data such as contacts list [1].

V. MALWARE ANALYSIS

Malware is a piece of code that is executed on the target machine like viruses, Trojans or worms. Sometime it is difficult to stop them since they use new signature, which prevents it from being detected.

Reverse Engineering process is used to analyze the Android Malware. It is a process that decompiling an application to understand its working and functionality by analyzing the codes and debugging it. Before explain the analysis, it is very important to understand how an APK (Android package) is made before reversing it. Figure 4 shows the process of building and reversing APK file.

Once the application is downloaded in the phone from Google Market, the file .apk is available. So, first of all, the file should be de-packaging by using command such as "unzip. Then, the following files and folders will be found [13,14]:

- **Meta-inf Folder:** this folder consists of information that allows users to make sure of the security of the system and integrity of the APK application.

- **Res Folder:** this folder contains XMLs defining the layout, attributes, languages, etc.
- **Classes.dex:** this file contains the entire Java source code that is compiled. This file is run on the Dalvik Machine. This file consists of the complete bytecode that the Dalvik Machine will interpret.
- **AndroidManifest.xml File:** this file is one of the most important XML file which contains information about the permissions that the application needs or accesses. In other words this file contains the Meta information concerning the application.
- **Resources.arsc:** this file is binary resource file that is obtained after compilation.

Here at this point, the focus will be on the classes.dex file, which is the compiled java classes and contains all the codes of the application. Then, decompiling the classes.dex file into readable code (Java files) using several tools such as JAD tool.

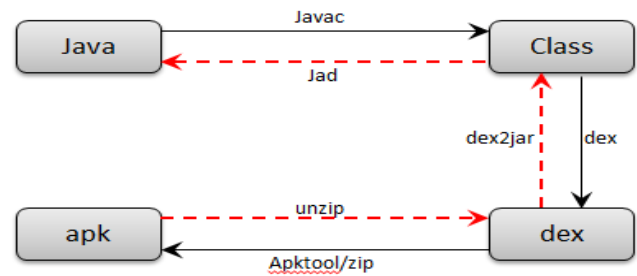


Fig. 4. Reverse APK file

The following sections describe the types of the malware analysis and the tools, which are used to perform a complete analysis.

A. Types of Malware Analysis

1) Dynamic Analysis

Dynamic analysis sometime is also called behavioral analysis, which is used to analyze and study the behavior of malware. Then studying how the malware interact with the system, services added, data capture, network connection made, open port and etc [11].

2) Static Analysis

Static analysis is called code analysis, which is used to analyze the code of the malicious software. The main purpose is to know the exact malicious code, which is embedded in the actual code [11].

B. Tools of Malware Analysis

Several tools were selected for both static and dynamic analysis. In this project, we use two methods to analysis the malware code. The first method involves the utilization of APKTOOL and editor such as Notepad++. The second method is performed using tools Dex2Jar and JD-GUI. The following is a list of tools used for reverse engineering Android malware [11,12]:

1) Create an isolation environment test

In this project, the VMware workstation 8 was used for the testing. Also, Linux back track 5 r2 on three laptops and Windows 7 in other laptop.

2) Apk Tool

This tool is used to analyze Android application binaries. It has a capability of disassembling applications to practically original form and repackaging them after certain modification. It also is used to debug the smali code [6].

3) Dex2Jar

Dex2Jar tool was developed and used in order to convert .dex file (Dalvik Virtual machine format) to .class format. It helps to view the source code of an application as a Java code [7].

4) Android SDK

The Android Software Development Kit (SDK) is a collection of development tools that are used to create applications. There are several components included on the SDK such as debugger, an emulator, sample source code, libraries and etc. [5].

5) JD-GUI

This tool is a java decompile that allows a user to view Java Source Codes of .class files. It shows log files and enables user to browse the hierarchy of the class files [8].

6) DroidBox

DroidBox is a dynamic analysis tool of Android applications. It is capable to identify information leaks of content, SMS data IMEI, GPS coordinates, and installed application, phone number and operation file [9].

C. Experimental

As mentioned earlier, there is a rapid growth of Android malware since 2011. In this section, we are explaining the basics of Android applications and showing the way to analyze them. Also, three different applications were chosen that were discovered back in 2010-2011. These sample applications can be downloaded from [10].

1) Android.FakePlayer

By start analyzing the first Android malware that was discovered in 2010 named Android.FakePlayer. It is a Trojan which sends SMS messages to certain numbers. It is distributed as an .apk file named "RU.apk". It pretends to be a movie player but does not actually play movies. This malware requires that the user install it on the device. To download the application sample it can be found here [10].

The file called RU.apk was found. It is a zip file that can be extracted using zip command. After extraction, several files were found. The most important file is AndroidManifest.xml. This is the metadata file that contains the information about the main class of the application as well as other information like permissions. Also, the file classes.dex was found which contains the actual compiled code on the dex file format.

The analysis starts by reading the AndroidManifest.xml file since it contains information about the main entry

points of the application as well as other useful information like permission and services used by the application. They can give a general overview of what the application is doing.

The Android Manifest.xml inside the APK file is a binary XML file. The apt tool was used to convert this format to a common XML format. After reviewing the file, it was obvious that the application is requesting the Android.permission.SEND_SMS that allows the application to send SMS messages.

It was noticed also that the activity that is launched when the application is executed is org.me.Androidapplication1.MoviePlayer. These are the entry points of the application. Since this is a movie player application, we know that the application does not need the send SMS messages.

Then, a disassembling Dex file was done by using Dex2Jar tool to get the readable original code to examine what the code is actually doing. We checked and reviewed the decompiled code of rg.me.Androidapplication1.MoviePlayer, which is the activity that will be launched first as seen in the AndroidManifest file. We also noticed that it contains a method named onCreate that on Android is called when the activity is starting. The code has several calls to Android.telephony.SmsManager.sendText Message (String destinationAddress, Strings cAddress, String text, PendingIntent sentIntent, PendingIntent deliveryIntent). So it was understood that the first time the application runs, it tries to send a numeric SMS text message to (3353, 3354) [11,12].

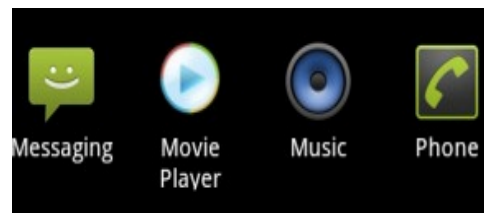


Fig. 5. Android.FakePlayer

To perform dynamic analysis of this malware, we install the application on Android 2.2. As shown in the figure 5, we have a new icon and application called Movie Player. When clicking on this application in our emulator, there is nothing happen. However, we know from our static analysis that after activating this application it will try to send out the SMS message.

2) Android.NickiSpy

The second malware that was analyzed is called Android Nickispy. It is a Trojan horse that steals information from Android devices and sends it to the remote server. It gets activated as soon as device finishes its boot. The package name of this malware is called "com.nicky.lyyws.xmlall".

After finishing analyzing the manifest file of this application, it was found that this malware requests a lot of permissions which some of them are related to the conversation recording capabilities (Figure 6) [8, 9].

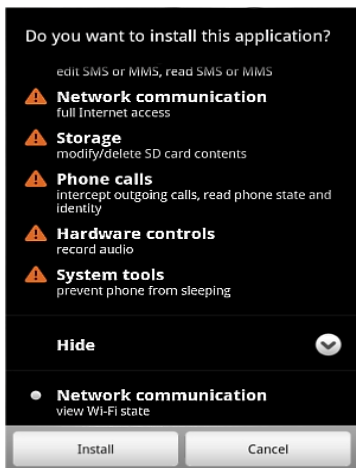


Fig. 6. NickiSpy Permissions

Also, It was noticed that there is a receiver declared in the Android Manifest file. The malware uses it to start the function after the device is booted. In addition, we found that there are many services declared to be run in the background. After the device boots and the malware activated, we found a list of service in the “running service”.



Fig. 7. NickiSpy Services

Those services are running in the background without the user noticing they exist. We craft all the entry point from the manifest file and then we began reviewing the Java code. We found in onCreate() method that the record service is been activated.

Also, it coded that after the malware is installed on the device, it creates a XML file names XM_ALL_setting in the shared preference file.

These file will contain all setting configuration of the remote server to let the device to send the information to the server. Also, we noticed that the address of the remote server is hard coded in the code. It also can be founded in the XM_ALL_setting file.

The following code was written which contains the send function which is used to send the information to remote server.

```
public void send(DataOutputStream paramDataOutputStream)
    throws IOException
{
    int i = this.Headinfo.getMessageLength();
    paramDataOutputStream.writeInt(i);
    int j = this.Headinfo.getCommandId();
    paramDataOutputStream.writeInt(j);
    int k = this.Headinfo.getVersion();
    paramDataOutputStream.writeByte(k);
    char[] arrayOfChar1 = this.Headinfo.getIENO();
    byte[] arrayOfByte1 = new String(arrayOfChar1).getBytes();
    paramDataOutputStream.write(arrayOfByte1);
    char[] arrayOfChar2 = this.UserNumber;
    byte[] arrayOfByte2 = new String(arrayOfChar2).getBytes();
    paramDataOutputStream.write(arrayOfByte2);
    char[] arrayOfChar3 = this.Date;
    byte[] arrayOfByte3 = new String(arrayOfChar3).getBytes();
    paramDataOutputStream.write(arrayOfByte3);
    int m = this.CallType;
    paramDataOutputStream.writeByte(m);
    int n = this.CallLen;
    paramDataOutputStream.writeInt(n);
    char[] arrayOfChar4 = this.Reserve;
    byte[] arrayOfByte4 = new String(arrayOfChar4).getBytes();
    paramDataOutputStream.write(arrayOfByte4);
    paramDataOutputStream.flush();
}
```

Fig. 8. Send Method in NickiSpy

Later, we performed a dynamic analysis of this malware. After installing the application on Android 2.2. We made a phone call between two emulators; the emulator which has Nickispy is installed on it records the conversation and save it in the SDCard under directories named “shangzhou/callrecord” [11,12].

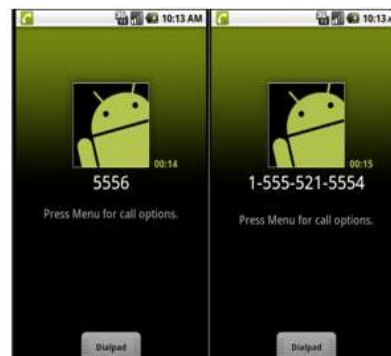


Fig. 9. Conversation Recording

The following code show that the malware sends SMS with the IMEI of the device to number ‘15859268161’

```
private class _cls1
    implements Runnable
{
    public void run()
    {
        Thread.sleep(60000);
        SmsManager smsmanager;
        String s1;
        smsmanager = SmsManager.getDefault();
        StringBuilder stringBuilder = new StringBuilder("IMEI:");
        String s = imei;
        s1 = stringBuilder.append(s).toString();
        smsmanager.sendTextMessage("15859268161", null, s1, null, null);
    }
}
```

Fig. 10. Code in Nickispy

The malware records other information other than the phone calls content. It also records GPS location and information and SMSs (received and sent).

3) Android. Seismic Application

The purpose of the analysis to reverse the Seismic application which is one of the most popular application in Google Market that allows a user to manage all the social networks . We aims to alter it by adding activities and more permission in the Android Manifest file and then recompile it and let it work on the phone. An attacker uses this process when to he downloads a legitimate application from the market and then embedded his/her malicious code.

We used Android 2.2 emulator and installed the Seismic application. The application has the following permissions:

- Access to the SD card
- Access to the GPS location
- Full Internet Access and
- Access to phone calls

It is possible to add more permission for this application in order to access to private data such as browser history, SMS and so on. We used the apktool to reverse the Seismic.apk file. The apktool generated a folder containing the AndroidManifest file and we modified the file by altering the version Name to be 1.6 and adding an activity which display a logo when the application is started and finally, adding more permissions to send and receive an SMS. Now that the Manifest file is altered. Then, we used the apktool to recompile the modified application. If the command executed successfully, a new folder named dist will be created which contain a file .apk. We renamed the file to Seismic1-6.apk and signed the application using a self-signed certificate which is generated by using openssl tool. Then, we installed the application on the phone. By checking the application information, it was noticed that the application version changed from (1.5 to 1.6) and that all permissions added in the AndroidManifest file are now available for the Seismic application. So the Seismic application has been successfully reversed, activity and permissions have been added on its source code, which was then recompiled to run on the phone [11].

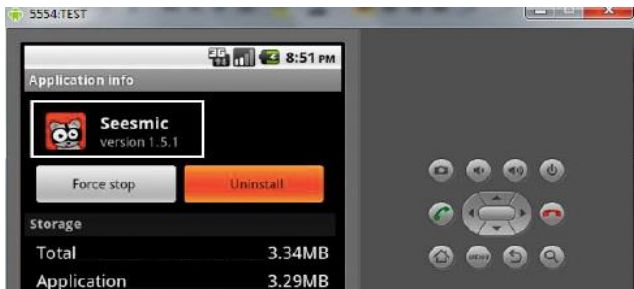


Fig. 11. Seismic Application

VI. MALWARE DETECTION

We attempt to measure the effectiveness of existing mobile anti- virus software. We choose four mobile anti-virus

software, i.e., AVG Antivirus Free, Lookout Security & Antivirus, BitDefender Mobile Security, and Avast Security Edition and download them from the official Android Market. We install each of them on a separate emulator running Android 2.2. We apply the default setting and enable the real-time protection. After that, we create a script that installs 104 applications in four emulators. If malware is detected, these anti-virus software will pop up an alert window, and then we recorded it down [1].

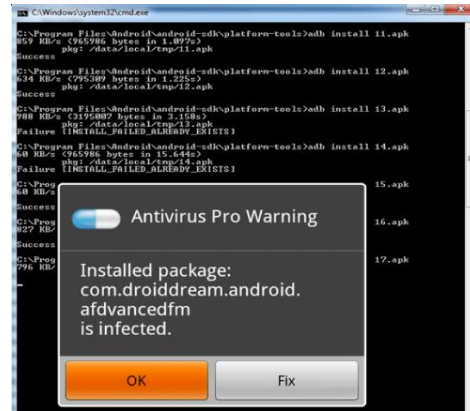


Fig. 12. AVG Detection

The table below shows the number of detected malware for each anti-virus, and its corresponding detection rate. The results are not encouraging: Avast detected 56 malwares; Lookout detected 51 malware; BitDefender detected 49 malware and finally AVG detected 46 malware.

TABLE I. THE NUMBER OF DETECTED MALWARE FOR EACH ANTI-VIRUS

	AVG	BitDefender	Avast	Lookout
Detected	46	49	56	51
%	44.2	47.1	53.8	49.0

There are some malwares were not detected by the four anti-virus software. One reason is that existing mobile anti-virus companies may not update the database signature for the free version anti- virus available in the market. Today, many people use free antivirus software and they have to know that they are not protected from malware.

VII. ANALYSIS RESULTS

After completed the test, we agreed that it becomes very important for every user to check the permissions that any application he/she is downloading really requires access to them or not. One of the major disadvantages of Android applications is that without agreeing to grant access to all the permissions, an application cannot be installed on the device. For example, the application Movie Player is only supposed to play movies and do nothing more. Hence it is obvious that it does not require permission to send messages or receive SMS. Similarly, the application Seismic accesses certain permissions in order for it to work normally. Since Seismic is one application that allows a user to manage all the social networks, hence it requires

access to Internet, Location in order to update status; but it does not require access to the user's SMS and private data. After we modify the manifest file and install the application, the user needs to have certain responsibilities while installing the application.

VIII. FUTURE WORK

The Android Manifest file is the only file that has been altered in order to allow the application to access more sensitive data. 83% of the malware they found in their analysis that they are repackaging applications. Therefore, if the developer can implement a method or implement an algorithm to detect the modified Android Manifest file before the application is installed to the phone, then the risk of those malware will be mitigated.

IX. CONCLUSION

In the past few years smartphone users have increased quickly. There are attackers who are now targeting smartphones. The main reason for this because the lack of user awareness regarding how their devices can be compromised. Today, smartphones like Android are not just used as a portable telephone. Android devices can access the internet, make online bank transmissions, manage social networks, etc. All these functionalities of a mobile phone seem very attractive for an attacker to gain information of the user and use it to his/her benefit. Therefore, users need to be aware enough and have full responsibilities to read and understand the permissions requested by the application before agreeing to grant access.

ACKNOWLEDGEMENT

This work is sponsored by Al-Baha University. Authors acknowledge the university for the kind support.

REFERENCES

- [1] Yajin Zhou, Xuxian Jiang, "Dissecting Android Malware: Characterization and Evolution," Proceedings of the 33rd IEEE Symposium on Security and Privacy (Oakland 2012), San Francisco, CA, May 2012
- [2] Jew Mark, "Android Market Reaching the Same Growth as App Store "gadgetoz.com, Dec 7, 2011. [Online]. Available: <http://www.gadgetoz.com/post/android-market-reaching-the-same-growth-as-app-store/>. [Accessed: July 25, 2012].
- [3] Technolgy, "Architecture of Android OS" [techneology.com](http://www.techneology.com/2011/11/architecture-of-android-os.html), Nov 2011. [Online]. Available: <http://www.techneology.com/2011/11/architecture-of-android-os.html>. [Accessed: July 25, 2012].
- [4] Frank Ableson, "Introduction to Android development the open source appliance platform "ibm.com, 12 May 2009. [Online]. Available:<http://www.ibm.com/developerworks/opensource/library/os-android-devel/>. [Accessed: July 25, 2012].
- [5] Android developer, Android-SDK. [Online]. Available: <http://developer.android.com/sdk/index.html>. [Accessed: July 3, 2012].
- [6] Reverse engineering tool for Android apk files, Android- Apktool. [Online]. Available: <http://code.google.com/p/android-apktool/>. [Accessed: July 3, 2012].
- [7] Tools to work with android .dex and java .class files, Dex2jar. [Online]. Available: <http://code.google.com/p/dex2jar/>. [Accessed: July. 3, 2012].
- [8] Decompiler, JD-GUI. [Online]. Available: <http://java.decompiler.free.fr/?q=jdgui>. [Accessed: July. 3,2012].
- [9] Android Application Sandbox, DroidBox.[Online].Available: <http://code.google.com/p/droidbox/>. [Accessed: July. 3, 2012].
- [10] Mobile malware mini dump, Malware dataset. Online]. Available: <http://contagiominidump.blogspot.ca/>. [Accessed: July. 3, 2012].
- [11] Vibha Manjunath, "Reverse Engineering of Malware on Android" [sans.org](http://www.sans.org), Aug 31, 2011. [Online]. Available: http://www.sans.org/reading_room/whitepapers/pda/reverse-engineering-malware-android_33769. [Accessed:July. 10,2012].
- [12] Jaime Blasco, "Introduction to Android Malware Analysis" [Magazine, Issue 34, June 2012]. Retrieved from :<http://net-security.org/insecuremag.php>. Last Accessed: 11 July, 2014

A Survey of Topic Modeling in Text Mining

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

Khalid Alfalqi
Information Systems Security
CIISE, Concordia University
Montreal, Quebec, Canada

Abstract—Topic Modeling provides a convenient way to analyze big unclassified text. A topic contains a cluster of words that frequently occurs together. A topic modeling can connect words with similar meanings and distinguish between uses of words with multiple meanings. This paper provides two categories that can be considered under the field of topic modeling. First one discusses the area of methods of Topic Modeling, which has four methods and can be considered under this category. These methods are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Correlated Topic Model (CTM). The second category is called Topic Evolution Model, it considers an important factor time. In this category, different models are discussed, such as Topic Over Time (TOT), Dynamic Topic Models (DTM), Multiscale Topic Tomography, Dynamic Topic Correlation Detection, Detecting Topic Evolution in scientific literatures, etc.

Keywords—Topic Modeling; Methods of Topic Modeling; Latent Semantic Analysis (LSA); Probabilistic Latent Semantic Analysis (PLSA); Latent Dirichlet Allocation (LDA); Correlated Topic Model (CTM); Topic Evolution Model

I. INTRODUCTION

To have a better way of managing the explosion of electronic document archives these days, it requires using new techniques or tools that deals with automatically organizing, searching, indexing, and browsing large collections. On the basis of today's research of machine learning and statistics, it has developed new techniques for finding patterns of words in document collections using hierarchical probabilistic models. These models are called "topic models". Discovering of patterns often reflect the underlying topics that are united to form the documents, such as hierarchical probabilistic models are easily generalized to other kinds of data; topic models have been used to analyze things rather than words such as images, biological data, and survey information and data [1].

The main importance of topic modeling is to discover patterns of word-use and how to connect documents that share similar patterns. So, the idea of topic models is that term which can be working with documents and these documents are mixtures of topics, where a topic is a probability distribution over words. In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated.

Create a new document by choosing a distribution over topics. After that, each word in that document could choose a topic at random depends on the distribution. Then, draw a word from that topic. [2]

On the side of text analysis and text mining, topic models rely on the bag-of-words assumption which is ignoring the information from the ordering of words. According to Seungil and Stephen, 2010, "Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, a corresponding low-rank representation of the high-dimensional histogram can be obtained for each document" [3]. The various kind of topic models, such as Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM) have successfully improved classification accuracy in the area of discovering topic modeling [3].

As time passes, topics in a document corpus evolve, modeling topics without considering time will confound topic discovery. Modeling topics by considering time is called topic evolution modeling. Topic evolution modeling can disclose important hidden information in the document corpus, allowing identifying topics with the appearance of time, and checking their evolution with time.

There are a lot of areas that can use topic evolution models. A typical example would be like this: a researcher wants to choose a research topic in a certain field, and would like to know how this topic has evolved over time, and try to identify those documents that explained the topic. In the second category, paper will review several important topic models.

These two categories have a good high-level view of topic modeling. In fact, there are helpful ways to better understanding the concepts of topic modeling. In addition, it will discuss inside each category. For example, the four methods that topic modeling rely on are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM). Each of these methods will have a general overview, the importance of these methods and an example that can describe the general idea of using this method. On the other hand, paper will mention the areas that topic modeling evolution provides such as Topic Over Time (TOT), Dynamic Topic Models (DTM), multiscale topic tomography, dynamic topic correlation detection, detecting topic evolution in scientific literature and the web of topics. Furthermore, it will going to present the overview of each category and provides examples, if any, and some limitations and characteristics of each part.

This paper is organized as follows. Section II provides the first category methods of topic modeling with its four methods and their general concepts as subtitles. Section III overviews of second category which is topic modeling evolution including its parts. Then it is followed by conclusions in Section IV.

II. THE METHODS OF TOPIC MODELING

In this section, some of the topic modeling methods will be discussed that deals with words, documents and topics. In addition, the general idea of each of these methods, and present some example for these methods, if any. Also, these methods involve in many applications so it will have a brief idea in what applications that can these methods work with.

A. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method or a technique in the area of Natural Language Processing (NLP). The main goal of Latent Semantic Analysis (LSA) is to create vector based representation for texts' to make semantic content. By vector representation (LSA) computes the similarity between texts' to pick the heist efficient related words. In the past LSA was named as Latent Semantic Indexing (LSI) but improved for information retrieval tasking. So, finding few documents that are close to the query has been selected from many documents. LSA should have many aspects to give approach such as key words matching, Wight key words matching and vector representation depends on occurrences of words in documents. Also, Latent Semantic Analysis (LSA) uses Singular Value Decomposition (SVD) to rearrange the data.

SVD is a method that uses a matrix to reconfigure and calculate all the diminutions of vector space. In addition, the diminutions in vector space will be computed and organized from most to the least Important. In LSA, the most significant assumption will be used to find the meaning of the text, otherwise least important will be ignored during the assumption. By searching about words that have a high rate of similarity will occur if those words have similar vector. To describe the most essential steps in LSA is firstly, collect a huge set of relevant text and then divide it by documents. Secondly, make co-occurrence matrix for terms and documents, also mention the cell name such as document x , terms y and m for dimensional value for terms and n dimensional vector for documents. Thirdly, each cell will be whetted and calculated. Finally, SVD will play a big roll to compute all the diminutions and make three matrices.

B. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (PLSA) is an approach that has been released after LSA method to fix some disadvantages that have found into LSA. Jan Puzicha and Thomas Hofmann introduced it in the year 1999. PLSA is a method that can automate document indexing which is based on a statistical latent class model for factor analysis of count data, and also this method tries to improve the Latent Semantic Analysis (LSA) in a probabilistic sense by using a generative model. The main goal of PLSA is to identifying and distinguishing between different contexts of word usage without recourse to a dictionary or thesaurus. It includes two

important implications: First one, it allows to disambiguate polysemy, i.e., words with multiple meanings. Second thing, it discloses typical similarities by grouping together words that shared a common context [3].

According to Kakkonen, Myller, Sutinen, and Timonen, 2008, "PLSA is based on a statistical model that is referred as an *aspect model*. An *aspect model* is a latent variable model for co-occurrence data, which associates unobserved class variables with each observation" [4]. The PLSA method comes to improve the method of LSA, and also to resolve other problems that LSA cannot do. PLSA has been successful in many real-world applications, including computer vision, and recommender systems. However, since the number of parameters grows linearly with the number of documents, PLSA suffers from over fitting problems. Even though, it will discuss some of these applications later [5].

On the other hand, PLSA is based on algorithm and different aspects. In this probabilistic model, it introduces a Latent variable $z_k \in \{z_1, z_2, \dots, z_K\}$, which corresponds to a potential semantic layer. Thus, the full model: $p(d_i)$ on behalf of the document in the data set the probability; $p(w_j | z_k)$ z_k representatives as defined semantics, the related term (word) of the opportunities are many; $p(z_k | d_i)$ represents a semantic document distribution. Using these definitions, will generate model, use it to generate new data by the following steps: [3]

- 1) Select a document d_i with probability $P(d_i)$,
- 2) Pick a latent class z_k with probability $P(z_k | d_i)$,
- 3) Generate a word w_j with probability $P(w_j | z_k)$.

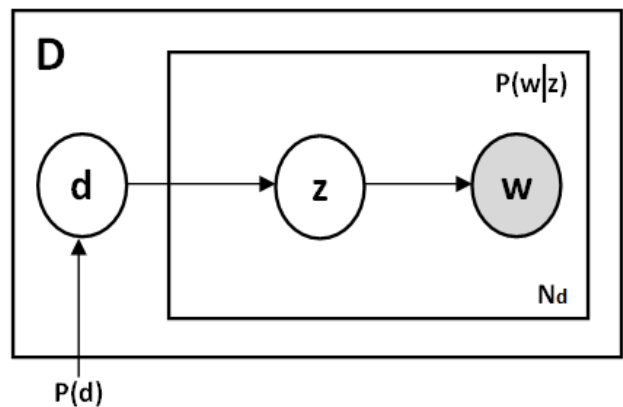


Fig. 1. High-Level View of PLSA

PLSA has two different formulations to present this method. The first formulation is symmetric formulation, which will help to get the word (w) and the document (d) from the latent class c in similar ways by using the conditional probabilities $P(d | c)$ and $P(w | c)$. The second formulation is the asymmetric formulation. In this formulation, each document d , a latent class, is chosen conditionally to the document according to $P(c | d)$, and the word can be generated from that class according to $P(w | c)$ [6]. Each of these two formulations has rules and algorithms that could be used for different purposes. These two formulations have been improved right now and this was happened when they released the Recursive Probabilistic Latent Semantic Analysis

(RPLAS). This method is extension for the PLAS; also it was improving for the asymmetric and symmetric formulations.

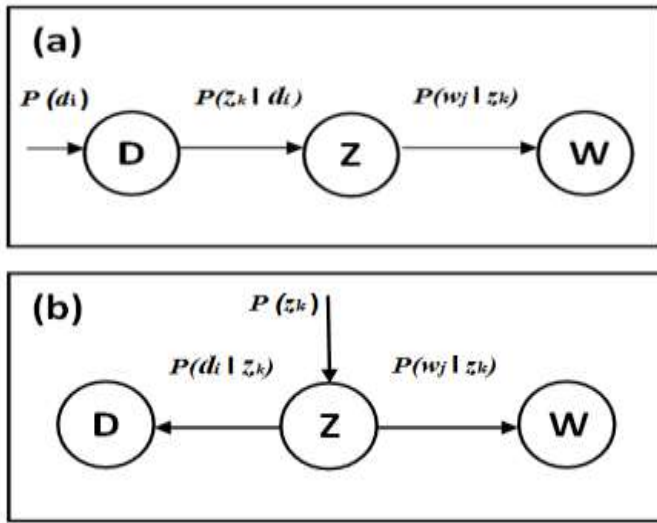


Fig. 2. A graphical model representation of the aspect model in the asymmetric (a) and symmetric (b) parameterization [3]

In the term of PLSA applications, PLSA has applications in various fields such as information retrieval and filtering, natural language processing and machine learning from text. In specific, some of these applications are automatic essay grading, classification, topic tracking, image retrieval and automatic question recommendation. Will discuss two of these applications as follows:

- Image retrieval: PLSA model has the visual features that it uses to represent each image as a collection of visual words from a discrete and finite visual vocabulary. Having an occurrence of visual words in an image is hereby counted into a co-occurrence vector. Each image has the co-occurrence vectors that can help to build the co-occurrence table that is used to train the PLSA model. After knowing the PLSA model, can apply the model to all the images in the database. Then, the pediment of the vector is to represent it for each image, where the vector elements denote the degree to which an image depicts a certain topic [7].
- Automatic question recommendation: One of the significant application that PLSA deal with is question recommendation tasks, in this kind of application the word is independent of the user if the user wants a specific meaning, so when the user get the answers and the latent semantics under the questions, then he can make recommendation based on similarities on these latent semantics. Wang, Wu and Cheng, 2008 reported that “Therefore, PLSA could be used to model the users’ profile (represented by the questions that the user asks or answers) and the questions as well through estimating the probabilities of the latent topics behind the words. Because the user’s profile is represented by all the questions that he/she asks or answers, we only need to consider how to model the question properly” [8]

C. Latent Dirichlet Allocation

The reason of appearance of Latent Dirichlet Allocation (LDA) model is to improve the way of mixture models that capture the exchangeability of both words and documents from the old way by PLSA and LSA. This was happened in 1990, so the classic representation theorem lays down that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture [9].

There are huge numbers of electronic document collections such as the web, scientifically interesting blogs, news articles and literature in the recent past has posed several new challenges to researchers in the data mining community. Especially there is a growing need of automatic techniques to visualize, analyze and summarize these document collections. In the recent past, latent topic modeling has become very popular as a completely unsupervised technique for topic discovery in large document collections. This model, such as LDA [10]

Latent Dirichlet Allocation (LDA) is an Algorithm for text mining that is based on statistical (Bayesian) topic models and it is very widely used. LDA is a generative model that tries to mimic what the writing process is. So it tries to generate a document on the given topic. It can also be applied to other types of data. There are tens of LDA based models including: temporal text mining, author- topic analysis, supervised topic models, latent Dirichlet co-clustering and LDA based bio-informatics [11], [18].

In a simple way, the basic idea of the process is, each document is modeled as a mixture of topics, and each topic is a discrete probability distribution that defines how likely each word is to appear in a given topic. These topic probabilities provide a concise representation of a document. Here, a "document" is a "bag of words" with no structure beyond the topic and word statistics.

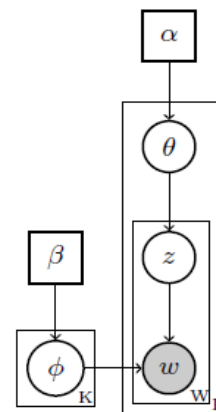


Fig. 3. A graphical model representation of LDA

LDA models each of D documents as a mixture over K latent topics, each of which describes a multinomial distribution over a W word vocabulary. Figure 3 shows the

graphical model representation of the LDA model. The generative process for the basic LDA is as follows:

For each of N_j words in document j

- 1) Choose a topic $z_{ij} \sim Mult(\theta_j)$
- 2) Choose a word $x_{ij} \sim Mult(\phi_{z_{ij}})$

Where the parameters of the multinomials for topics in a document θ_j and words in a topic ϕ_k have Dirichlet priors [12]

Indeed, there are several applications and models based on the Latent Dirichlet Allocation (LDA) method such as:

- Role discovery: Social Network Analysis (SNA) is the study of mathematical models for interactions among people, organizations and groups. Because of the emergence connections among the 9/11 hijackers and the huge data sets of human on the popular web service like facebook.com and MySpace.com, there has been growing interest in social network analysis. That leads to exist of Author-Recipient-Topic (ART) model for Social Network Analysis. The model combines the Latent Dirichlet Allocation (LDA) and the Author-Topic (AT) model. The Idea of (ART) is to learn topic distributions based on the direction-sensitive messages sent between the senders and receivers [13].
- Emotion topic: The Pairwise-Link-LDA model, which is focused on the problem of joint modeling of text and citations in the topic modeling area. It is built on the idea of LDA and Mixed Membership Stochastic Block Models (MMSB) and allows modeling arbitrary link structure [14].
- Automatic essay grading: The Automatic essay grading problem is closely related to automatic text categorization, which has been researched since 1960s. Comparison of Dimension Reduction Methods for Automated Essay Grading. LDA has been shown to be reliable methods to resolve information retrieval tasks from information filtering and classification to document retrieval and classification [15].
- Anti-Phishing: Phishing emails are ways to theater the sensitive information such as account information, credit card, and social security numbers. Email Filtering or web site filtering is not the effective way to prevent the Phishing emails. Because latent topic models are clusters of words that appear together in email, user can expect that in a phishing email the words "click" and "account" often appear together. Usual latent topic models do not take into account different classes of documents, e.g. phishing or non-phishing. For that reason the researchers developed a new statistical model, the latent Class-Topic Model (CLTOM), which is an extension of Latent Dirichlet Allocation (LDA) [16].
- Example of LDA: This section is to provide an illustrative example of the use of an LDA model on real data. By using the subset of the TREC AP corpus containing 16,000 documents. First, remove the stop-words in TREC AP corpus before running topic modeling. After that, use the EM algorithm to find the

Dirichlet and conditional multinomial parameters for a 100-topic LDA model. The top words from some of the resulting multinomial distributions are illustrated in Figure 4. As a result, these distributions seem to capture some of the underlying topics in the corpus (it is named according to these topics [9].

"ARTS"	"BUDGET"	"CHILDREN"	"EDUCATION"
New	Million	Children	School
Film	Program	Women	Students
Show	Tax	People	Schools
Music	Budget	Child	Education
Movie	Billion	Years	Teachers
Play	Federal	Families	High
Musical	Year	Work	Public
Best	Spending	Parent	Teacher
Actor	New	Says	Bennett
First	State	Family	Manigat
York	Plan	Welfare	Namphy
Opera	Money	Men	State
Theater	Programs	Percent	President
Actress	Government	Care	Elementary
Love	Congress	Life	Haiti

Fig. 4. Most likely words from 4 topics in LDA from the AP corpus: the topic titles in quotes are not part of the algorithm

D. Correlated topic model

Correlated Topic Model (CTM) is a kind of statistical model used in natural language processing and machine learning. Correlated Topic Model (CTM) used to discover the topics that shown in a group of documents. The key for CTM is the logistic normal distribution. Correlated Topic Models (CTM) is depending on LDA.

TABLE I. THE CHARACTERISTICS OF TOPIC MODELING METHODS [17]

Name of The Methods	Characteristics
Latent Semantic Analysis (LSA)	* LSA can get from the topic if there are any synonym words. * Not robust statistical background.
Probabilistic Latent Semantic Analysis (PLSA)	* It can generate each word from a single topic; even though various words in one document may be generated from different topics. * PLSA handles polysemy.
Latent Dirichlet Allocation (LDA)	* Need to manually remove stop-words. * It is found that the LDA cannot make the representation of relationships among topics.
Correlated Topic Model (CTM)	* Using of logistic normal distribution to create relations among topics. * Allows the occurrences of words in other topics and topic graphs.

TABLE II. THE LIMITATIONS OF TOPIC MODELING METHODS [17]

Name of The Methods	Limitations
Latent Semantic Analysis (LSA)	- It is hard to obtain and to determine the number of topics. - To interpret loading values with probability meaning, it is hard to operate it.
Probabilistic Latent Semantic Analysis (PLSA)	- At the level of documents, PLSA cannot do probabilistic model.
Latent Dirichlet Allocation (LDA)	- It becomes unable to model relations among topics that can be solved in CTM method.
Correlated Topic Model (CTM)	- Require lots of calculation - Having lots of general words inside the topics.

III. METHODS ABOUT TOPIC EVOLUTION MODELS

A. Overview of topic evolution models

When time goes by, the themes of a document corpus evolve. Modeling topics without considering time will cause problems. For example, in analyzing topics of U.S. Presidential State-of-the-Union addresses, LDA did not correctly do it by confounding Mexican-American War with some aspects of World War I, since LDA did not notice that there were 70-years of separation between the two events.

It is important to model topic evolution, so people can identify topics within the context (i.e. time) and see how topics evolve over time. There are a lot of applications where topic evolution models can be applied. For example, by checking topic evolution in scientific literature, it can see the topic lineage, and how research on one topic influences on another.

This section will review several important papers related to model topic evolutions. This paper will review model topic evolution by using different models, but all of them have considered the important factor 'time'. For example, probabilistic time series models are used to handle the issues in paper "dynamic topic models" and "non-homogeneous Poisson processes" and "multi-scale analysis" with "Haar wavelets" being employed in paper "multiscale topic tomography".

B. A Non-Markov Continuous-Time Method

Since most of the big data sets have dynamic co-occurrence patterns, word and topic co-occurrence patterns change over time, TOT model topics and their changes are done over time by taking into account both the word co-occurrence pattern and time [19]. In this method, a topic is considered as being associated with a continuous distribution over time.

In TOT, for each document, multinomial distribution over topics is sampled from Dirichlet, words are generated from multinomial of each topic, and Beta distribution of each topic generates the document's time stamp. If there exists a pattern of a strong word co-occurrence for a short time, TOT will create a narrow-time-distribution topic. If a pattern of a strong word co-occurrence exists for a while, it will generate a broad-time-distribution topic.

The main point of this paper is that it models topic evolution without discretizing time or making Markov assumptions that the state at time $t + 1$ is independent of the state at time t . By using this method on U.S. Presidential State-of-the-Union address for two centuries, TOT discovers topics of time-localization and also improves the word-clarity over LDA. Another experimental result on the 17-year NIPS conference demonstrates clear topical trends.

C. Dynamic Topic Models (DTM)

The authors in this paper developed a statistical model of topic evolution, and develop approximate posterior inference techniques to decide the evolving topics from a sequential document collection [20]. It assumes that corpus of documents is organized based on time slices, and the documents of each time slice are modeled with K-component model, and topics associated with time slice t evolve from topics corresponding to slice time $t-1$.

Dynamic topic models estimate topic distribution at different epochs. It uses Gaussian primarily for the topic parameters instead of Dirichlet, and can capture the topic evolution over time slices. By using this model, it has been inferred that what words are different from the previous epochs can be predicted.

D. Multiscale Topic Tomography

This method assumes that the document collection is sorted in the ascending order, and that the document collection is grouped into equal-sized chunks, each of which represents the documents of one epoch. Each document in an epoch is represented by a word-count vector, and each epoch is associated with its word generation Poisson parameters, each of which represents the expected word counts from a topic. Non-homogeneous Poisson process was used to model word counts, since it is a natural way to do the task, and also because it is amendable to sequence modeling through Bayesian multi-scale analysis. Multi-scale analysis was also employed to the Poisson parameters, which can model the temporal evolution of topics at different time-scales.

This method is similar to DTM, but provides more flexibility by allowing studying the topic evolution with various time-scales [21].

E. A Non-parametric Approach to Dynamic Topic Correlation Detection (DCTM)

This method models topic evolution by discretizing time [22]. In this method, each corpus contains a set of documents, each of which contains documents with the same timestamp. It assumes that all documents in a corpus share the same time-scale, and that each document corpus shares the same vocabulary of size d .

Basically, DCTM maps the high-dimensional space (words) to lower-dimensional space (topics), and models the dynamic topic evolution in a corpus. A hierarchy over the correlation latent space is constructed, which is called temporal prior. The temporal prior is used to capture the dynamics of topics and correlations.

DCTM works as follows: First of all, for each document corpus, the latent topics are discovered, and this is done by

first summarizing the contribution of documents at certain time, which is done by aggregating the features in all documents. Then, Gaussian process latent variable model (GPLVM) is used to capture the relationship between each pair of document and topic set. Next, hierarchical Gaussian process latent variable model (HGP-LVM) is employed to model the relationship between each pair of topic sets. They also use the posterior inference of topic and correlations to identify the dynamic changes of topic-related word probabilities, and to predict topic evolution and topic correlations.

An important feature of this paper is that it is non-parametric model since it can marginalize out the parameters, and it exhibits faster convergence than the generative processes.

F. Detecting Topic Evolution of Scientific Literature

This method employs the observation that citation indicates important relationship between topics, and it uses citation to model topic evolution of scientific literature [23]. Not only papers that are in a corpus $D(t)$ but cited papers are also considered for topic detection. It uses Bayesian model to identify topic evolution.

In this method, “a document consists of a vocabulary distribution, a citation and a timestamp”. Document corpus is divided into a set of subsets based on the timestamp, for time unit t , the corresponding documents are represented with $D(t)$. For each time unit, topics are generated independently. The topic evolution analysis in this paper is specified to analyze the relationship between topics in $D(t)$ and those in $D(t-1)$. In other words, it models topic evolution by discretizing time.

They first proposed two citation-unaware topic evolution learning methods for topic evolution: independent topic evolution learning method and accumulative topic evolution learning method. In independent topic evolution learning method, topics in $D(t)$ are independent from those in $D(t-1)$, while in accumulative topic evolution learning method, topics in $D(t)$ are dependent on those in $D(t-1)$. Then, Citation is integrated into the above two approaches, which is an iterative learning process based on Dirichlet prior smoothing. The iterative learning process takes into account the fact that different citations have different importance on topic evolution. Finally an inheritance topic model is proposed to capture how citations can be employed to analyze topic evolution.

G. Discovering the Topology of Topics

A topic is semantically coherent content that is shared by a document corpus. When time passes, some documents in a topic may initiate a content that differs obviously from the original content. If the initiated content is shared by a lot of later documents, the content is identified as a new topic. This paper is to discover this evolutionary process of topics. In this paper, a topic is defined as “a quantized unit of evolutionary change in content”.

This method develops an iterative topic evolution learning framework by integrating Latent Dirichlet Allocation into citation network. It also develops an inheritance topic model by using citation counts.

It works as follows: first, it tries to identify a new topic by identifying significant content changes in a document corpus. If the new content is different from the original content and is shared by later documents, it is being identified as a new topic.

The next step is to explore the relationship between the new topics and the original topics. It works by finding member documents of each topic, and examining the relationship. It also uses citation relationship to find member documents of each topic. That is, if a paper is being cited as start paper, this will be considered as the member paper of the start paper. In addition, papers that are textually close to the start paper are also considered as member paper of the start paper. The relationship between the original topics and the new discovered topics is identified by citation count. Their experimental results demonstrate that citations can better understand topic evolutions.

H. Summary of topic evolution models

This paper summarizes the main characteristics of topic evolution models discussed in section 3, which is listed as follows:

TABLE III. THE MAIN CHARACTERISTICS OF TOPIC EVOLUTION MODELS

Main characteristics of models	Models
Modeling topic evolution by continuous-time model	1)“Topics over time: a non-markov continuous-time model of topical trends”
Modeling topic evolution by discretizing time	1)“Dynamic topic models” 2)“Multiscale topic tomography” 3)“ANon-parametric Approach to Pair-wise Dynamic Topic Correlation Detection”
Modeling topic evolution by using citation relationship as well as discretizing time	1) “Detecting topic evolution in scientific literature: How can citations help” 2) “The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus”

I. Comparison of Two Categories

The main difference of the two categories is as follows: In the first category, model topics are considered without time and are basically model words. While in the second category, model topics are considering time viz. Continuous time, discretizing time, or by combining time discretization and citation relationship.

Due to the different characteristics of these two categories, the methods in the second category are more accurate in terms of topic discovery.

IV. CONCLUSION

This survey paper, presented two categories that can be under the term of topic modeling in text mining. In the first category, it has discussed general idea about the four topic modeling methods including Latent Semantic Analysis (LSA),

Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM). In addition, it explained the difference between these four methods in terms of characteristics, limitations and the theoretical backgrounds. Paper does not go into specific details of each of these methods. It only describes the high-level view of these topics that relates to topic modeling in text mining. Furthermore, it has also mentioned some of the applications being involved in these four methods. Also, it has been mentioned that each of these four methods has improved and modified over the previous one. Model topics without taking into account 'time' will confound the topic discovery. In the second category, paper has discussed the topic evolution models, considering time. Several papers have used different methods of model topic evolution. Some of them have used discretizing time, continuous-time model, or citation relationship as well as time discretization. All of these papers have considered the important factor 'time'.

REFERENCES

- [1] Blei, D.M., and Lafferty, J. D. "Dynamic Topic Models", *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.
- [2] Steyvers, M., and Griffiths, T. (2007). "Probabilistic topic models". In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum
- [3] Hofmann, T., "Unsupervised learning by probabilistic latent semantic analysis", *Machine Learning*, 42 (1), 2001, 177- 196.
- [4] Kakkonen, T., Myller, N., Sutinen, E., and Timonen, J., "Comparison of Dimension Reduction Methods for Automated Essay Grading", *Educational Technology & Society*, 11 (3), 2008, 275-288.
- [5] Liu, S., Xia, C., and Jiang, X., "Efficient Probabilistic Latent Semantic Analysis with Sparsity Control", *IEEE International Conference on Data Mining*, 2010, 905-910.
- [6] Bassiou, N., and Kotropoulos C. "RPLSA: A novel updating scheme for Probabilistic Latent Semantic Analysis", *Department of Informatics, Aristotle University of Thessaloniki, Box 451 Thessaloniki 541 24, Greece* Received 14 April 2010.
- [7] Romberg, S., Hörster, E., and Lienhart, R., "Multimodal pLSA on visual features and tags", *The Institute of Electrical and Electronics Engineers Inc.*, 2009, 414-417.
- [8] Wu, H., Wang, Y., and Cheng, X., "Incremental probabilistic latent semantic analysis for automatic question recommendation", *ACM New York, NY, USA*, 2008, 99-106.
- [9] Blei, D.M., Ng, A.Y., and Jordan, M.I., "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, 2003, 993-1022.
- [10] Ahmed, A., Xing, E.P., and William W. "Joint Latent Topic Models for Text and Citations", *ACM New York, NY, USA*, 2008.
- [11] Zhi-Yong Shen, Z.Y., Sun, J., and Yi-Dong Shen, Y.D., "Collective Latent Dirichlet Allocation", *Eighth IEEE International Conference on Data Mining*, pages 1019-1025, 2008.
- [12] Porteous, L., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M., "Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation", *ACM New York, NY, USA*, 2008.
- [13] McCallum, A., Wang, X., and Corrada-Emmanuel, A., "Topic and role discovery in social networks with experiments on enron and academic email", *Journal of Artificial Intelligence Research*, 30 (1), 2007, 249-272.
- [14] Bao, S., Xu, S., Zhang, L., Yan, R., Su, Z., Han, D., and Yu, Y., "Joint Emotion-Topic Modeling for Social Affective Text Mining", *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference*, 2009, 699-704.
- [15] Kakkonen, T., Myller, N., and Sutinen, E., "Applying latent Dirichlet allocation to automatic essay grading", *Lecture Notes in Computer Science*, 4139, 2006, 110-120.
- [16] Bergholz, A., Chang, J., Paaß, G., Reichartz, F., and Strobel, S., "Improved phishing detection using model-based features", 2008.
- [17] Lee, S., Baker, J., Song, J., and Wetherbe, J.C., "An Empirical Comparison of Four Text Mining Methods", *Proceedings of the 43rd Hawaii International Conference on System Sciences*, 2010.
- [18] X. Wang and A. McCallum. "Topics over time: a non-markov continuous-time model of topical trends". In *International conference on Knowledge discovery and data mining*, pages 424-433, 2006.
- [19] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *International conference on Machine learning*, pages 113-120, 2006.
- [20] R. M. Nallapati, S. Dittmore, J. D. Lafferty, and K. Ung. Multiscale topic tomography. In *Proceedings of KDD'07*, pages 520-529, 2007.
- [21] *A Non-parametric Approach to Pair-wise Dynamic Topic Correlation Detection*. In *Proceedings of the Eighth IEEE International Conference on Data Mining (ICDM 2008)*, Pisa, Italy, December 2008.
- [22] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles. Detecting topic evolution in scientific literature: How can citations help? In *CIKM*, 2009.
- [23] Yookyung Jo, John E. Hopcroft, and Carl Lagoze. The Web of Topics: Discovering the Topology of Topic Evolution in a Corpus, *The 20th International World Wide Web Conference*, 2011.

Effectiveness of iPhone's Touch ID: KSA Case Study

Ahmad A. Al-Daraiseh
IS dept. King Saud University
Riyadh, Saudi Arabia

Diana Al Omari
IS dept. King Saud University
Riyadh, Saudi Arabia

Hadeel Al Hamid
IS dept. King Saud University
Riyadh, Saudi Arabia

Nada Hamad
IS dept. King Saud University
Riyadh, Saudi Arabia

Rawan Althemali
IS dept. King Saud University
Riyadh, Saudi Arabia

Abstract—A new trend of incorporating Touch ID sensors in mobile devices is appearing. Last year, Apple released a new model of its famous iPhone (5s). One of the most anticipated and hailed features of the new device was its Touch ID. Apple advertised that the new technology will increase the security of its device, and it will also be used in different applications as a proof of identity. To make the issue more controversial, Apple announced a new financial service (Apple Pay) that allows iPhone 6 users to use their iPhone as a replacement to credit cards. The minute the new technology was introduced; many questions appeared that needed immediate answers. Users were concerned about how it will work? Is it easy to use? Is it really safe? And whether it will be effective in protecting their private data or not? In this paper we provide a comprehensive study of this feature. We discuss the advantages and disadvantages of using it. Then we analyze and share the results of a survey that we conducted to measure the effectiveness of such feature in the Kingdom of Saudi Arabia (KSA). In this study, we only focus on users from KSA, because if the device fails to protect mobile's data, severe consequences might happen. Due to cultural believes in KSA, releasing mobile contents to unauthorized people could lead to crimes. Survey analysis revealed somewhat controversial results, while 76% of all participants believe that this technology will improve the device security, only 33% use it to lock/unlock their devices, and even a smaller percentage use it to make purchases.

Keywords—iPhone 5s; iPhone 6; iPhone 6 plus; Fingerprint; Touch ID

I. INTRODUCTION

Nowadays, one of the main concerns in the mobile computing industry is the mobile security. Smartphones and other mobile devices can store and process a large amount of data in different formats. The majority of such data is private and confidential. Moreover, Hardware and software advances in this field made mobile devices an essential part of almost every activity we carry on in our lives. Storing large amount of data about such activities made mobile device a target for all types of attacks.

Attackers used vulnerabilities in communication protocols (such as, GSM, WIFI, and Bluetooth), Hardware, and software

to attack mobile devices. Therefore, securing such devices from all types of attacks became a priority to all manufacturers and software developers. One of the modern security methods used in securing Smartphones against unauthorized users is the fingerprint technology. It was originally introduced to the mobile industry by Apple Company in its iPhone 5s device, and was re-used again in the new mobile editions iPhone 6, and iPhone 6 plus.

Fingerprint is the most widely used biometric to identify different individuals. It is impossible to find two persons with an identical fingerprint pattern. Also, fingerprint patterns never change during an individual's life span, which make them ideal means for identification purposes. [1]

The concept was introduced for the first time by the Chinese who invented a new technique called fingerprint to identify people. The idea received more attention in Europe during the 17th and 18th centuries, were European scientists began their interest in the human skin especially friction ridge skin. Later, in the 19th century England published many books about fingerprint. In the 20th century exactly in 1902 fingerprint evidence has started to be used in the courts of England. In 1903, New York developed the first system that uses fingerprinting for criminal purposes. Then in the year of 1921, Federal Bureau of Infestation (FBI) used fingerprinting as an identification method and built special section for that. In 1992, identification section was rebuilt as the Criminal Justice Information Services division (CJIS). [2]

Since the 80's of the last century, the usage of computing devices increased rapidly. Such devices stored and processed very sensitive data. Immediately, scientists realized the need for a strong authentication mechanism to protect those devices from an unauthorized user. While passwords and smart cards are good means for authentication, a human fingerprint might be the most unique and hardest to fake or break [3].

The important questions now are, to what extent can this technology help securing mobile devices? Do users have any concerns when using it? Will it be used openly or selectively? These questions and others will be discussed later in the

analysis section.

The rest of the paper is organized as follows: in section 2, related work is discussed. In section 3, the methodology is presented. In section 4, a comprehensive analysis is provided. Finally in section 5, conclusions are drawn.

II. RELATED WORK

Steve Gold [4] wrote on how the future of payment authentication will be through biometric means. He explained that multiple agencies will be involved and that any standardization effort needs to consider all of them. Steve stated that using such technology will simplify the authentication process. He concluded that in order to protect users' privacy there shouldn't be a central database for biometrics, and network tracking of such devices shouldn't be allowed.

Stephen Tipton et al. [5] investigated the iOS security issues. The authors pointed out that the scanned biometric data could be recorded by Apple, in addition to problems related to faking fingerprints and usability issues. They concluded few measures Apple took to protect such data; for example, keeping the data away from app developers, turning tracking off ability, providing the iCloud Key Chain which uses different PIN, and the utilization of strong encryption to prevent any group from accessing such data.

Shri et al. [6] did a study on the usability of Smartphone fingerprinting. The Authors did a task oriented experiment to see whether PIN authentication or fingerprint Authentication was more usable. Their results indicate that Fingerprint authentication was more appealing and that it could reduce the number of Smartphones that was left unsecured without a PIN.

N. Yildirim and A. Varol [7] investigated the different biometric features that could be utilized to protect mobile devices; for example, face, voice, and fingerprint. They also listed different methods and applications of such features. They concluded that fingerprint authentication will be used heavily and in different applications.

Ming Gao et al. [8] focused on the benefits the fingerprinting technology in Smartphones will bring, and challenges it will face. They concluded that this technology will be the mainstream in the future.

S. SaintGermain [9] discussed a new law in California that required a warrant to search any Smartphone. This law is considered a victory for privacy activists. The author concludes that by law, the victim shouldn't be forced to unlock his own Smartphone, and hence, the police need to be able pass the biometric authentication, even with a search warrant, by other means.

Hugh and Lorie [10] claim that using fingerprint as an authentication mechanism may reduce the system's security. The authors did a little experiment. They prepared two groups of people and asked them to create passwords to protect an e-banking account. One of the groups was only allowed to choose passwords, the other one was allowed to use fingerprints as well as passwords. By examining the length and the strength of the passwords they had chosen, the results

showed that the group that was given the fingerprint option created less secure passwords than the other group. That led the authors to say that the group who had (password-with-fingerprint) account felt more secure, which made them create less secure passwords. In conclusion, using the fingerprint authentication shouldn't seduce us to select weak passwords.

Tarika and Bhawna [11] indicate that fingerprint authentication shouldn't be used. Their reasoning is that, we leave our fingerprint everywhere, and that it is very easy to reproduce such fingerprints. Hence, using them is not safe.

J. Hu [12] discussed different methods for the protection of fingerprint templates. Specifically, he considered biometric key generation, fuzzy schemes and noninvertible transforms. He concluded that the first two methods don't require the storage of a template, and the third one easily produces cancellable fingerprint templates.

It is very clear from all of the above that there are mixed opinions regarding this technology. Given the peculiar nature of Saudi Community, this research aims at finding out in which direction KSA's users will go? And how deep they will utilize the technology?

III. METHODOLOGY

In order to produce a comprehensive study of iPhone's fingerprint technology, a large amount of information was gathered and analyzed from different resources; such as, papers, newspapers, and electronic articles. After that, a survey was published to see whether Saudi people can trust this technology for securing their sensitive data or not. The reason why only Saudi participants were selected is that we wanted to see how the most private and protected society accepted the technology. The results will be discussed in the analysis section.

The main challenge was the lack of resources especially that the fingerprinting in Smartphones is new. Only few articles discussed the technology. Also, most of the conclusions were opinions rather than facts.

IV. ANALYSIS

Apple Company, one of the largest well-known companies in the computing and Smartphone industry, has released the new version of smart phones "iPhone 5s" with a new feature added to it. The purpose of this new feature as Apple states is to improve the security of mobile phones, make it easier to their customers to protect their phones, and use it as a way to verify and accept orders done by users from the iTunes Store, and in iPhone 6, use the phone to replace credit cards.

Using this technology, iPhone mobile users can secure and lock their phones by a touch of their finger, as simple as that. So, before actually getting into the privacy details of this feature, let's give a general view over it by talking about this feature and how it works.

Currently, the technology exists in the latest releases of the iPhone (5s and 6), some iPad versions, and other Smartphones from different manufacturers. In order to activate this feature on your device, all you have to do is to put your fingerprint on the button and through this touch; your fingerprint will be

saved through an embedded sensor. It is important to mention here that this button is made of hard glass material in order to protect it. It is also used as a lens to generate a clear picture of your fingerprint. The more you use it on your mobile, the better the scanner will recognize your fingerprint [13].

The using of fingerprint was expanded in iPhone 6 to include purchasing products by using fingerprint as a way to pay. Apple realized how hard it is to carry and manage multiple credit cards. They also realized the danger that threatens our safety when carrying them. "Apple Pay" is a new service introduced by Apple. It is a way to pay by phone using fingerprints and NFC technology. Apple has promised a high level of security so that all transactions are confidential, and no one can track what we buy using this service. The service is now working in the United States and had a strong commencement. Apple made agreements with a large number of shops and officially began the service in October 2014 using the iPhone 6 and 6 Plus devices only. More than 220000 shops and popular restaurants in America will support this service [14]

The following sections discuss this feature from different security perspectives.

A. Safety and usability

Firstly, regarding the security and the safety of the saved fingerprints, Apple's senior vice president of hardware engineering, Dan Rico illustrated how the company's technique used to save the fingerprint information is very secure, Apple utilized one of its security techniques called "Secure Enclave".

Generally speaking, secure enclave is like a vault where information can be stored and this information cannot be accessed without the touch ID of the user. Also, the fingerprint will be saved after it has been encrypted. As Mr. Dan emphasized, the fingerprint will never ever be used in other software nor it will be saved on the company's servers.

This was regarding where the fingerprints will be saved, but actually in our daily activities, our fingerprints can be anywhere. Wherever we put our hands, our fingerprints will be. So what if someone tries to simulate our fingerprint? Will he be able to open our mobile? The answer is definitely "NO" because according to Apple Company, the sensor senses the shapes on our fingerprint from specific layers of the skin that only works on a live finger.

Secondly, regarding the usability of this iPhone 5s' fingerprint feature, is it easy to use? Absolutely yes. A user only needs to register his/her fingerprint for the first time, then start using it each time he/she wants to unlock the phone. When a user wants to unlock the phone, he/she has two options: either enter the PIN or push the home button by one of registered fingers. Both methods produce the same result. So what's the difference and why would someone use the fingerprint feature? Actually, the answer of this question will be in the next section, where a list of advantages and disadvantages of this feature will be shown. [15]

B. Advantages and disadvantages

Just like any other new technology, iPhone's touch ID has

some advantages and some disadvantages. Advantages will be listed first:

- The first and most important advantage of this feature is its uniqueness. And hence it gives us a peace of mind that no one else will be able to unlock our devices. Based on this we can also assume that our data is more protected.
- Fingerprint recognition is fast. The device unlocks almost instantaneously.
- Ease of use. The phone will unlock by putting the owner's finger over the Home button.
- Convenient. Unlocking the phone doesn't require much attention, and hence users can be doing other tasks as they unlock the phone.
- Universal. iPhone's fingerprint recognition system allows the user to enrol multiple fingers which let the user use any other finger to unlock the phone if one of his fingers is injured.
- Long lasting. A person's fingerprint does not disappear by aging, but as people get older they usually lose their collagen which makes it harder to recognize their fingerprint. [16]
- Another advantage is that when the owner wants to buy music or any other material from the iTunes store, he doesn't need to enter the password, he can only use his fingerprint and this will be as a verification of his identity. [17]

On the other hand, the following are weaknesses or disadvantages of this technology:

- Fingerprints can be easily recreated. Tarika [11] indicated that fake fingerprints can be used to unlock the device.
- Overconfidence. Using the fingerprint option makes us feel more secure and hence we tend to choose weak passwords as a backup. As suggested by [10].
- Fears of wrong storage or usage. Many researchers and users expressed their fears and lack of trust. Losing such information can lead to severe consequences.
- Sensor's sensitivity. Dirty or oily skin might affect the accuracy of the sensor. Also, fingerprint recognition is affected by what the finger is exposed to of injuries or burns.

C. Reliability

Is it reliable or not? Can people rely on it as they did with the PIN? The Touch ID is very reliable and durable. Although, some people have found that sometimes the sensor may not respond to their fingerprint if the hand is wet or has a high temperature. It does work for the majority of people with no issues.

D. People's perspective toward fingerprint feature

Generally speaking, some people like this feature and find

it as an interesting new feature to protect their mobiles, and even if there is a password, they would like to use it as a way of following the technology without thinking about any privacy concerns. But actually, these are the minority, whereas the majority of people have high concerns regarding the real aim of such feature. Why to have our fingerprints saved at a specific place even if no one can share or use it. As long as the password is still there, why does Apple Company and others release such feature? Moreover, with all of Apple's efforts to convince people that their fingerprints information will be secured and not saved on their servers, people still have high fears of Apple's other objectives of this feature and whether Apple will share any of their analytical information about the Touch ID system to Apple or any other party [18].

E. People's fears and concerns

Most of people's concerns are centred on privacy and identity tracking. One main concern is that Apple stores the users' fingerprints in its servers, creating a huge database of users' biometric information for people from all around the world. If it happens, it will pose a huge threat to all users, especially if this data is handed to governments of different countries. These fears have increased dramatically after the United States' National Security Agency (NSA) spying scandal was uncovered. NSA collected personal information of citizens and residents in USA through a program called PRISM. Regarding this matter, Apple confirmed that it will not store the fingerprints in its servers and they will not be synchronized with iCloud even. Instead, it will be stored only on the encrypted chip A7. Also, it will not be stored as an image, but instead it will be stored as fingerprint data. It is worth mentioning that Apple calls the technique as (Touch ID) not (finger scan) which is an accurate description of what it does, so it doesn't scan the fingerprint but it reads features that distinguish one person from another. So, it divided the fingerprint into three parts (whorl, loop, arch) and then picked up the finer details such as the path of the blood veins. [19]

Another concern is about the recreation of one's fingerprint. These fears have increased even more since the media published a story about a German hacker who was able to hack iPhone5s Touch ID and unlock the device using fake finger from a fingerprint's photo. [20]

People are also concerned that a thief is forced to cut off the victim's finger to be able to unlock the phone. Such concerns may seem exaggerated, but we can't ignore that it already happened. In 2005, a car thief in Malaysia cut off part of the owner finger to steal a car, Mercedes S-Class, which was protected by fingerprint recognition system. Regarding this, Apple confirmed that it has developed the technology, so that fingerprint recognition happens by scanning the finger skin dermal layer, which requires the finger to be alive and in its natural state. After all, the real concern would be "do thieves know that?" [21]

In the next section, results of the conducted survey will be explained in detail.

F. Survey results:

According to the survey, the majority are using iPhone for more than 3 years. And most of them are using iPhone 5s.

A survey was used to see the prevalence of this new feature amongst Saudis, whether they have liked it or not? And what are their fears and concerns about it? The survey was filled by 2230 persons living in different Saudi Arabia regions. Our sample consists of 780 females and 1450 males. The majority of all participants held a bachelor degree and between 25 to 34 years old. Thus, they are overwhelmingly young. Most of the participants in the sample are from Riyadh region. The demographic questions answers are in table 1, 2 and 3.

TABLE I. PARTICIPANT REGION

Region	Number	Percent
Riyadh region	1310	59%
Mekkah region	300	13%
Eastern region	180	8%
Qassim region	140	6%
Asir region	90	4%
Medina region	70	3%
Al Jawf region	40	2%
Northern Border region	30	1%
Jazan region	20	1%
Tabuk region	20	1%
Al Bahah region	10	1%
Hail region	10	1%
Najran region	10	1%

TABLE II. PARTICIPANT AGE

Age	Number	Percent
18-24	600	27%
25-34	1060	48%
35-44	450	20%
45-54	90	4%
55+	30	1%

TABLE III. PARTICIPANT EDUCATION

Education	Number	Percent
No high school degree	40	2%
High school degree only	460	21%
Bachelor degree	1560	70%
Master degree or higher	170	8%

Also, by asking the participants if they use the same mobile device password PIN for their online account, 21% answered “Yes”, 46% answered “No” while 29% answered “Sometimes”.

Regarding the usage of the fingerprint feature, the results were somewhat controversial, although 55% of all users think that password PIN is not secure enough, and 76% agree that the use of biometric can improve the mobile security, only 33% use fingerprint to unlock their iPhone device, while 17% use it sometimes. Besides that, only 16% use the fingerprint to buy from iTunes usually, and 5% use it sometimes, while 77% do not use it at all. The questions along their answers in details are found in table 4.

When asked the participants if they have concerns about using the fingerprint feature, 31% answered “Yes” while 67% answered “No”. Then, people who answered “Yes” were asked about their concerns. The majority of all concerns were from breach of privacy. By comparing the answers based on

the range of ages, we found that the majority of people in the ages between 18 to 44 were concerned from a breach of privacy, while the major concerns for the people who are in the ages between 45 to 54, are releasing their fingerprint’s details to governmental agencies. The comparison details can be seen in figure 1.

When asked “What makes you comfortable with protecting your iPhone?” 49% answered “Having strong password”, 34% answered “Using biometrics”, 23% answered “Having antivirus” and 17% answered “Having security software” as showed in figure2.

Finally, participants were asked about the most important things in their phones, 88% of all females which is the majority answered “personal photos” and 63% of males answered “personal photos” and “personal information” as seen in figure 3. The questions and their answers in details are shown in table 5.

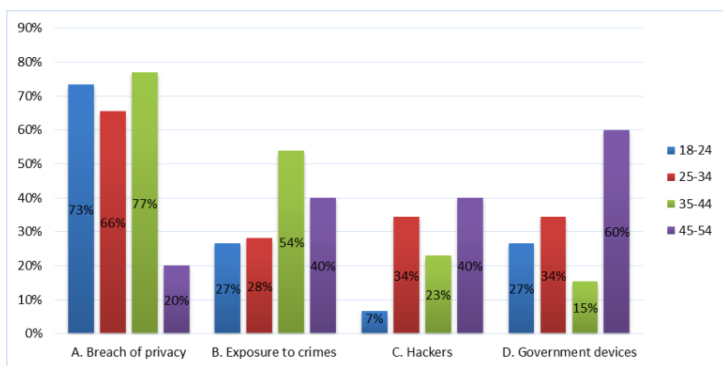


Fig. 1. participant concerns

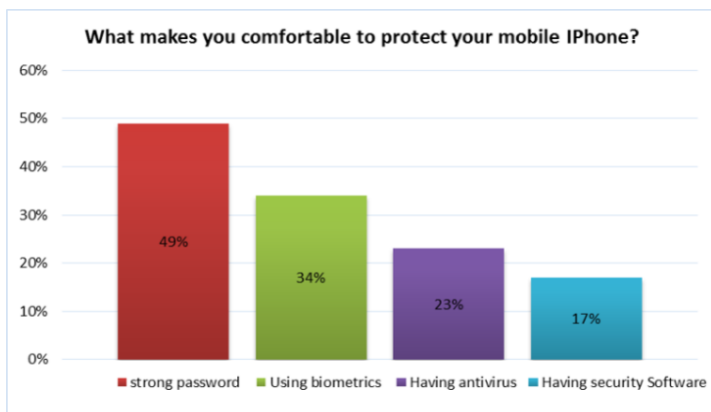


Fig. 2. protection methods results

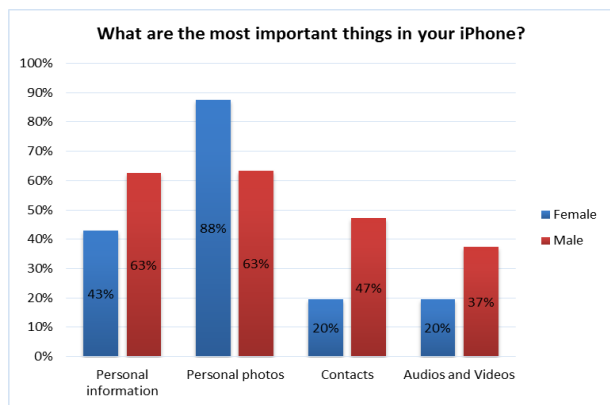


Fig. 3. participant important mobile content

TABLE IV. SUMMARY OF THE IMPRESSIONS OF FINGERPRINTS

No	Question	Responses				Total
		Yes	No	Sometimes	No answer	
1	Do you own an iPhone?	1950	180		91	2230
2	How long have you been using it? a. One year b. Two years c. More than 3 years	380 270 1400			180	2230
3	Which iPhone model do you have? a. iPhone 5s b. iPhone 6 c. iPhone 6 plus d. other	1630 80 40 480				2230
4	Do you use the same mobile password PIN for your online account?	490	1030	670	40	2230
5	Is the password "PIN" secure enough?	960	1230		40	2230
6	Do you think using biometrics improves the mobile security?	1690	490		50	2230
7	Do you use the fingerprint feature to unlock your device?	720	1080	380	41	2230
8	Do you use the fingerprint feature to buy from iTunes?	350	1710	110	51	2230
9	Do you face any difficulties while using it?	310	1420	380	120	2230
10	Do you have any concerns related to the fingerprint feature?	690	1490		50	2230

TABLE V. SUMMARY OF SECURITY ISSUES

No	Question	Responses	Total
1	Which of the following makes you comfortable protecting your iPhone? a. Having antivirus b. Having security Software c. Having strong password is good enough d. Using biometrics e. All f. A & B g. C & D h. A & B & C i. No answer	280 200 790 480 60 50 150 20 80	2290
2	What are the most important things in your iPhone? a. Personal information b. Contacts c. Personal photos d. Audios and Videos e. All f. A & C g. A & D h. A & B & C i. No answer	230 110 340 70 40 280 40 10 310	1430

V. CONCLUSIONS

In conclusion, there is no doubt that using Touch ID in Smartphones is an attractive and somewhat secure feature. Apple and other mobile manufacturers are racing to include in their products and find more ways to utilize it.

Different communities have mixed views regarding this technology. Some of them think that it is the most secure and convenient feature, while others think it not secure and can be used for tracking purposes.

In this paper, we covered this feature from all aspects. We discussed the pros and cons of this technology and the different views of users and researchers.

In KSA, the survey results show that people extremely care about their mobile data. Although ostensibly, the majority believes that the Touch ID will improve the security of their phones, only a small percentage fully trusts it.

ACKNOWLEDGMENT

At the end of this paper, we would like of course to thank Allah who supports us in doing everything in our lives and of course in completing this paper specially. Also, we would like to thank Dr. Ahmad Darayseh - Our instructor in the Security course - for his support and help in suggesting this interesting topic for us and also giving us many advices and ideas. Last but not least, we would like to thank our families and friends and everyone supported us with any idea and helps us in completing this paper.

REFERENCES

- [1] T. Trimpe. "Fingerprint Basics" [Online]. Available: <http://sciencespot.net/Media/FrnsScience/fingerprintbasicscard.pdf>
- [2] The International Association of Identification in partnership with NIJ, The Fingerprint Sourcebook, Washington, DC 20531, The Department of Justice's National Institute of Justice (NIJ), 2011.
- [3] Jansen Wayne, Daniellou Ronan, Cilleros Nicolas, "Fingerprint Identification and Mobile Handheld Devices: An Overview and

Implementation." *National Institute of Standards and Technology*, March 2006, 18 pages

- [4] Steve Gold. (2013, Nov-Dec) "Meeting the Biometrics Payment Security Challenge". *Biometric Technology Today*, [Online] Vol. 2013, Issue 10, pp. 5-8. Available: <http://www.sciencedirect.com/science/article/pii/S0969476513701759> [Nov. 12, 2014]
- [5] Stephen J. Tipton, Daniel J. White II, Christopher Sershon, and Young B. Choi. (2014, May) "iOS Security and Privacy: Authentication Methods, Permissions, and Potential Pitfalls with Touch ID" , *International Journal of Computer and Information Technology*, [Online] Vol. 03 - Issue 03. Available: <http://www.ijcit.com/archives/volume3/issue3/Paper030302.pdf> [Nov. 12, 2014]
- [6] Shri Karthikeyan, Sophia Feng, Ashwini Rao, Norman Sadeh. "Smartphone Fingerprint Authentication versus PINs: A Usability Study" in CMU-CyLab-14-012, July 31, 2014.
- [7] N. Yildirim, A.Varol. (May. 2014) "Mobile Biometric Security Systems for Today and Future", *2nd International Symposium on Digital Forensics and Security*. [Online] ISDFS'14. Available: <http://asafvarol.com/makaleler/NilayYAsafV8691.pdf> [Nov. 15, 2014]
- [8] Ming Gao, Xihong Hu, Bo Cao and Dianxin Li. "Fingerprint Sensors in Mobile Devices", in *Industrial Electronics and Applications (ICIEA)*, 2014 IEEE 9th Conference, 2014, pp. 1437 - 1440.
- [9] SaintGermain, Sonthonax Bolivar. (September 19, 2014) "Is the Battle Over for Smart-Phones?" Available: SSRN: <http://ssrn.com/abstract=2498707> or <http://dx.doi.org/10.2139/ssrn.2498707>
- [10] Hugh Wimberly, Lorie M.Liebrock, "Using Fingerprint Authentication to Reduce System Security: An Empirical Study", *IEEE Symposium on Security and Privacy*, 2012, Pp.32-46.
- [11] Tarika Bhutani, Bhawna Bhutani. (Oct. 2013) "No To Fingerprint Security System", *International Journal of Computer Science and Management Research*. [Online] Vol. 2 Issue 10. Available: <http://www.ijcsmr.org/vol2issue10/paper532.pdf> [Nov. 12, 2014]
- [12] J. Hu, " Mobile Fingerprint Template Protection: Progress and Open issues", third IEEE Conference on Industrial Electronics and Applications, Singapore, RMIT University, 2008.
- [13] Matt Reeder. "iPhone 5S fingerprint scanner explained: 10 questions about Apple's new smartphone feature answered." *Financial Post* (Sept. 13, 2013), sec. FP Tech Desk.

- [14] Apple. (2014, October 16). Apple Pay Set to Transform Mobile Payments [Online]. Available: <http://www.apple.com/apple-pay/>
- [15] Dave Tach. "How Apple will keep your fingerprints safe in the iPhone 5S". Internet: <http://www.polygon.com/2013/9/17/4741030/apple-a7-iphone5s-fingerprints-secure-enclave>, Sep 17, 2013 [Mar. 17, 2014]
- [16] J. Angeline Rubella, B. Santhosh Kumar, "Fingerprint Based License Checking for Auto-Mobiles", IEEE-Forth International Conference on Advanced Computing (ICOAC), MIT, Anna University, China, December, 2012.
- [17] Andrea Peterson and Hayley Tsukayama. "Fingerprint scanner for iPhone 5s raises privacy, security concerns." *The Washington Post* (Sept. 21, 2013), sec. PostTV.
- [18] Dan Farber. "Sen. Franken questions privacy of iPhone 5S fingerprint scanner." Internet: http://news.cnet.com/8301-13579_3-57603947-37/sen-franken-questions-privacy-of-iphone-5s-fingerprint-scanner/. Sept. 20, 2013 [Mar. 12, 2014]
- [19] Stephen Braun, Anne Flaherty, Jack Gillum, Matt Apuzo. (Jun. 15, 2013). PRISM is just part of much larger, Scarier Government Surveillance program. [Online]. Available: <http://businessinsider.com/prism-is-just-the-start-of-nsa-spying-2013-6>.
- [20] Charles Arthur. (Sept. 23, 2013). iPhone 5s fingerprint sensor hacked by Germany's Chaos Computer Club. [Online]. Available: <http://theguardian.com/technology/2013/sep/22/apple-iphone-fingerprint-scanner-hacked>
- [21] Jonathan Kent. (Mar. 31, 2005). Malaysia car thieves steal finger. [online]. Available: <http://news.bbc.co.uk/2/hi/asia-pacific/4396831.stm>

A Comparison of OSPFv3 and EIGRPv6 in a Small IPv6 Enterprise Network

Richard John Whitfield
University of Derby

Shao Ying Zhu
University of Derby

Abstract—As the Internet slowly transitions towards IPv6, the routing protocols that are used to forward traffic across this global network must adapt to support this gradual transition. Two of the most frequently discussed interior dynamic routing protocols today are the IETF’s OSPF and Cisco’s EIGRP routing protocol. A wealth of papers have compared OSPF and EIGRP in terms of converge times and resource usage, however few papers have assessed the performance of each when implementing their respective security mechanisms. Therefore a comparison of OSPFv3 and EIGRPv6 will be conducted using dedicated Cisco hardware. This paper will firstly introduce each protocol and its security mechanisms, before conducting a comparison of OSPFv3 and EIGRPv6 using Cisco equipment. After discussing the simulation results, a conclusion will be drawn to reveal the findings of this paper and which protocol performs the best upon implementing their respective security mechanisms within a small IPv6 enterprise network.

Index Terms—IPv4; IPv6; OSPFv3; IPsec; ESP; EIGRPv4; EIGRPv6; MD5

I. INTRODUCTION

Two of the most discussed IPv6 routing protocols amongst researchers are the IETF’s Open Shortest Path First Version 3 (OSPFv3) and Cisco’s Enhanced Interior Gateway Routing Protocol for IPv6 (EIGRPv6). A number of papers such as [1,2,3,4] have reviewed both protocols countless times with respect to their resource usage and convergence speed. However, no comparisons have been conducted to assess the additional effects when implementing the respective authentication and encryption mechanisms of OSPFv3 and EIGRPv6.

Therefore, due to the popularity of OSPFv3 and EIGRPv6, it is critical that a through comparison is conducted to comprehensively assess both protocols when operating within a small IPv6 enterprise network.

In addition, it should also be noted that in recent years, a key drawback of EIGRP has been its proprietary nature. However, as discussed by [5], EIGRP is being opened up to the IETF and will soon no longer be a drawback.

This paper contributes to the ongoing comparisons of OSPFv3 and EIGRPv6 by testing both protocols and assessing the additional impact of both protocol’s security mechanisms when they are implemented in a Cisco hardware based test environment.

II. OSPFV3

OSPFv3 is a dynamic routing protocol that uses the

Shortest Path First (SPF) algorithm and has been specifically designed to run within an IPv6 environment. Compared to its IPv4 equivalent OSPFv2, OSPFv3 incorporates a number of key changes necessary to operate in an IPv6 network [6].

As discussed by [7], a key change that has been performed for OSPFv3 is that the packet header of which has been restructured. OSPFv3’s packet header is now far less complex compared to that of OSPFv2 and also includes the “Instance ID” field [7]. The Instance ID field also reflects another dramatic change, in that routing protocols for IPv6 are more concerned about the links they are enabled on, rather than the nodes they are enabled on [7]. This “per-interface” concern means that multiple addresses can be configured on the same interface [8]. This is because rather than establishing neighbourhood using IP subnets, OSPFv3 uses link local addresses to establish its adjacencies.

Furthermore, the changes to the OSPFv3 packet header have also had an additional effect on the OSPFv3 Hello Packet. To reflect the changes made for IPv6, the OSPFv3 Hello packet structure has been changed [8].

These changes are as follows [7]:

- The addresses of 224.0.0.5 and 224.0.0.6 are used for passing traffic between the DR and DROther routers is now FF02::5 and FF02::6.
- IPv6 addresses in OSPFv3 are located within the payload rather than the packet header.
- Network-LSA’s do not contain any IPv6 addresses compared to OSPFv2.
- OSPFv3 requires that a router ID is configured before routing can begin.
- DR and BDR routers are now identifiable by their router ID’s instead of their IP addresses as with OSPFv2.

In addition, a key change for OSPFv3 is the security mechanisms that the protocol uses to protect its routing updates. As discussed by [7,9], whereas OSPFv2 used MD5 authentication, OSPFv3 uses the services provided by IPsec, of which is used within an IPv6 environment [10].

III. EIGRPV6

Designed by Cisco, the Enhanced Interior Gateway Routing Protocol for IPv6 (EIGRPv6) uses the Diffusing Update Algorithm (DUAL) which is also used by EIGRP in an IPv4 environment. However, unlike OSPFv3, the majority of

EIGRP's features for IPv4 have been integrated into IPv6. As discussed by [7,11], these similarities include:

- The use of DUAL to compute EIGRP Successor and Feasible Successors.
- Using bandwidth and delay as the default metrics.
- The use of Reliable Transport Protocol (RTP).
- No periodic updates.
- EIGRPv6 implements the same authentication mechanism (MD5) as EIGRP.

However, despite the almost identical properties between EIGRP and EIGRPv6, a few changes have been implemented to prepare the protocol for routing within an IPv6 environment. As further discussed by [11], these changes include:

- The use of Link Local Addresses to establish neighbor adjacencies instead of using an IP subnet. This is also the case with OSPFv3.
- EIGRP routers will use the IPv6 multicast address FF02::10 rather than the previous 224.0.0.10 multicast address.
- Like OSPFv3, EIGRPv6 is also configured on a per-interface basis rather than been globally enabled.
- The creation of a router ID is required to successfully start routing operations.

However, unlike OSPFv3, EIGRPv6 does not incorporate the use of IPsec to encrypt its routing updates, but instead uses the MD5 authentication method that was previously used in EIGRP for IPv4 [7].

IV. METHODOLOGY

To ensure that the most relevant information can be gathered to accompany the research undertaken for this paper, a clear and concise methodology is required. Therefore, a number of specific goals will be implemented to deliver the most accurate conclusion possible. These goals include:

- 1) To investigate which protocol initialises quickest from a cold start-up.
- 2) To assess OSPFv3 and EIGRPv6's ability to recover from unforeseen failures.
- 3) Analyse which protocol re-converges with minimal packet drops.
- 4) Investigate the response times of each protocol when detecting idle link changes.
- 5) Examine each protocols security mechanism and implement them to compare their operational differences.
- 6) Observe any differences upon implementing both protocols.

Furthermore, to meet the goals designated above, a series of controlled experiments will be carried out by implementing four Cisco 1841 routers and one Cisco 2960 switch, all connected through fast Ethernet ports.

Moreover, data for this paper will be gathered by using

outputs from the router's command line and packets captured in Wireshark. It should also be made clear that each test performed for either protocol will be conducted three times and then averaged to promote result reliability. In addition, each specific test will be done again to assess the additional impact upon implementing OSPFv3 and EIGRPv6's security mechanism. This test strategy ensures that the additional effects of OSPFv3 and EIGRPv6's security mechanisms can be measured, while performing each test three times to ensure result reliability.

Lastly, it should be mentioned that both protocols will be operating using the default Hello and Dead timers to ensure that the results best reflect the default behavior of both OSPFv3 and EIGRPv6.

V. EXPERIMENT SCENARIOS

So that a comprehensive and thorough comparison can be conducted, two test scenarios have been designed to assess the performance of OSPFv3 and EIGRPv6.

As figure 1 illustrates, test Scenario 1 implements a four router point to point test scenario. The purpose of this scenario is to assess the performance of both protocols when the routers are connected directly and not through another device such as a switch.

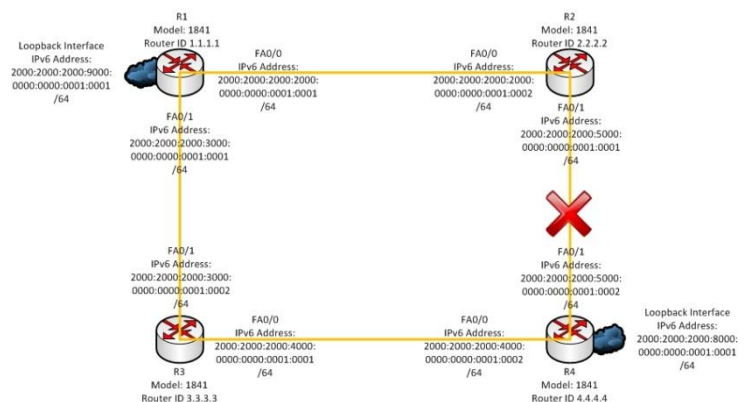


Fig. 1. Test Scenario 1

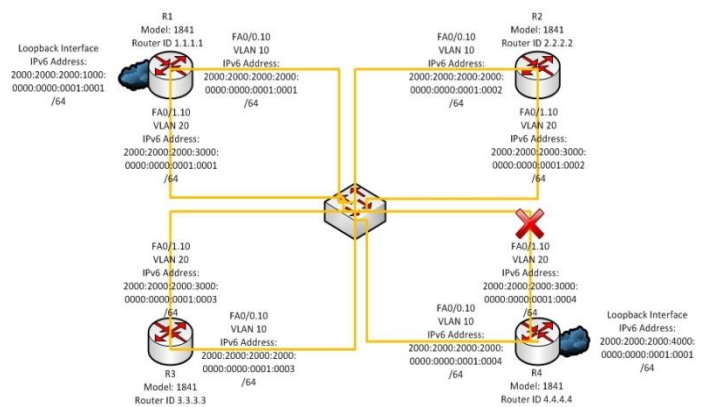


Fig. 2. Test Scenario 2

Furthermore, figure 2 demonstrates the second scenario that has been implemented to test OSPFv3 and EIGRPv6. Compared to the point to point topology of Scenario 1,

Scenario 2 implements a LAN environment where all routers connect to a switch. In Scenario 2, the switch will be configured with two VLANs to ensure the traffic for each interface is separated and kept in their own subnets. Therefore, compared to Scenario 1 where each router operates with different subnets, the routers in Scenario 2 will operate in the same subnet (one for the primary and one for the secondary link) and therefore enable an assessment of OSPFv3 and EIGRPv6's performance in a LAN environment.

Also, so that an active interface is always available to send and reply to ICMP Ping messages, a Loopback interface [12] will be implemented onto routers R1 and R4. By implementing Loopback address on R1 and R4, the traffic can be routed via another path depending on the link failed in the topology.

It should be noted for the purposes of this paper that all tests will be executed and monitored from R4's perspective and the preferred interface is FA0/1 towards R2.

VI. SCENARIO RESULTS AND ANALYSIS

This section will discuss the results generated by testing OSPFv3 and EIGRPv6 and their security mechanisms, in Scenarios 1 and 2. The results are as follows:

TABLE I. TEST SUMMARY FOR SCENARIO'S 1 AND 2

Test Details	With Auth	Scenario 1		Scenario 2	
		OSPFv3	EIGRPv6	OSPFv3	EIGRPv6
Convergence time from a cold router start-up.		181.3s	143.4s	182.2s	163.5s
	Y	180.3s	142.4s	180.6s	163.6s
Neighbour down detection after an unexpected link failure.		9.7s	7.5s	9.1s	4.7s
	Y	9.7s	9.1s	9.3s	8.3s
Traffic re-sent after an unexpected link failure.		14.4s	8s	14.9s	13.5s
	Y	13.6s	11.0s	14.3s	14.9s
Time for Protocol to detect neighbour down after cable removal		10.1s	7.2s	9.8s	8.2s
	Y	8s	10.3	7.9s	9.3s
Time to detect neighbourship re-establishment after cable replacement		49s	6.8s	43.5s	35.2s
	Y	50.5s	6.9s	43.4s	36.8s
Peak CPU utilisation		70%	70%	70%	70%
	Y	70%	70%	70%	70%

The first goal set in the methodology section previously was to investigate which protocol initialises the fastest from a cold start-up.

As figure 3 demonstrates, the testing performed for this paper reveals a series of key findings through testing in

Scenarios 1 and 2. These findings have been extracted from table I shown earlier in this section.

As shown by figure 3 below, the startup times for EIGRPv6 are significantly faster than that of OSPFv3, irrespective of the test scenario. However, a key finding is that compared to its Scenario 1 (P2P) result, EIGRPv6 took longer to start up in Scenario 2 (LAN) test environment. In addition, figure 3 also reveals that whereas EIGRPv6 performed worse in Scenario 2, OSPFv3 performed marginally better and better still when its IPSec encryption mechanism was enabled. Interestingly, EIGRPv6's MD5 authentication mechanism affected the protocols performance in Scenario 1, but had no additional effect in Scenario 2.

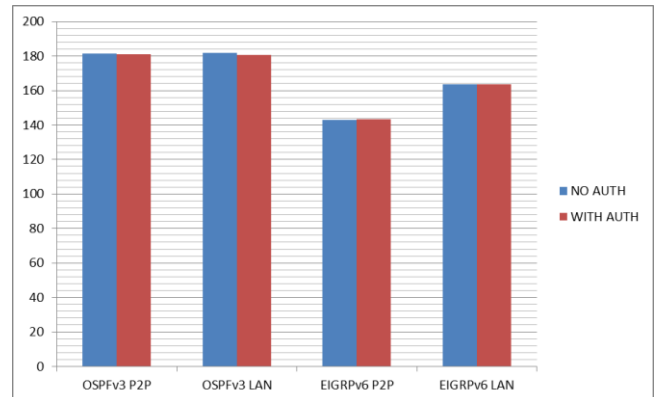


Fig. 3. OSPFv3 and EIGRPv6 Cold Start-up Time Comparison

The second and third goals that were defined in the methodology of this paper are to assess OSPFv3 and EIGRPv6's ability to recover from unforeseen failures and analyse which protocol re-converges with minimal packet drops. Therefore using the results collected in table I in addition to figures 4 and 5, this shows the averaged results from the convergence tests conducted in this paper.

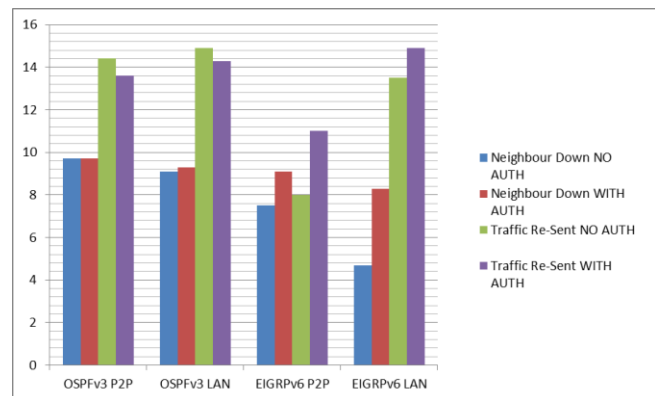


Fig. 4. OSPFv3 and EIGRPv6 Re-Convergence Times

As revealed by figure 4, a number of key findings can be found. Firstly, figure 6 continues the trend observed in figure 4 in that EIGRPv6 performance is better in Scenario 1 compared to that of Scenario 2. Although EIGRPv6 detected the neighbor was down faster in Scenario 2, it took longer to resend the traffic in Scenario 2 and also with a much bigger margin when MD5 authentication was activated.

Furthermore, figure 4 also shows that OSPFv3 performed better in Scenario 2 than Scenario 1 for neighbor detection, but was able to resend the traffic faster in Scenario 1. Moreover, when OSPFv3's IPsec encryption was configured in Scenario 2, the time taken to detect the neighbor was down increased.

However as figure 5 shows, this result may have been caused by packet drops.

Figure 5 reveals that compared to Scenario 1, OSPFv3 in Scenario 2 dropped on average more packets compared to Scenario 1. However, an interesting finding from this test is that whereas IPsec encryption improved the performance of OSPFv3, EIGRPv6's MD5 authentication adversely affected the protocols performance in both test Scenarios. In addition, figure 5 also supports the trend identified throughout this paper, in that EIGRPv6 performs better in the point to point topology of Scenario 1 compared to that of Scenario 2's LAN topology.

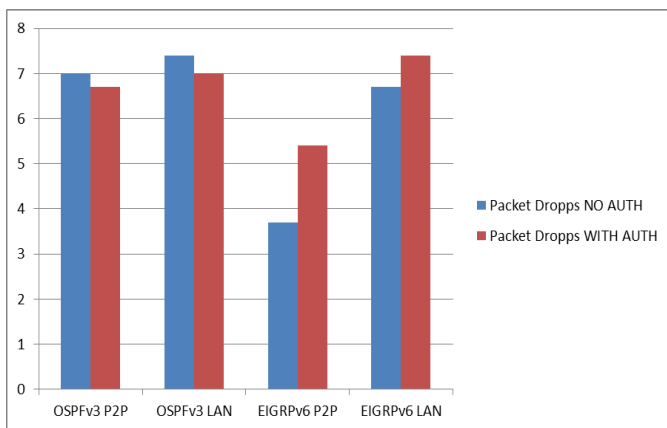


Fig. 5. OSPFv3 and EIGRPv6 Packet Drop Comparison

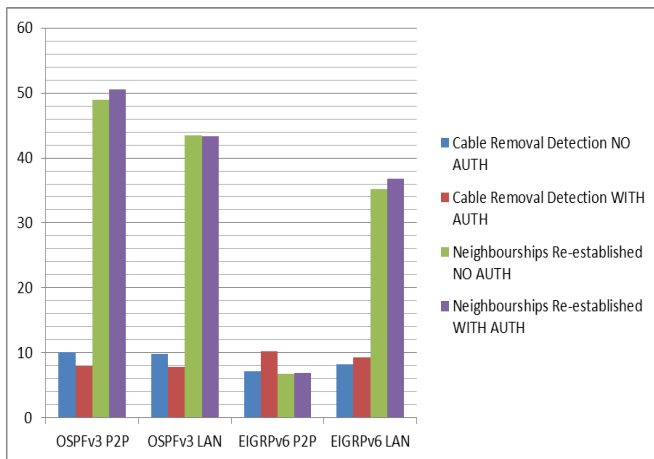


Fig. 6. Protocol Response Testing

In addition, a test to measure the responsiveness times of OSPFv3 and EIGRPv6 was carried out by deliberately failing a link over an idle link. This test differs to the convergence tests discussed earlier as the purpose is to measure the time taken for each protocol to detect and re-establish a neighbourhood, rather than detecting and re-sending traffic. This test satisfies the fourth goal in the methodology which is to analyse the response times of each protocol when detecting idle link changes.

As shown in figure 6, a number of interesting findings can be found from the results extracted from table I.

Firstly, the two ongoing trends identified throughout this paper are whereas IPsec improves the performance of OSPFv3, EIGRPv6's MD5 negatively affects its performance and that EIGRPv6 performs better in the point to point environment of Scenario 1 compared to that of Scenario 2. Figure 6 also agrees with this trend with the exception of the point to point cable removal with authentication time average.

Furthermore, a packet inspection using Wireshark was performed in addition to attempting to crack the type seven passwords implemented upon encrypting the running configurations within the Cisco 1841 routers. These tests meet the criteria for the fifth goal which is to examine each protocols security mechanism and implement them to compare their operational differences.

As shown in figures 7 and 8, the results of the packet inspection can be observed when a packet is captured and analysed using Wireshark.

```

Frame 251: 130 bytes on wire (1040 bits), 130 bytes captured (1040 bits)
Ethernet II, Src: Cisco_6f:da:77 (00:22:55:6f:da:77), Dst: Cisco_6f:dc:
Internet Protocol Version 6, Src: fe80::222:55ff:fe6f:da77 (fe80::222:5
0110 .... = Version: 6
.... 1110 0000 .... = Traffic class: 0x000000e0
.... 0000 0000 0000 0000 0000 = Flowlabel: 0x00000000
Payload length: 76
Next header: ESP (50)
Hop limit: 1
Source: fe80::222:55ff:fe6f:da77 (fe80::222:55ff:fe6f:da77)
[Source SA MAC: Cisco_6f:da:77 (00:22:55:6f:da:77)]
Destination: fe80::222:55ff:fe6f:dc53 (fe80::222:55ff:fe6f:dc53)
[Destination SA MAC: Cisco_6f:dc:53 (00:22:55:6f:dc:53)]
[Source GeoIP: unknown]
[Destination GeoIP: unknown]
Encapsulating Security Payload
ESP SPI: 0x000003ea (1002)
ESP Sequence: 4
    
```

Fig. 7. OSPFv3's IPsec ESP Encrypted Packet

As shown by figures 7 and 8, the major difference between the two packets is that compared to OSPFv3's ESP IPsec encrypted packet, EIGRPv6's MD5 authenticated packet makes no attempt to hide the information within the packet. Therefore as figure 8 shows, information such as the autonomous system number, the "K" values in use and the Hello times used by EIGRPv6 can be discovered by capturing one packet.

```
Frame 361: 134 bytes on wire (1072 bits), 134 by
Ethernet II, Src: Cisco_6f:dc:74 (00:22:55:6f:dc
Internet Protocol Version 6, Src: fe80::222:55ff
Cisco EIGRP
  Version: 2
  Opcode: Hello (5)
  Checksum: 0xd390 [correct]
  Flags: 0x00000000
  Sequence: 0
  Acknowledge: 0
  Virtual Router ID: 0 (Address-Family)
  Autonomous System: 1
  Authentication MD5
    Type: Authentication (0x0002)
    Length: 40
    Type: MD5 (2)
    Length: 16
    Key ID: 1
    Key Sequence: 0
    Nullpad: 0000000000000000
    Digest: 6fbb7488bea033b3026589f62f548dbc
  Parameters
    Type: Parameters (0x0001)
    Length: 12
    K1: 1
    K2: 0
    K3: 1
    K4: 0
    K5: 0
    K6: 0
    Hold Time: 15
  Software Version: EIGRP=5.0, TLV=3.0
```

Fig. 8. EIGRPv6 MD5 Authenticated Packet

In addition using the tools provided by [13], a test was also performed to attempt to reverse the type seven passwords stored in the routers running configuration. This type seven password is used by both protocols authenticate a neighboring router and is encrypted using the “service password-encryption” command. The Cisco Type 7 Reversing tool found on [13] will be used to reverse the passwords stored in the routers running configuration file.

Upon attempting to reverse the passwords used by both protocols, it was discovered that whereas OSPFv3’s passwords was only partially decrypted, EIGRPv6’s MD5 authentication passwords was completely decrypted and therefore revealed the authentication password required to peer with a router in the topology. The reason for this is because whereas OSPFv3’s ESP IPsec encryption requires a minimum password length of 40 characters, EIGRPv6’s MD5 authentication mechanism specifies no minimum password length. Therefore, EIGRPv6 can be configured with potentially very weak passwords and making the implemented password prone to decryption as a result.

Lastly, the final goal set in the methodology was to observe any differences upon implementing both protocols.

```
R1(config)#
R1(config)#int fa0/0.10
R1(config-subif)#ipv6 ospf 1 area 0
R1(config-subif)#
*Mar 21 16:51:16.503: %OSPFV3-4-NORTRID: OSPFV3 process 1 could not pick a router-id,
please configure manually
```

Fig. 9. OSPFv3 Router ID Prompt

As revealed in figure 9, a key difference noticed when implementing OSPFv3 and EIGRPv6 is whereas OSPFv3 generates a router ID prompt upon first configuration, EIGRPv6 does not generate this prompt for the creation of a router ID. As discussed previously, both protocols will not begin routing traffic until a router ID is created. As a result, a

network administrator may spend time debugging EIGRPv6 only to find that the protocol would not route traffic due to the lack of a router ID.

VII. CONCLUSION

This paper finds that upon comparing the performance of OSPFv3 and EIGRPv6 using the tests that have been conducted throughout this project, EIGRPv6 was the faster performing protocol. However aside from the overall conclusion, a series of thought provoking results have been found in this project. These include:

- That EIGRPv6 performed better in every test when it was configured for the point to point topology of Scenario 1. EIGRPv6’s performance was noticeably different when it was implemented into Scenario 2’s LAN environment, taking longer to recover from simulated failures and dropping considerably more packets. It can therefore be assumed from the findings that EIGRPv6 performs better within a point to point configuration, rather than a LAN environment.
- OSPFv3’s performance was relatively similar when implemented into Scenarios 1 and 2, but on average performed consistently better when IPsec was enabled. By comparison, EIGRPv6’s performance was always degraded when its MD5 authentication mechanism was enabled.
- However despite this degradation, EIGRPv6 still outperformed OSPFv3 in terms of sheer speed while converging and adjusting to failures and therefore wins the performance comparison.

Therefore, the principle conclusion from the results of this paper is that when comparing OSPFv3 and EIGRPv6 within a small flat IPv6 enterprise network, EIGRPv6 outperforms OSPFv3 in terms of start-up and re-convergence speed and is therefore the faster protocol. This conclusion has been generated by testing OSPFv3 and EIGRPv6 in both a point to point and LAN based network environment, where OSPFv3 took consistently longer to complete its operations than that of EIGRPv6.

However whereas the MD5 authentication mechanism used by EIGRPv6 negatively affected its performance, IPsec noticeably improved OSPFv3’s performance. This therefore makes OSPFv3 an attractive option to network administrators who wish to implement a routing protocol that integrates a strong security mechanism and operates within a hierarchical network topology. By comparison, EIGRPv6 is designed to operate on a typically flat network structure which may still limit its application.

REFERENCES

- [1] Wijaya, C. (2011) Performance Analysis of Dynamic Routing Protocol EIGRP and OSPF in IPv4 and IPv6 Network, Informatics and Computational Intelligence (ICI), 2011 First International Conference, pp. 335-360.
- [2] Thorenoor, S.G. (2010) Dynamic Routing Protocol Implementation Decision Between EIGRP, OSPF and RIP Based on Technical Background using OPNET Modeler, Computer and Network Technology (ICCNT), 2010 Second International Conference, pp. 191-195.

- [3] Krishnan, Y.N., G, Shobha. (2013) Performance Analysis of OSPF and EIGRP Routing Protocols for Greener Internetworking, Green High Performance Computing (ICGHPC), 2013 IEEE International Conference, pp. 1-4.
- [4] Fitigau, I., Todorean, G. (2013) Network Performance Evaluation for RIP, OSPF and EIGRP Routing Protocols, Electronics, Computers and Artificial Intelligence (ECAI) 2013 International Conference, pp. 1-4.
- [5] Savage, D., Slice, D., Ng, J., Moore, S., White, R. (2013) Enhanced Interior Gateway Routing Protocol Draft-Savage-EIGRP-00, IETF, February 2013.
- [6] Hinds, A., Atojoko, A., Zhu, S. (2013) Evaluation of OSPF and EIGRP Routing Protocols for IPv6, International Journal of Future Computer and Communication, 2(4), pp. 287-291.
- [7] Teare, D. (2010) Implementing Cisco IP Routing (Route). 4th edn. Indianapolis: Cisco Press.
- [8] Coltun, R., Ferguson, D., Moy, J., Lindem, A. (2008) RFC 5340 - OSPF for IPv6, IETF, July 2008.
- [9] Gupta, M., Melam, N. (2006) RFC 4552 - Authentication / Confidentiality for OSPFv3, IETF, June 2006.
- [10] Wen, X., Xu, C., Guan, J., Su, W., Zhang, H. (2010) Performance Investigation of IPsec Protocol Over IPv6 Network, Advanced Intelligence and Awareness Internet (ALAI 2010), 2010 International Conference, pp. 174-177.
- [11] Graziani, R. (2012) IPv6 Fundamentals. 1st edn. Indianapolis: Cisco Press.
- [12] Deering, S., Haberman, B., Jinmei, T., Nordmark, E., Zill, B. (2005) RFC 4007 - IPv6 Scoped Address Architecture, IETF, March 2005.
- [13] Packet Life (2008) PacketLife.net. Cisco Type 7 Reverser. [Online]. Available at: <http://packetlife.net/toolbox/type7/> Date of access: (January 24th 2014).

Give a Dog ICT Devices: How Smartphone-Carrying Assistance Dogs May Help People with Dementia

Chika Oshima

Faculty of Science and Engineering,
Saga University
Saga, Japan

Kiyoshi Yasuda

Chiba Rosai Hospital
Chiba, Japan
Kyoto Institute of Technology
Kyoto, Japan

Toshiyuki Uno

Akebono Day Service
Chiba, Japan

Kimie Machishima

Graduate School of Science and Engineering,
Saga University
Saga, Japan

Koichi Nakayama

Graduate School of Science and Engineering,
Saga University
Saga, Japan

Abstract—People with dementia suffer from memory loss, speech disabilities, and many other problems. A smartphone could benefit them, because it offers functions and applications that may alleviate their disabilities. However, some people with dementia refuse to carry a smartphone. Many of them dislike doing the tasks ordered by such devices due to a lack of psychological interaction. Therefore, we are exploring the concept of having a dog carry a smartphone on its back to assist these people with their daily lives. In this paper, we first show that, with a little training, a dog can be made to run to its owner when the smartphone on its back emits an alarm. This result suggested that the concept will allow applications and devices for the people with dementia to become the more useful things of their daily lives. Then, we propose an application wherein people with mild cognitive impairment can be reminded what they were going to do a few minutes ago. We also propose a support method using a vibration-sensing device that causes a dog to run up to its severe-dementia person who is trying to open a door to go outside. Finally, we describe an experiment that examined how a person with dementia might respond to a dog who “talks” to them. (Of course, the talker was a person at a different location speaking through the smartphone on the dog’s back.) These suggestions and the results of the experiment show that, with the help of a dog, a smartphone can offer better assistance for dementia patients.

Keywords—Android; Care facility; Memory loss

I. INTRODUCTION

Dementia is a collection of symptoms that include deterioration of mental abilities and cognitive functions such as memory, language, reasoning, planning, recognizing, and identifying people and objects¹. Dementia is caused by more than a hundred diseases and injuries that primarily or secondarily affect the brain. About 36 million people have dementia, and there are about 8 million new cases worldwide every year².

Many people with dementia are tended by caregivers, including their families. Yet as their symptoms become more

severe, the caregivers bear a greater burden. When the patients begin to suffer memory loss, their caregivers have to supervise their behavior. Even at the stage just before dementia, these people may forget what they were going to do a few minutes ago. As for people with severe dementia, they are apt to wander outside, requiring their caregivers to watch constantly for their safety. Caregivers also need to communicate frequently with their charges, even if it means saying the same things over and over. In short, there are plenty of things for the caregivers to do in the daily life of a person dementia, even at a time when staffing, time, and finances are becoming scarce.

Many studies have explored the use of information communication technology (ICT) devices to aid with memory [1][2][3]. Yasuda et al. [4] evaluated the use of a digital voice recorder as a voice output memory aid. Their results showed that such a recorder assists patients with prospective memory impairment. Kamimura et al. [5] examined the efficacy of a medication reminder device. They found that it can improve medication adherence in elderly patients with mild cognitive impairment.

The Global Positioning System (GPS) is often used to search for persons who have wandered off [6][7]. “iWander [8]” is a device that collects GPS and other sensor data about location, weather conditions, stage of illness, etc. This data is then evaluated using Bayesian network techniques to determine the probability the person is wandering. Lin et al. [9] proposed a real-time method for detecting wandering based on an individual’s GPS traces. This method is able to detect loop-like traces on the fly. The experimental results showed its effectiveness in detecting wandering behavior. However, GPS cannot prevent the person from going outside in the first place.

At present, some robots can communicate with people and alert them when it is time to take their medication. They were developed for healing and therapeutic use in private and nursing homes. “Paro³” is a therapeutic robot that can locate the source of a voice and recognize words such as its name, greetings, and even praise. By interacting with people, Paro

¹Janssen: Dementia, <http://www.dementia.com/index.html>

²World Health Organization: Dementia, <http://www.who.int/mediacentre/factsheets/fs362/en/>

³PARO Robots U.S., Inc.: Paro, <http://www.parorobots.com/index.asp>

acts as if it is alive, moving its head and legs, and making sounds. “Palro⁴” is a small, autonomous humanoid robot which can have an intelligent conversation and walk on two legs. Once the user programs information into its computer, Palro alerts him or her at the appropriate time. “Pepper⁵” is a humanoid robot that can converse with a person, recognize and react to their emotions, and move and live autonomously. As the person continues to interact with Pepper, it will recognize its person and learn new things about his or her tastes. People who carry GPS-equipped smartphones can find friends by using a mapping service. “emopa⁶” is a smartphone service that talks a user of this smartphone like family or friends.

Today, even health professionals use their smartphones to alert them about important tasks. However, these applications have limited use for people with dementia [10]. First of all, most of them forget where their smartphones are located [11]. Second, due to a lack of personal interaction, some people are reluctant to perform the daily tasks instructed by the devices [12]. Third, the cost of these robots is still too high for most homes and facilities. It is also difficult for a robot at present to chase after a person and run up stairs. Although these robots might be accepted by some people with dementia, others may consider them as “alien invaders.”

To overcome these shortcomings, we had the idea of mounting an ICT device on a dog [13]. Now people with dementia would not have to remember to carry their phones. Dogs are always happy to accompany their owners, even those with dementia. With a little training, a dog can be taught to rush to its owner when the smartphone on its back emits an alarm. Dogs can run to their owners even up a flight of stairs. Dogs have already been widely used in therapy [14]. Animal-assisted therapy is effective for the treatment of agitation/aggression and depression in patients with dementia [15][16][17]. People with dementia might be more willing to perform tasks if their dogs brought the smartphones.

In the next section, we compare the effectiveness of a smartphone on a dog to that of a stationary smartphone. We built an application where the user can set an alarm and display a message highlighting particular tasks that have to be performed at specific times. In the third section, we show another application for smartphones. Even people with mild cognitive impairment are prone to forget a task that they just did and what they were going to do next. With our app, they can easily set an alarm to remind them of what they had planned to do. In the fourth section, we propose a support method for caregivers. People with dementia might try to open a door to wander outside. But if a smartphone-equipped dog runs up, it may distract the person from going out. In the fifth section, we describe an experiment where a person with dementia converses with someone in another location via a smartphone on a dog. In the sixth section, we discuss the benefits of affixing smartphones to dogs. The final section concludes this paper.



Fig. 1: The dog mounts the smartphone on its back.

II. A SMARTPHONE ON A DOG’S BACK VS. A STATIONARY SMARTPHONE

In this section, we compare the effectiveness of a smartphone attached to a dog’s back to a stationary smartphone [18].

A. Development of the Application

We built an application for an android smartphone, the FleaPhone CP-D02. It was developed by Java Version 7 Update 21 using a development kit, Android SDK 1.0. The display of the application consists of three parts: setting the alarm, inputting a message, and a completion button. A user (usually the caregiver) can set an alarm for a particular time and input a message telling the person with dementia to begin or complete a task.

B. Subject

The subject in this experiment is a healthy person in her 50s. She has a five-year old female toy poodle that is kept indoors. Fig. 1 shows the dog with the smartphone on its back. It took one week for the dog to become accustomed to having the smartphone tied to its back. The subject trained the dog to run to her when the smartphone emitted a specific sound. This training took only three days. The experiment was conducted after a month of continuous training.

C. Method

Two identical smartphones were prepared with the application. Both phones used in the study were the same. Fig. 2 shows the two conditions of the experiment. One smartphone (named “Set-A”) was mounted on the dog, and the other (“Set-B”) was placed in a predetermined location in the living room. The sound of the alarm was different for Set-A and Set-B. The volume of the alarm was the same for each. The volume was low enough that someone sitting in the next room could not hear it.

An experimenter set the time when each smartphone would emit a sound on that day. The study was conducted for five days over the course of one week. Each study day lasted from 9 a.m. to 9 p.m. These 12 hours were divided into four parts. In each part, each smartphone emitted an alarm at a random time. The subject had a maximum of eight chances of hearing the alarm. The subject did not know when the alarms would sound. The subject was required to turn off the alarm and to

⁴Fujisoft Incorporated: Palro. <http://palro.jp/>

⁵Softbank: Pepper. <http://www.softbank.jp/robot/>

⁶Sharp: emopa. <http://www.sharp.co.jp/products/sh01g/service/emopa/>

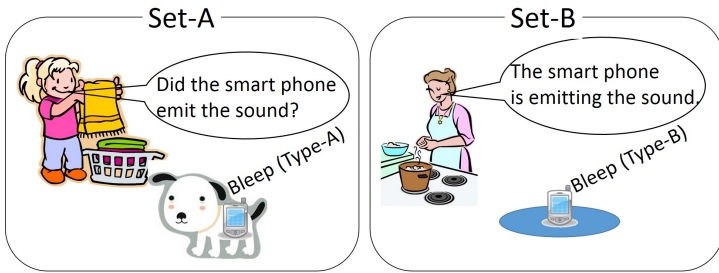


Fig. 2: Setup for the experiment.

perform an allotted task. The smartphones recorded the length of time before the alarm was turned off.

We employed the Kraepelin test⁷ as a task. This test requires an individual to perform calculations as fast and accurately as possible. The test is a boring task and involves mental stress similar to that experienced by people with dementia who have to take medication. Each test was allowed 30 seconds to complete. The application recorded the number of questions answered and the number of correct answers. After the experiment, the subject was asked to fill out a questionnaire. The question items are as follows:

- 1) Did you experience any difficulties when you used the smartphone in the experiment?
- 2) How did you feel about your dog before the experiment?
- 3) How did you feel about your dog when it responded to the alarm and came to you?
- 4) How did you feel when the smartphone fixed in the living emitted the alarm? The answers are scored from 1 to 5, with 1 denoting "I did not think at all" and 5 denoting "I thought so very much."
 - a) I was happy.
 - b) I was nervous.
 - c) I wanted to turn off the alarm as soon as possible.
 - d) I felt that it was troublesome to turn off the alarm.
 - e) I considered leaving the emitting alarm.
- 5) Please state how you felt when the fixed smartphone emitted the alarm.
- 6) How did you feel when the smartphone mounted on the dog emitted the alarm? The answers are scored from 1 to 5, with 1 denoting "I did not think at all" and 5 denoting "I thought so very much."
 - a) I was happy.
 - b) I was nervous.
 - c) I wanted to turn off the alarm as soon as possible.
 - d) I felt that it was troublesome to turn off the alarm.
 - e) I considered leaving the emitting alarm.

- 7) Please state how you felt when the smartphone mounted on the dog emitted the alarm.
- 8) How did you feel when you performed the Kraepelin test on the smartphone fixed in the living room? The answers are scored from 1 to 5, with 1 denoting "I did not think at all" and 5 denoting "I thought so very much."
 - a) I enjoyed calculating.
 - b) I enjoyed manipulating the smartphone.
 - c) I was nervous.
 - d) I thought that I should calculate as quickly as possible.
 - e) I thought that I should correctly answer as many questions as possible.
 - f) I reluctantly performed the test.
- 9) If you had other feelings or an emotional shift when you were performing the test on the fixed smartphone, please write those down.
- 10) How did you feel when you performed the Kraepelin test using the smartphone that was mounted on the dog? The answers are scored from 1 to 5, with 1 denoting "I did not think at all" and 5 denoting "I thought so very much."
 - a) I enjoyed calculating.
 - b) I enjoyed manipulating the smartphone.
 - c) I was nervous.
 - d) I thought that I should calculate as quickly as possible.
 - e) I thought that I should correctly answer as many questions as possible.
 - f) I reluctantly performed the test.
- 11) If you had other feelings or an emotional shift when you were performing the test on the smartphone mounted on the back of the dog, please write those down.
- 12) What was the dog doing while you performed the test using the smartphone mounted on its back?
- 13) What did the dog do after you completed the test?

D. Results

Table I shows the times that the phones sounded the alarm and the length of time until the subject silenced it. A blank space means that the subject did not turn off the alarm within 60 seconds. The subject silenced the alarm within 22-54 seconds in Set-A and within 22-60 seconds in Set-B. The average time to silence the alarm for Set-A and Set-B was 35.77 and 37.44 seconds, respectively (SDs = 10.0, 13.8, respectively).

Fig. 3 shows the number of times that the subject turned off the alarm. The subject turned off the alarm 13 times in Set-A and nine times in Set-B. The rate of silencing the alarm (success) was 76.47% and 52.94%, respectively ($z = 1.08$, no difference).

In Set-A, the number of times the subject did not turn off the alarm was four. In two of these, although the dog responded to the alarm, it was unable to bring the smartphone to the subject within 60 seconds. On these occasions, the owner was either out of the house, or upstairs. On the other two occasions, the alarm sounded while the dog was taking a nap.

⁷Google Play: Kraepelin test for job hunting and brain training (in Japanese). <https://play.google.com/store/apps/details?id=jp.lumireis.kraepelin>

TABLE I: The times that the smartphones emitted the sounds.

Day	section	Set-A (dog)		Set-B (stationary)	
		time	length of time (sec.)	time	length of time (sec.)
1	1	15:28	25	16:15	-
2	2	9:42	39	9:39	40
	3	13:48	35	14:25	60
	4	16:00	22	17:41	-
3	5	18:31	54	18:03	22
	6	10:12	35	9:34	32
	7	12:37	22	14:53	-
4	8	16:16	53	16:52	24
	9	18:44	-	19:42	-
	10	11:52	46	11:34	60
5	11	14:14	37	13:08	-
	12	17:25	-	16:57	37
	13	20:19	-	19:25	-
M	14	11:37	36	9:44	40
	15	12:13	31	13:32	-
	16	15:02	30	17:51	-
	17	18:08	-	19:43	22
M	-	35.77	-	37.44	-
SD	-	10.0	-	13.8	-

TABLE II: The number that the subject calculated and the number of correct answers.

day	section	Set-A (dog)				Set-B (stationary)			
		time	correct	sum	rate (%)	time	correct	sum	rate (%)
1	1	15:28	37	40	92.5	16:15	-	-	-
2	2	9:42	45	46	97.8	9:39	35	39	89.7
	3	13:48	42	42	100.0	14:25	40	43	93.0
	4	16:00	44	46	95.7	17:41	-	-	-
3	5	18:31	41	42	97.6	18:03	34	38	89.5
	6	10:12	43	46	93.5	9:34	39	43	90.7
	7	12:37	43	47	91.5	14:53	-	-	-
4	8	16:16	32	39	82.1	16:52	42	45	93.3
	9	18:44	-	-	-	19:42	-	-	-
	10	11:52	43	47	91.5	11:34	42	45	93.3
5	11	14:14	34	41	82.9	13:08	-	-	-
	12	17:25	-	-	-	16:57	32	36	88.9
	13	20:19	-	-	-	19:25	-	-	-
M	14	11:37	38	44	86.4	9:44	31	41	75.6
	15	12:13	44	49	89.8	13:32	-	-	-
	16	15:02	42	45	93.3	17:51	-	-	-
	17	18:08	-	-	-	19:43	44	46	95.7
M	-	-	40.62	44.15	92.0	-	37.67	41.78	90.2
SD	-	-	4.0	3.0	-	-	4.5	3.3	-

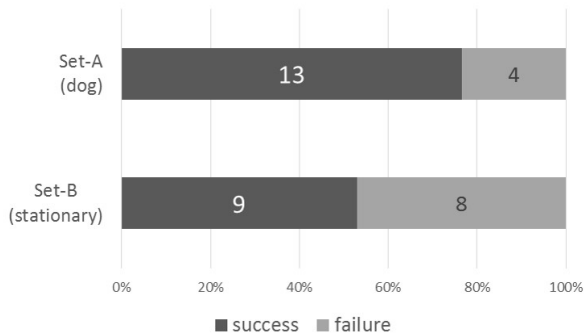


Fig. 3: the number of times that the subject turned off the alarm.

In Set-B, the subject failed to silence the alarm eight times. The subject could not hear these alarms because she was not in the living room.

Table II shows the number of calculations made by the subject and the number of correct answers in the Kraepelin test. The number of calculations was 39-49 in Set-A and 36-46 in Set-B. The average number of questions answered in Set-A and Set-B was 44.15 and 41.78, respectively (SDs = 3.0, 3.3, respectively).

In Set-A, the number of correct answers was 32-45, whereas it was 31-44 in Set-B. The average number of correct answers was 40.6 in Set-A and 37.7 in Set-B (SDs = 4.0, 4.5, respectively).

We compared the results of Set-A and Set-B. In both, questions in some sections were not answered because the

subject had not heard the alarm and thus had not performed the test. Hence, we include only the results of those sections where the subject completed both Set-A and -B: sections 2, 3, 5, 6, 8, 10, and 14. We compared the medians of the number of all answers between Set-A and Set-B (signed-rank test). The results did not show a significant difference between sets (two-sided test, $T=7.0$, $P=0.297$). In addition, we compared the medians of the number of correct answers between Set-A and Set-B (signed rank test). The results revealed no significant difference between the sets (two-sided test, $T=6.5$, $P=0.297$).

Table III shows the responses of the subject to the questionnaire. The subject did not find it difficult to operate the smartphone (see (1)). Before the experiment, she expressed a mixture of anticipation and anxiety about her dog (see (2)). But when the dog responded to the alarm and came to her, she was impressed (see (3)). When the dog's smartphone sounded, she was a little happier than when the stationary smartphone sounded (see (4) and (6)). She explained why she could not always silence the smartphones (see (5) and (7)). In the case of the living room smartphone, when she was cooking she did not hear the alarm. In the case of the dog-mounted phone, the dog was taking a nap in the early evening and did not wake up by the alarm. Therefore the dog did not bring the phone to the subject when the alarm sounded.

She was happy to perform the Kraepelin test on either set (see (8),(9), and (10)). The Kraepelin test was meant to simulate the boredom/stress encountered in repetitive-drug taking. We wanted to simulate the situation where an individual with dementia would not want to take their medication. We expected that the subject would be more willing to perform the tests presented by the dog-mounted smartphone than the tests on the fixed phone. However, the results showed no difference in the subject's willingness to perform the tests on either phone

TABLE III: Questionnaire response.

(1)	None.					
(2)	I believed that my dog ran to me when the smartphone emitted an alarm because it was trained to do so. On the other hand, as the dog was unwilling to put the wear, I wondered if the dog runs.					
(3)	As the dog is small, it might find the smartphone too heavy. I was impressed that the dog came to me when the alarm sounded. I was delighted that our daily training yielded results. I realized how important trust is between people and dogs.					
(4)	(a) 3	(b) 2	(c) 2	(d) 1	(e) 1	
(5)	I did not notice that the smartphone was emitting an alarm while cooking because the alarm was quiet.					
(6)	(a) 4	(b) 2	(c) 2	(d) 1	(e) 1	
(7)	The dog did not respond to the alarm in the early evening because it took a nap at that time.					
(8)	(a) 5	(b) 3	(c) 2	(d) 5	(e) 4	(f) 1
(9)	I enjoyed the test because it was a brain-training exercise. Sometimes, I touched the different answer from what I considered because of hasty move. I was concerned about the accuracy of my answers.					
(10)	(a) 5	(b) 3	(c) 2	(d) 5	(e) 4	(f) 1
(11)	None.					
(12)	The dog sat quietly when I performed the test.					
(13)	The dog made a point of shaking itself after it stood up.					

(see (8) and (10)).

E. Discussion

There was no significant difference between Set-A and -B with regard to the number of times that the subject turned off the alarm. Moreover, there was no significant difference between the sets in the number of questions answered or in the number of correct answers. If the same experiment was conducted with a large number of subjects, we would expect to observe significant differences between Set-A and -B.

It is clear that the subject was quicker to turn off the alarm on Set-A because the dog was trained to run to her when the alarm emitted. Contrary to our expectations, it took only three days to train the dog. However, the time of training may differ depending on the character of the dog and the relationship with its owner beforehand .

We expected that a person with dementia would be more willing to perform tasks if his/her dog brought over the smartphone. In our experiment, we assumed that the Kraepelin test would be boring. Surprisingly, the subject, a healthy person, enjoyed the test on either Set-A or -B; she described it as a brain-training exercise. On the other hand, she said that when the dog's phone emitted the alarm, it made her a little happier than hearing the fixed phone. She was also impressed when her dog took the smartphone to her. Therefore, we conclude that individuals with dementia will be more willing to perform daily tasks if influenced through their dogs. This result suggested that other applications for the people with dementia also become the more useful things of their daily lives by the dogs who mount the smartphones.

III. AN APP TO REMIND PEOPLE OF THE TASKS THEY MUST DO NOW

In this section, we develop a smartphone application that reminds people with mild cognitive impairment of which tasks they have to do at this moment.

A. Background

Mild Cognitive Impairment (MCI) is a stage just before dementia. People with MCI have a risk of progressing to dementia. MCI lies between decline of general aging and that of dementia. People with MCI experience difficulties with memory, language, thinking, decisions, planning, and judgment. In most cases, these changes hardly impair their day-to-day living. Most homemakers with MCI can continue to perform their household chores. However, homemakers with MCI often forget the task which they need to do now or were planning to do soon.

We offer an example of a homemaker with MCI who forgot an important task. This case was based on an event that was written up by one of the authors. It dramatically shows the need to assist people with MCI.

Mary is a homemaker with MCI. While she was making dinner in her kitchen, the telephone rang. She stopped chopping a carrot and answered the phone. The caller was a neighbor who is treasurer of their local neighborhood association. After the call, Mary went to the neighbor's home and paid her annual membership fee. When she returned home, she sat down in her living room. She opened her notebook PC and started checking her E-mails. Her daughter came into the living room and said, "Mom, I'm hungry." Only then did Mary realize that she still had to make dinner. She hurried back to the kitchen. She had just placed the pot of vegetable soup over the fire, when her smartphone rang. Her husband was calling to say he would be home soon. Then she cleaned a bathtub, went into the living room and sit down in the front of her PC and started checking her E-mails. Meanwhile, the soup in the pot was beginning to burn. She never noticed it.

People with MCI may experience an impairment in short-term memory. However, if someone tells them what they have to do, they can do that task. Most of us write down our plans on a calendar or in an appointment organizer. However, we usually don't need to write down tasks that we are going to do within a few minutes. Yet these are the very near-term tasks that people with MCI often forget. Fortunately, the kinds of the jobs that homemakers do within the household are few and can be listed. In this section, we construct an application that lets people with MCI set an alarm to remind them of what they have to do now. And if they have a dog in the house, that dog can carry the smartphone and run to the householder when the alarm goes off.

B. Structure of the Application

We built an app for an android smartphone, an XPERIA C6903. It was developed by Eclipse SDK 4.3 and works on Java Version 8. Fig. 4 shows a flow chart for operating the app. When some new task requires a person with dementia to interrupt the current task, he/she touches the start button on the app's display. This display shows twelve pictures and a

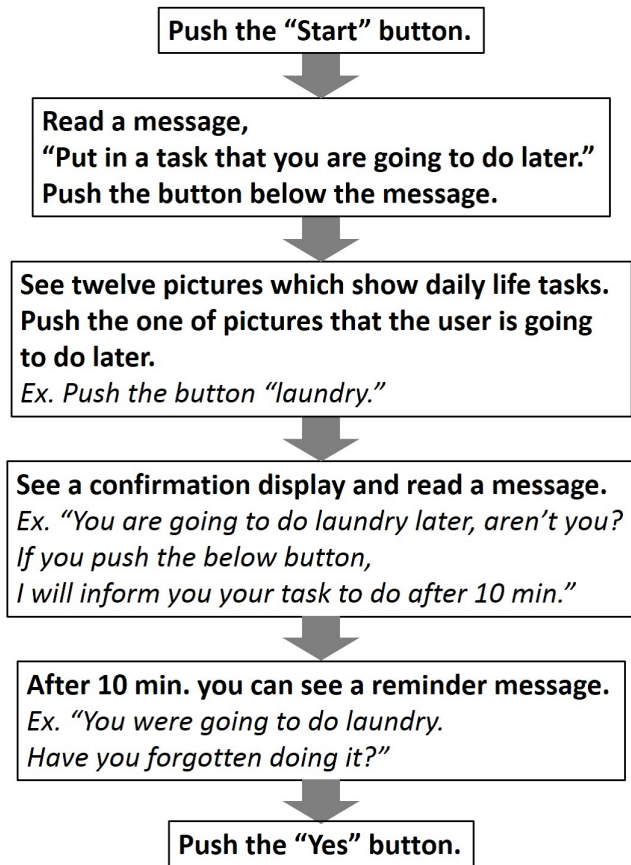


Fig. 4: How to operate the application.

message “What are you planning to do soon? Please touch the button matching your task.” The twelve pictures show typical daily chores: cooking, doing laundry, ironing, house cleaning, feeding the dog, taking a bath, making an important phone call, taking medication, preparing a meal, going shopping, brewing tea, and using the PC. The user pushes the button which corresponds to what he or she intends to do after a few minutes. After a predetermined time (5 minutes, 10 minutes, or some other preset amount), the smartphone alerts the user and displays another message, for example: “You were planning to do laundry. Do you still wish to do it?”

C. Using the Application

The app can be downloaded to any smartphone. The user invokes the application whenever something interrupts his or her planned task. When the smartphone gives the alarm, the dog carrying it runs up to the user. The user is reminded of what he or she was going to do.

IV. HOW A DOG CAN PREVENT A PERSON FROM OPENING A DOOR

In this section, we propose a method whereby a dog equipped with a smartphone distracts a person from attempting to open a door.

A. Background

In some homes and institutions, caregivers must watch over those people with dementia. That is because these people often try to open a door to go outside. Locking everything down is an imperfect solution, since it interferes with the caregivers’ own movements.

A person with dementia will stop opening a door if the caregiver speaks to them. However, caregivers cannot watch everyone all the time. Many minutes might pass before they discover that someone has gone outside.

However, if a dog can be trained to play with the patient until the caregiver shows up, then the dog may prevent that person from going through the door.

B. Setting the Devices

We used two different devices: a vibration sensor and a smartphone. For this experiment, we used a loss-prevention tag, an REX-SEEK1-X, as the vibration sensing device. To test the feasibility of the concept, we hung a single tag on a selected door. An application linked to the Bluetooth-compatible tag causes a smartphone to sound an alarm when the door is vibrated.

In an extended application, the smartphone can be made to play predetermined messages to the person with dementia. For example, the smartphone says “Don’t go away!” “Play with me.” or “Shall we go over there?”

C. Training the Dog

A few things are needed to train a dog for this application. First, when the smartphone on the animal’s back sounds off, its owner should call it to the desired door. If the dog runs to the owner, it gets a little treat from a box in front of the door. For the next step, the dog’s treat is placed in the box in advance. When the alarm sounds, the dog becomes accustomed to finding the treat in front of the door. For the last step, the owner reduces the size of the treat. At last, the dog will run to the door whenever the alarm goes off, even if there is no treat.

D. Using the Devices

Fig. 5 shows the setup of the devices. The smartphone-equipped dog runs up to the person when the alarm goes off. The dog might twirl around or bark. The person with dementia will notice the dog and forget about the door. A caregiver also notices the dog making a commotion and rushes over to the scene. Moreover, we expect that the person with dementia hears the messages from the smartphone and may come back by him/herself.

V. PEOPLE WITH DEMENTIA TALKING THROUGH A SMARTPHONE ON A DOG

In this section, we describe an experiment where a person with dementia converses through a smartphone on a dog’s back. A second person at a different location chatted with the first person via Skype over the smartphone.

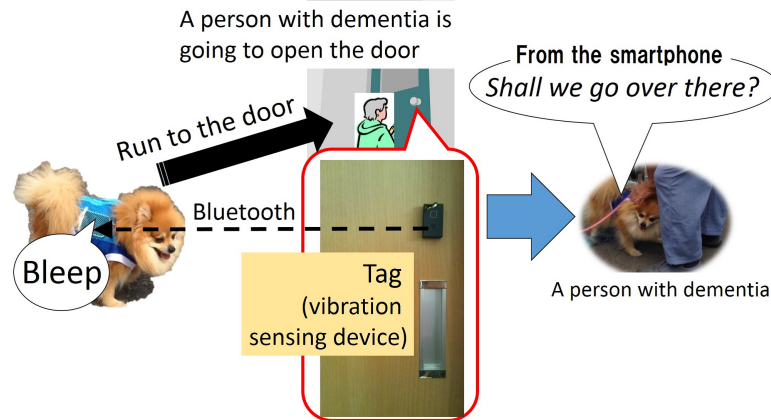


Fig. 5: Dog runs to the door when the alarm emits.

A. Method

We conducted an experiment in which one of authors, who was at a different location, spoke with a person with dementia at facility X through Skype on the smartphone that was on the dog. We wanted to see if a person with dementia would respond to a human voice that seemingly came from a dog. Facility X is a day-care facility in Japan. People with mild as well as medium dementia attend the facility for a few days every week.

The director of this facility has two dogs, both female Pomeranians. One of them cooperated with the experiment. She is young and easily attaches to a person. She was equipped with a harness that had a smartphone in one pocket.

Eight people with dementia (the users) were enjoying a snack at the facility after taking walk. They sat in a half-circle in the garden and ate ice shavings from a cup. The author talked to different users for thirty minutes. Fig. 6 shows the users eating ice shavings while hearing someone talking to them through the smartphone on the dog. The staff of the facility were given an explanation of the situation and consented to our experiment. However, the users of the facility were not told that someone would try to talk to them through a smartphone.

B. Result

The dog accepted the harness the very first time. The dog walked around or sat at the feet of the users who were snacking on ice shavings.

Some users responded to the voice through the smartphone. We offer a few samples of conversations. “D” is the author speaking via the dog. “U” is one of the users of the facility. “S” is a staff member.

Phrases in parentheses indicate what the speaker was doing.

The following conversation took place with a male user (U1) having mild dementia.

D: Did you like it?

U1: Is it tasty? (He spoke to the dog.)



Fig. 6: People with dementia hear that someone talks to them through a smartphone mounted by a dog.

D: It's yummy.

U1: Ha ha ha!

S1: It just said “It's yummy,” didn't it?

S2: It said, “It is yummy.”

U1: Is it really tasty? (He spoke to the dog.)

The staffer's question elicited the user's final response.

The next conversation occurred when a female user with mild dementia (U2) showed a cup of ice shavings to the dog.

D: What are you going to eat?

U2: We are going to eat this, here. (She showed a cup of ice to the dog.)

The dog was not furnished with a camera. The author could not see the cup. However, U2 showed the cup to the dog and

said “this” quite naturally.

The next conversation took place with a female user who had mild dementia (U3) and a twisted sense of humor.

D: I would like to eat ice shavings with you.

U3: Do you want to eat with me?

D: Yeah.

U3: I want to eat alone.

When U3 returned from her walk, and before eating the ice shavings, she heard a voice coming from nowhere. A staff member told her that it was the dog’s voice. U3 said “No, it is ghost-like, ha-ha-ha!” She must have realized that the speaker was not a dog.

C. Discussion

Most of the users with mild dementia responded to the dog “speaking.” However, the user with medium dementia hardly uttered a word. We can think of some reasons for this. Even when a human unexpectedly talks to someone with medium dementia, he or she does not always react at once. Furthermore, it is not easy for one with medium dementia to comprehend every situation, especially if that situation is something unexpected such as a talking dog. They might be a little confused or avoid confronting the situation because their cognitive functioning has been reduced.

We think that U3 should have realized that the voice did not belong to the dog. Apparently she did not know or want to know where the voice was coming from. However, she seemed to enjoy responding to it.

A dog, as a medium of conversation, may allow a user to be more open with their feelings. U3 reacted to the dog with wry humor. But if the conversation partner had been a child, the child might have felt hurt. And if the conversation partner was a staff member, U3 would never have said such a thing. This result suggests that a dog allows elderly people to be more playful in their talk.

The day will soon come when a smartphone will be able to hold a conversation autonomously. The smartphone will also present topics of general interest during elderly people [19][20]. But for the present, a lack of psychological interaction means people with dementia are reluctant to talk with a smartphone. But when a dog carries the phone, these people may have an enjoyable conversation. A smartphone can do more than give people with dementia a lot of information. In the presence of a dog, it can offer such people the pleasure of conversation. In general, we talk to a dog and infer its feelings from its responses and actions. A voice from a smartphone will support this conversation. The smartphone will encourage new relationships between people and dogs who work for those with dementia every day.

One problem revealed in this study was that it is not easy to hear a smartphone outdoors. This problem is especially acute for elderly people whose hearing is diminished. This requires extra effort and cognitive ability to for anyone assisting elderly people. There were cases where the staff at the facility had to tell the users what the dog (the author) had just said. We need a better speaker that will allow the voice to be amplified in the same direction as the head of the dog. Although U3 heard the

voice before eating the ice shaving, she did not know where the voice was coming from. This way, the users will easily recognize to whom the dog is talking.

VI. DISCUSSION

The symptoms of dementia are varied and increase with time. Some devices can compensate for the disabilities that arise with dementia. However, these devices must be kept on hand according to the person’s stage or symptoms of dementia. Today, many people possess a smartphone that can be equipped with many useful functions and applications. The smartphone offers a lot of possibilities to provide support through only one device. However, some people with dementia forget to carry their smartphone. Also, many of them do not like to perform the tasks instructed by their device.

The results in Section II showed that a person with a dog could enjoy the benefits of a smartphone, even if she did not have it in her possession. A dog can be trained to bring the smartphone to its owner whenever an alarm sounds. Our results showed that an owner is happier when the smartphone on the dog gave the alarm. We submit that a dog can overcome the problems of non-possession of the smartphone and disinclination of doing daily tasks.

In Sections III, IV, and V, we presented an application and two methods that can assist people with dementia. Generally, dogs love to accompany their owners. If an owner goes to another place, the dog will trot after him. He can then be reminded to input necessary information into a smartphone app. Dogs are good conversation partners. An app which presents topics of general interest during elderly people will support the conversation between the dogs and the people with dementia [19][20]. Moreover, a dog can be trained to go to a door when a specific alarm goes off, and deter a person from going outside. In this way a smartphone, with a little help from a dog, can offer superior support for people who need it.

These results suggest that the concept of having a dog carry a smartphone on its back will allow other applications [21] and devices for the people with dementia to become the more useful things of their daily lives.

VII. CONCLUSION

For this paper, we conducted an experiment to compare the effectiveness of a smartphone mounted on a dog’s back to that of a stationary smartphone. The results showed that, after a little training, a dog will rush up to its owner when the smartphone emits an alarm. Then, we presented an application that people with mild cognitive impairment can easily set to remind them what they are going to do inside of a few minutes. Because dogs usually follow their owners, these people can be encouraged to remember to input the necessary information into an application on the smartphone carried on the dog. We also proposed a method in which a dog runs up to a person who is trying to open a door and go outside. This dog is trained to run to the door when an alarm sounds on the smartphone. The dog then distracts the person’s attentions from the door until a caregiver shows up. For the last part, we conducted an experiment to examine how people with dementia respond to a dog who “talks” to them. People with mild (but not medium) dementia responded to the dog’s “voice.” We considered that

they would realize that the words were not coming from the dog. We posited that if a dog carries a smartphone, people with dementia might enjoy to talking to it, thus overcoming the psychological barriers to such devices.

For the future, we plan to give a dog other devices; a small camera, a RFD tag, an acceleration sensor, and conduct experiments through the cooperation of a care facility and a home with a dog.

ACKNOWLEDGMENT

The contents of Section II was presented at the 16th International Conference on Human-Computer Interaction in 2014.

REFERENCES

- [1] J. Hoey, P. Poupart, A. V. Bertoldi, T. Craig, C. Boullier, A. Mihailidis: Automated Handwashing Assistance for Persons with Dementia Using Video and a Partially Observable Markov Decision Process, *Computer Vision and Image Understanding*, 114(5), 503-519, 2009.
- [2] N. Kuwahara, K. Yasuda, N. Tetsutani, K. Morimoto: Remote Assistance for People with Dementia at Home Using Reminiscence Systems and a Schedule Prompter, *International Journal of Computers in Healthcare*, 1(2), 126-143, 2010.
- [3] HH. Huang, H. Matsushita, K. Kawagoe, Y. Sakai, Y. Nonaka, Y. Nakano, K. Yasuda: Toward a Memory Assistant Companion for the Individuals with Mild Memory Impairment, *Cognitive Informatics & Cognitive Computing*, 2012 IEEE 11th International Conference on, 295-299, 2012.
- [4] K. Yasuda, T. Misu, B. Beckman, O. Watanabe, Y. Ozawa, and T. Nakamura: Use of an IC Recorder as a Voice Output Memory Aid for Patients with Prospective Memory Impairment, *Neuropsychol Rehabil*, 12(2), 155-166, 2002.
- [5] T. Kamimura, R. Ishiwata, and I. Inoue: Medication Reminder Device for the Elderly Patients With Mild Cognitive Impairment, *American Journal of Alzheimer's Disease & Other Dementias*, 27(4), 238-242, 2012.
- [6] Y. Dahl, K. Holbø: "There are no secrets here!": Professional Stakeholders' Views on the Use of GPS for Tracking Dementia Patients, *Proceeding of MobileHCI '12*, 133-142, 2012.
- [7] B. McKinstry, A. Sheikh: The use of global positioning systems in promoting safer walking for people with dementia, *Journal of Telemedicine and Telecare*, 19(5), 288-292, 2013.
- [8] F. Sposaro, J. Danielson, and G. Tyson: iWander: An Android application for dementia patients, *Conference Proceedings of Engineering in Medicine and Biology Society (EMBC)*, IEEE, 3875-3878, 2010.
- [9] Q. Lin, D. Zhang, H. Xiaodi, N. Hongbo, and Z. Xingshe: Detecting wandering behavior based on GPS traces for elders with dementia, *Control Automation Robotics & Vision (ICARCV)*, 672-677, 2012.
- [10] M. H. Mohd, A. N. Mohamad: A Study of Smartphone Usage and Barriers among the Elderly, *User Science and Engineering*, 2014 3rd International Conference on, 109-114, 2014.
- [11] N. Armstrong, C. Nugent, G. Moore, D. Finlay, W. Burns: Inactivity Monitoring for People with Alzheimer's Disease Using Smartphone Technology, *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 83, 313-321, 2012.
- [12] K. Yasuda: Rehabilitation through Portable and Electronic Memory Aids at Different Stages of Alzheimer's Disease, *Les Cahiers De La Fondation Médéric Alzheimer*, 3, 97-107, 2007.
- [13] K. Yasuda, N. Kuwahara, M. Nakamura, K. Morimoto, K. Nakayama, C. Oshima, and J. Aoe: Assistance Dogs for Individuals with Dementia Using ICT Devices: Proposal of Human-Computer-Animal Interface, *Proceedings of International Conference on Humanized Systems 2012*, CD-ROM OS01-1010, 2012.
- [14] I. Dimitrijevi: Animal - assisted therapy: A new trend in the treatment of children and adults, *Psychiatria Danubina*, 21, 236-241, 2009.
- [15] S. L. Filan and R. H. Llewellyn-Jones: Animal-assisted therapy for dementia: a review of the literature, *International Psychogeriatrics*, 18(4), 597-611, 2006.
- [16] J. Perkins, H. Bartlett, C. Travers, and J. Rand: Dog-assisted therapy for older people with dementia: A review, *Australasian Journal on Ageing*, 27(4), 177-182, 2008.
- [17] T. Majić, H. Gutzmann, A. Heinz, U. E. Lang, and M. A. Rapp: Animal-Assisted Therapy and Agitation and Depression in Nursing Home Residents with Dementia: A Matched Case Control Trial, *The American Journal of Geriatric Psychiatry*, 21(11), 1052-1059, 2013.
- [18] C. Oshima, C. Harada, K. Yasuda, K. Machishima, and K. Nakayama: The Effectiveness of Assistance Dogs Mounting ICT Devices: A Case Study of a Healthy Woman and Her Dog, *Lecture Notes in Computer Science (LNCS)*, 8522, Springer, 467-478, 2014.
- [19] K. Yasuda, J. Aoe, and M. Fuketa: Development of an Agent System for Conversing with Individuals with Dementia, *The 27th Annual Conference of the Japanese Society for Artificial Intelligence*, 3C1-10S-1b-2, 2013.
- [20] K. Yasuda, M. Fuketa, and J. Aoe: An anime agent system for reminiscence therapy, *Gerontechnology*, 13(2):118-119, 2014.
- [21] C. Oshima, K. Nakayama, N. Itou, K. Nishimoto, K. Yasuda, N. Hosoi, H. Okumura, E. Horikawa: Towards a System that Relieves Psychological Symptoms of Dementia by Music, *International Journal on Advances in Life Sciences*, 5(3&4), 126-136, 2013.

Orientation Capture of a Walker's Leg Using Inexpensive Inertial Sensors with Optimized Complementary Filter Design

Sebastian Andersson
School of Software Engineering
Tongji University
Shanghai, China

Liu Yan
School of Software Engineering
Tongji University
Shanghai, China

Abstract—Accelerometers and gyroscope are often referred to as inertial sensors. They detect movement and are used for motion tracking systems in many fields. In recent years they have become much smaller, lighter and cheaper which makes them attractive for use in consumer electronics. The goal of this research is to use all these advantages to create a cheap, low cost and accurate motion tracking system. The system that will be developed is using two pairs of accelerometer + gyroscope sensors which communicates with an iOS device using BLE. The sensors are attached to a persons leg to capture the orientation of the leg while walking or running. Studying the movements of a persons leg can be useful regarding both performance and health aspects. To create the system, usage of inertial sensors and how to combine their data using the complementary filter have been studied. Further, several experiments were made to optimize the filter design for this kind of movement. The results shows how the orientation estimation differs in accuracy depending on different values of how the filter is designed. However, by using the right values, a fairly accurate orientation of the leg can be estimated which is proved by the simple visualization of the iOS application.

Keywords—Motion capture, Complementary Filter, Inertial sensors, Bluetooth Low Energy, iOS.

I. INTRODUCTION

Motion tracking systems with inertial sensors have been used for many years in fields that includes military, health care, navigation and flight technologies [1]. But it is just in the last decade that the market of inertial sensors in consumer electronics has rapidly increased [2]. The main reason for this is advances in Micro-Electro-Mechanical-System (MEMS) technology which makes the sensors small, light, low cost and with low power consumption [2,3]. Together with Bluetooth Low Energy (BLE, Bluetooth Smart), which also have reduced power consumption compared to the classic Bluetooth, it makes the sensors very convenient to wear on the body for applications in sports, fitness and health. One product that use the advantages of both these technologies is the CC2541 SensorTag by TI. With the CC2541 SensorTag the development process for smart phone applications that uses inertial sensors gets a lot simpler since no hardware implementation is required. Inertial sensors includes accelerometers and gyroscopes. For a successful motion tracking system, data from both these sensors is combined and thereby creating an

inertial measurement unit (IMU) [4-6].

The improvements of inertial sensor technologies opens up a lot of possibilities for developers to create cheap consumer electronics in ways that was not possible before. It could be anything from tracking the movement of a specific body part to a completely different device or vehicle. One interesting example is to capture the orientation of the legs while walking or running. This can be done for many different reasons. One might want to study the movement of the legs to improve the running technique which can increase the performance and avoid injuries [7]. This paper is considering orientation capturing of the leg while walking or running by using sensors in the CC2541 SensorTag. To make the orientation estimation as accurate as possible, data is combined from both accelerometer and gyroscope sensors by using a complementary filter. Experiments are then made to optimize the accuracy of the filter by having a test subject walking while wearing the sensors.

The rest of the paper is organized in the following way: Section II gives an overview of the application and details about the setup of the experiments, Section III introduces usage of inertial sensors while Section IV introduces the complementary filter, the procedure of how to optimize the filter for this application is described in Section V and finally, the result and conclusion is presented in Sections VI and VII respectively.

II. EXPERIMENTAL ENVIRONMENT

Two CC2541 SensorTags and an iOS device is used for implementation of the orientation tracking system. Each SensorTag represents one IMU and they are attached to the lower leg and the upper leg respectively to get an orientation estimation of the whole leg. Only one angle is being tracked as it is the most interesting while walking or running, this is the angle of the leg straight forward or backward of the walker. In all tests the test subject that is wearing the sensors starts from a standing position with the legs 90° relative to the ground. The subject then starts walking in a speed of 4km/h for around 1 minute and then ends the test by stop walking and goes back to the starting position.

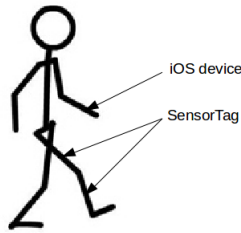


Fig. 1: Placement of the sensors

A. TI CC2541 SensorTag

The CC2541 SensorTag is a device which includes several commonly used sensors, including accelerometer and gyroscope, on one single board and uses Bluetooth Low Energy for communication. It is specially targeted for smart phone application development since a lot of configurations can be done over the GATT¹ server which is already included in the SensorTag's firmware.

B. iOS application

The responsibilities of the iOS device is to connect, configure, read the sensor values from the SensorTags and, most important, make all calculations of the orientation estimation. Connecting, configuring and reading the SensorTag is done using the following steps:

- 1) Scan and discover the SensorTag
- 2) Establish a connection
- 3) Discover the available services and characteristics
- 4) Write to characteristic value
- 5) Read from characteristic value

In iOS this is done by using the *CBCentralManagerDelegate* and *CBPeripheralDelegate* protocols. These protocols makes step 1-3 very straight forward. Step 4 activates and configures the wanted sensors to suit with the application needs. When that is done the SensorTag will advertise at the specified intervals with data from the active sensors. Calculations of the angle estimation is done after each time a new value is updated from a characteristic.

III. INERTIAL SENSORS

A. Gyroscope

Gyroscope sensors measures the velocity of angular movement around one axis. The angle of a rotating gyroscope can be obtained by integrating this data. Of course, since there will be a time interval between each reading from the sensor the integration will not be 100% accurate and as time goes the angle estimation will be less and less accurate. This problem is referred to as drift which is a big problem with inertial sensors [8]. To measure the angular movement in a 3D space it is required to use three different gyroscopes placed orthogonal to each other. In this experiment only one gyroscope is used since it is only the angle in one direction that is of interest.

¹Framework for transporting data between two Bluetooth Low Energy devices

B. Accelerometer

An accelerometer measures the acceleration in G forces. Similar to the gyroscope, three accelerometers placed orthogonal to each other is needed to measure acceleration in a 3D space. In a constant speed or in a resting state the only output will be gravity, which is 1G straight down towards the earth. The gravity can be used to calculate the angle of the accelerometer using trigonometry. In Objective-C the function *atan2f* can be used which is the arctangent function with two arguments.

$$zAngle = atan2f(x, y) * 180.0/M_PI;$$

The above code calculates the angle around the accelerometer *z* axis. *X* and *y* is the acceleration measured on the respective axis. The resulted angle from the *atan2f* function is in radians so it needs to be converted to degrees. Unlike the gyroscopes drift problem when it comes to angle estimation, the accelerometer angle calculation is very accurate as long as it is not exposed to any kind of acceleration other than gravity. In this experiment a 2D accelerometer is used to measure the one angle that is interesting.

Table I shows the update frequency as well as the range of which each sensor is operable. There are some limitations

TABLE I: Sensor details

	Updates	Range
Accelerometer	100ms	±8G
Gyroscope	100ms	±1000°/s

in the range of the gyroscope. For the usage within this paper (walking, slow running) the range is enough but faster movements will need a gyroscope with higher range. Another limitation is, of course, the updates. It is obvious that faster updates will give better results. The SensorTag doesn't allow a lower update frequency to be set over the GATT server and the reason for this is to keep the power consumption at an acceptable level.

IV. COMPLEMENTARY FILTER

Basically, there are two different kinds of filters that have become very popular to use when combining accelerometer and gyroscope data for angle calculation. The more complex one is the Kalman filter. It was first introduced in 1960 by R.E. Kalman [9]. It uses a set of complex mathematical equations to estimate the past, present and future state of a process in a way that minimizes the errors [10]. The other one is the complementary filter. It is much simpler to understand and contains a lot less computations and is therefore much easier to modify and optimize for a specific problem. In its most basic form it takes integrated data from the gyroscope and combines it with data from the accelerometer [11,12] in the following way:

$$angle = (1 - ii) * (angle + gyroData * dt) + ii * accData \quad (1)$$

Where *gyroData* is the angular movement in °/s, *dt* is the time passed since the last reading and *accData* is the angle calculated by the accelerometer. The variable *ii* is a value

between 0 and 1 and will directly decide the influence of the different sensors. Bigger ii will result in an angle where the accelerometer has more influence. How to choose the best possible ii depends a lot of the kind of movement that the IMU will be used for. As mentioned before, the gyroscope is good for reading fast movements during a short time but will drift over time and the accelerometer is very reliable when the speed is constant. How can these principles be applied to optimize the filter for a walking movement?

V. OPTIMIZING THE FILTER

The basic idea for the optimization is simple: use data from the two different sensors when it is the most reliable. This means using gyroscope data when the movement is big and many disturbing forces is acting on the accelerometer and using accelerometer data as much as possible when the sensor is in a resting state to correct the drift caused by the gyroscope. To do this we must know when and how much (varying values of ii) to use the data according to when it is the most reliable. Figure 2 shows the angle estimation of both sensors alone, without combining any data and with no filters. The angle from the gyroscope is very smooth and behaves in a logical way for a walking motion. But one can see that already after 10 seconds the angle has drifted away and after 48 seconds the angle has drifted around 60° . The angle from the accelerometer on the other hand is very noisy during a lot of movement but when standing still again after 48 seconds the angle is back to where it started, no drift at all.

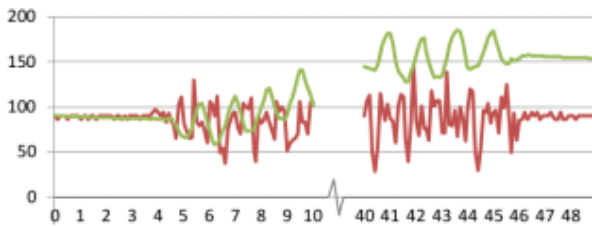


Fig. 2: Angle estimation from gyroscope (green) and accelerometer (red) (y axis: angle, x axis: seconds)

First, the accelerometer data alone can be filtered to get rid of some noise during movement. We know now that angle estimation by the accelerometer is almost 100% accurate when gravity is the only force that is acting on it. This means that in a position where one axis is directed straight to earth that axis will measure an acceleration of 1G and the other two axes 0G which makes the total acceleration 1G. If the device is tilted 45° , two of the axes will measure an acceleration of 0.75G and the third 0G which makes the total 1.5G. So, when gravity is the only force the total acceleration must measure between 1G and 1.5G and that is when the accelerometer angle should have the most influence. This can be programmed as shown below.

```
totalAcc = fabsf(x) + fabsf(y) + fabsf(z);
if totalAcc ≥ 1.0 && totalAcc ≤ 1.5 then
    {Calculations of the desired angles}
    {according to the algorithm}
    {presented in section IV}
end if
```

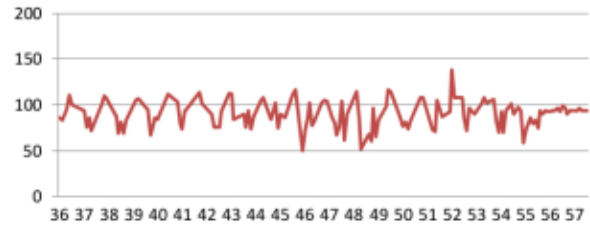


Fig. 3: Improved angle estimation of accelerometer (y axis: angle, x axis: seconds)

Figure 3 shows the result of this method. The improvements from figure 2 where no filtering of the accelerometer data was done is very clear. The spikes are almost gone, the data is less noisy and the angle estimation behaves more logical (more like the gyroscope in figure 2). After knowing how to use the accelerometer data the next step is to find out how much of the data to use. If the accelerometer angle estimation have too much influence there will be too much disturbance in the filtered angle. If it has too little influence the angle will start to drift away with the gyroscope angle estimation. Experiments were made with different values of ii (0.5, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35) and the result is presented in the next section.

VI. RESULT

The figures 4 to 10 presents the resulted angles of the complementary filter discussed in previous sections using different values of ii and with the accelerometer filter. The data is extracted from the same test session, starting from the 22:nd second.

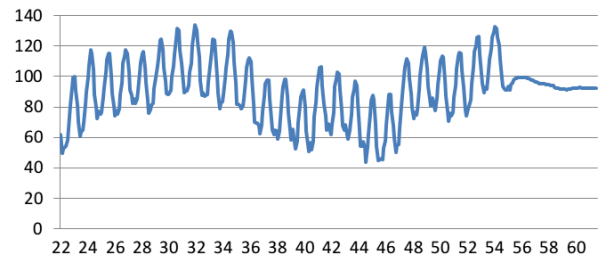


Fig. 4: $ii=0.05$ (y axis: angle, x axis: seconds)

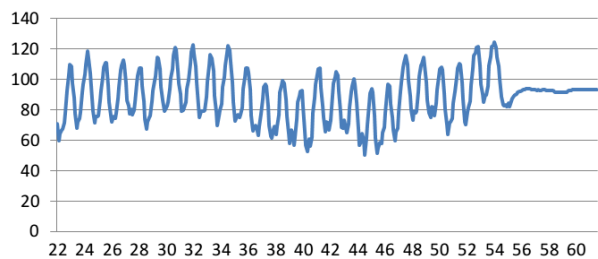


Fig. 5: $ii=0.1$ (y axis: angle, x axis: seconds)

If ii is too small like in figure 4 where it is 0.05 we can see that the angle is drifting away with the gyroscope. On the

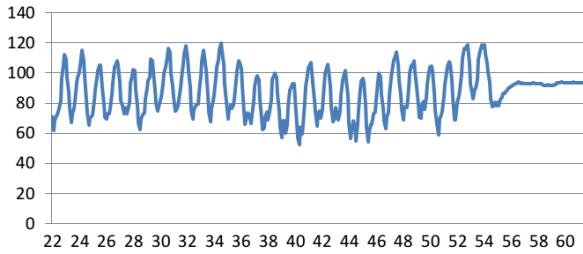


Fig. 6: $ii=0.15$ (y axis: angle, x axis: seconds)

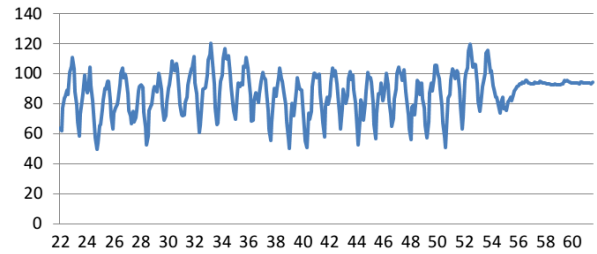


Fig. 10: $ii=0.35$ (y axis: angle, x axis: seconds)

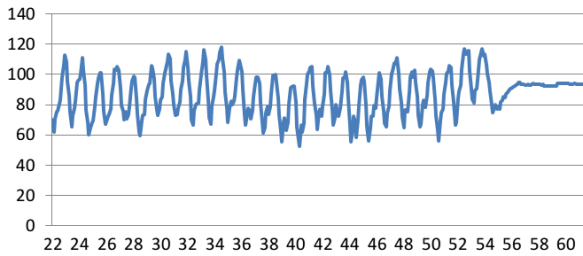


Fig. 7: $ii=0.2$ (y axis: angle, x axis: seconds)

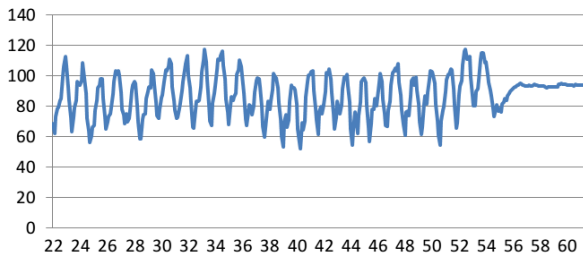


Fig. 8: $ii=0.25$ (y axis: angle, x axis: seconds)

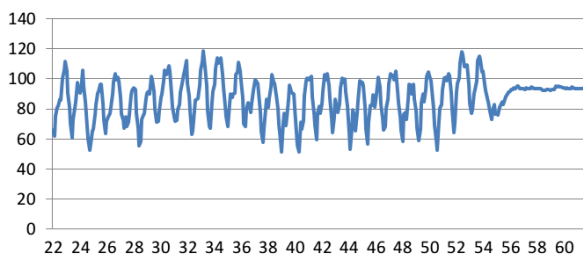


Fig. 9: $ii=0.3$ (y axis: angle, x axis: seconds)

other hand, we can also see the effects of ii being too big as in figure 10 where it is 0.35. The angle doesn't drift at all but the angle estimation when movement is occurring is way too noisy because of the accelerometers big influence. When $ii = 0.1$ there is still some drift on the angle but at $ii \geq 0.15$ there is not much difference when considering the drift problem. From $ii = 0.2$ and up we mostly just gain noise and making the angle estimation less and less smooth.

The iOS application includes a simple visualization of a leg which is wearing the sensors. In figure 11, three samples

has been taken of the visualization, starting when the foot is about to leave the ground and ends when it is making contact again. The data is the same as in the previous figures of the complementary filter output and the filter is using $ii = 0.15$.

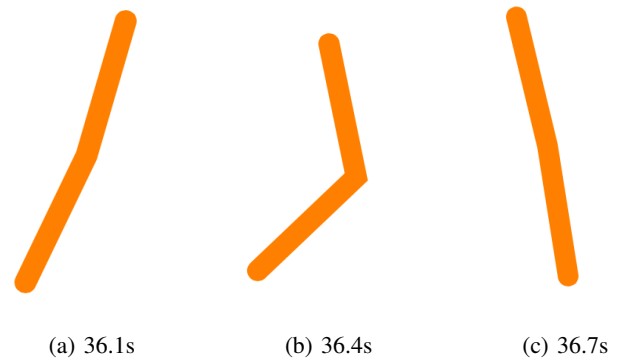


Fig. 11: (a) The leg is behind the body and the foot is just about to get released from the ground. (b) The leg is in the middle of the forward swing. No contact with the ground. (c) The forward swing is complete and the foot is just about to make contact with the ground again.

VII. CONCLUSION

A system with a pair of IMU's that's connected with an iOS device have been developed for angle estimation of a persons leg while walking. Optimizations of the sensor data was applied considering the specific application. The complementary filter has been studied and with help of several experiments we can see how the performance of the filter changes by adjusting how much of each sensor's data is used. The result shows how different values of ii in the filter affects the outcome of the angle estimation. For the best performance we end up with ii around 0.15-0.2, depending a little on what factor is most important: no drifting or making the angle smoother while moving. After 36 seconds of walking, the system is still capturing the angles of the leg which is shown by the simple leg visualization in figure 11.

The filter can be improved even further by analyzing more accurately when each sensor is reliable and from there we can make ii change with the reliability. For example, if at one point the accelerometer data is 80% reliable then 80% of that data should be used but at another point it is calculated to be 5% reliable then only that much is used.

REFERENCES

- [1] Oliver J. Woodman, *An Introduction to Inertial Navigation*. Technical Report, University of Cambridge, 2007
- [2] K. Maenaka, *MEMS inertial sensors and their applications*. 5th International Conference on Networked Sensing Systems - INSS, 2008, 71-73, doi: 10.1109/INSS.2008.4610859
- [3] N. Yazdi, F Ayazi and K. Najafi, *Micro-machined Inertial Sensors*. Proc. of the IEEE, 1998, vol. 86, 1640-1659.
- [4] D. Hazry, M. Sofian and A. Zul Azfar, *Study of Inertial Measurement Unit Sensor*. Proc. of the International Conference on Man-Machine Systems, 2009, doi: <http://hdl.handle.net/123456789/7317>
- [5] P. Corke, J. Lobo and J. Dias, *An Introduction to Inertial and Visual Sensing*. The International Journal of Robotics Research, 2007, vol. 26, 519-535.
- [6] C. Fischer, P. T. Sukumar, M. Hazas, *Tutorial: Implementing a Pedestrian Tracker Using Inertial Sensors*. IEEE Pervasive Computing, 2013, vol. 12, 17-27. doi: <http://doi.ieeecomputersociety.org/10.1109/MPRV.2012.16>
- [7] B. Martin, *Running Technique*. Smashwords, 2011, (E-book)
- [8] G. Welch and E. Foxlin, *Motion tracking: no silver bullet, but a respectable arsenal*. Computer Graphics and Applications, IEEE, 2002, vol. 22, 24-38.
- [9] R. E. Kalman, *A New Approach to Linear Filtering and Prediction Problems*. Trans. of the ASME-Journal of Basic Engineering, 1960, vol. 82, 35-45.
- [10] G. Welch and G. Bishop, *An Introduction to the Kalman Filter*. Department of Computer Science, University of North Carolina at Chapel Hill, 2006, www.cs.unc.edu/welch/media/pdf/kalman_intro.pdf (Accessed 2014-11-03).
- [11] Hyung Gi Min and Eun Tae Jeung, *Complementary Filter Design for Angle Estimation using MEMS Accelerometer and Gyroscope*. Department of Control and Instrumentation, Changwon National University, http://www.academia.edu/6261055/Complementary_Filter_Design_for_Angle_Estimation_using_MEMS_Accelerometer_and_Gyroscope (Accessed 2015-01-19).
- [12] W. T. Higgins, *A Comparison of Complementary and Kalman Filtering*. IEEE Trans. on Aerospace and Electronic Systems, 1975, vol. 11, 321-325.

Technical Perspectives on Knowledge Management in Bioinformatics Workflow Systems

Walaa N. Ismail
Information Systems Dept.
Faculty of Computers and Information
King Saud University.
Saudi Arabia

M.Sabih Aksoy
Information Systems Dept.
Faculty of Computers and Information
King Saud University.
Saudi Arabia

Abstract—Workflow systems by its nature can help bioinformaticians to plan for their experiments, store, capture and analysis of the runtime generated data. On the other hand, the life science research usually produces new knowledge at an increasing speed; Knowledge such as papers, databases and other systems knowledge that a researcher needs to deal with is actually a complex task that needs much of efforts and time. Thus the management of knowledge is therefore an important issue for life scientists. Approaches has been developed to organize biological knowledge sources and to record provenance knowledge of an experiment into a readily resource are presently being carried out. This article focuses on the knowledge management of in silico experimentation in bioinformatics workflow systems.

Index Terms—Bioinformatics Workflows; Knowledge Management; Ontologies; Scientific Workflows; Semantic Web.

I. INTRODUCTION

Scientific experimentation in science domains contains all required aspects of the experimentation process including data analysis, modeling, and testing [1].

A workflow is a well-defined organization of activities or patterns designed to achieve a certain data transformation [2]. Workflow systems by its nature can help bioinformaticians to plan for their experiments, store, capture and analyses of the runtime generated data [2]. Complex workflow systems that integrate programs, methods, agents, and services coming from diverse organizations or sites requires a more flexible framework that can execute such complex scenario [3]. In such a way, the execution sequence and the scheduling of algorithms, data, services, and other software components are orchestrated in a single virtual framework [3].

Workflow systems construction process include but not limited to the following steps [2]:

- (a) Users define typical execution patterns for computational process.
- (b) The system store the generated pattern form step 1.
- (c) Later the users can retrieve such pattern for modifications and re-execute them in different scenarios.

Figure 1 shows a simple workflow example for constructing phylogenetic tree to a given sequence. The steps required to construct phylogenetic tree [4] which begin with the user

input sequence query is:



Figure 1: A Simple Workflow For Constructing Phylogenetic Tree

- Step 1 Choose an appropriate markers for the phylogenetic analysis from the workflow database.
- Step 2 Perform multiple sequence alignments for the matched sequences in step1.
- Step 3 Select an evolutionary model.
- Step 4 Reconstruct the Phylogenetic tree.
- Step 5 Evaluate the phylogenetic tree.

Consider for example **Step 2**; a wide range of algorithms can perform the alignment like

(a) synchronous Blast services, and (b) Blast services. Hence, the workflow enable the user simply specifies that a *Sequence Alignment* is desired. On the second hand expert users can choose to specify all the analysis required for every step in the workflow.

Our level of understanding will be increased to more effectively solve problems and make required decision via knowledge management (KM) discipline [5]. KM is subject that provides strategy, process and technology to share information and expertise among users.

KM has been an important subject disciplines for many fields today which needs understanding knowledge processes and selecting the most appropriate KM systems that can help in creating, storing, and more effectively sharing knowledge [5].

A bioinformatics workflow system seems by its nature could benefit from KM principles and methodology [6] [7]. The main reasons are bioinformatics workflow systems usually interact with [8]:

- 1) A modern infrastructure that are frequently change.
- 2) General community knowledge.
- 3) The biologists who generally prefer to share their knowledge with each others.
- 4) Also those workflows usually acquire fast accessible knowledge sources.

Therefore, in the bioinformatics, KM can be defined as a systematic process that allow creating, capturing, sharing, and analyzing knowledge in ways that affect system performance and availability [9].

With the vast amount of the available bioinformatics tools, services and algorithms that can execute the biologists tasks; it's a must to have certain technology that allow automation and discovery of such resources, in addition to that the bioinformaticians need to create complex workflows from a wide range of available web services knowledge base. So far the emergence of the semantic Web technology (SW) [10] is starting to have a significant impact on knowledge integration, querying, and knowledge sharing in the life science domain [10], [11].

The success of knowledge management system (KMS) in Bioinformatics can be achieved by the assistance of knowledge technology. Knowledge technology is a part of KM, refers to an unclear set of available tools that enable better representation, organization and exchange of information and knowledge [8] [1]. Among the existence technologies are knowledge mapping, collaborative technologies, semantic technologies and social computing tools [12].

The growing acceptance of the semantic web as a means to manage biological knowledge is noteworthy [1] as SW technology offers more flexibility in data modeling by integration of large amounts of data [11]. Therefore, This paper will discuss the technical perspectives on KMS in Bioinformatics that focus on technology, ideally those that enhance knowledge sharing and growth in bioinformatics domain.

The rest of the paper is organized as follows: Section II, presents how semantic Web technology is an effective knowledge management technology in life science domain. Section III, discussed knowledge management efforts in Bioinformatics workflow systems and presents the knowledge management life cycle in bioinformatics workflow systems. Section IV explain semantic system biology cycle. Section V presents related work about workflow and workflow systems in life science. Finally, section VI concludes and outlines directions for future work.

II. TOWARDS EFFECTIVE KNOWLEDGE MANAGEMENT IN THE LIFE SCIENCES

Semantic Web (SW) technology is an effective knowledge management technology in life science, since it allow automatic discovery and execution of web services that can handle the workflow tasks, which prevents biologists from

the need to working with similar or time-consuming tasks, such as taking manual copy of one tool and then pasting that tool to another tool [10].

SW depends on a set of web technologies specifically designed to facilitate automated machine interoperability [10]. It promises to meet the challenge of integrating and querying highly diverse and distributed resources [13].

Systems based on SW would provide sophisticated frameworks to manage and retrieve knowledge. Ontologies in biology (bio-ontologies) and the semantic Web are playing a vital role in the integration of data and knowledge by offering an explicit, unambiguous and rich data and knowledge representation mechanisms [14] [10].

Biomedical ontologies are playing an important role in life sciences semantic web since they help in capturing the semantics of entities and their interrelationships within biology domain, thereby reducing conceptual ambiguity, increasing re-usability and computational automation that aids in knowledge gathering and discovery [15].

Ontologies can be classified according to the degree of conceptualization which includes [12]:

- 1) **Upper-level ontology:** Ontologies that describes general concepts which are independent of a particular domain. Their applicability is in providing support to a large number of ontologies. The Basic Formal Ontology¹ is a widely used upper level ontology in a number of sub-domains within the life sciences.
- 2) **Domain ontology:** The knowledge represented in this type of ontology serves a particular domain by providing vocabularies about concepts and their relationships governing the domain such as The Gene Ontology (GO)².
- 3) **Application ontology:** These ontologies are typically used to define concepts for a particular use case. For instance, EFO³ is used to represent concepts and sample variables from gene expression experiments. An ontology that captures knowledge related to the cell cycle processes.

III. WORKFLOWS AT THE KNOWLEDGE LEVEL

Bioinformatics workflow systems could benefit from KM efforts that define strategies to capture the vast amount of available bioinformatics tools, services and algorithms that can execute a certain biologists tasks. Knowledge management (KM) processes encompasses many tasks such as knowledge formulation, storage and distribution [14]. Figure 2 represent knowledge management life cycle in bioinformatics [14]

- 1) The **Creation stage** identify the major bioinformatics system components including rich knowledge base about services/tools , algorithms and data conversion methods.

¹<http://www.ifomis.org/bfo/>

²<http://geneontology.org/>

³<http://bioportal.bioontology.org/ontologies/EFO>

- 2) The **Identify or collect** stage collects the local and shared knowledge, algorithms, other workflows provenance, scientific theories and available scientists experience to create the selected main knowledge domains components.
- 3) The **Select** stage takes the composed collected knowledge and evaluate its value. For organizing and classifying knowledge that will be stored in the knowledge repositories; one framework should be selected..
- 4) The **Store** stage classifies the collected knowledge and adds them to the workflow system.
- 5) The **Share** stage retrieves knowledge from the workflow system and makes it available to the system users. Scientists often needs to share and use ideas, results of experiments, knowledge expertise over the network or from other workflow systems.
- 6) The **Apply** stage reuses the collected knowledge in executing workflow tasks, building new workflow, discovering new research ideas, taking important decisions and learning new thought.
- 7) The **Update** stage provides a creative update and automated knowledge discovery platform by investigating uncovered new knowledge, such as new methods, algorithms, scientists feedback, analysis, research, and experimentation.



Figure 2: Knowledge Management Life Cycle

Workflow systems can be defined as repositories of scientific knowledge [2] [16]; so Does describing workflow systems at the knowledge level could define new concepts? if so, we have to ask what should workflow systems expected to achieve by using that knowledge?

Figure 3 shows a set of layers in the workflows specification process, the layers organized such that from more abstraction level to more specific level. The information contained on each layer can be used to implement the layer below it. Workflows specify what data will be used as well as the services or codes that are to be used to execute each workflow task. Those refers to layers 2 and 1.

The data and services that has been specified in *level 1* workflows are then mapped to actual execution resources in the execution environment, resulting in level 0 workflows.

Moving up in the figure levels, some workflow steps can be ignored if they are not central to the experiment a workflow can then be described not by the specific resource but by identifying classes of services to be used instead.

For example, if a workflow to Construct phylogenetic tree [4] is needed; user query sequence is first processed with normalization step followed by sequence alignment step, followed by selection of an evolutionary model, and then Phylogenetic tree reconstruction without specifying any algorithm or methods to be used.

Workflows at level 3 does not specify how each operation will be executed in relation to other operation in the workflow instead it specify how data will be carried out. At a highest level of workflow abstraction, only the desired results would be specified without any other details. For example, Construct phylogenetic tree to a reference sequence without any details are provided about how to construct the tree or what workflow to be used or what type of data to be generated.

Having scientific workflow means to have a wide range of methods, algorithms, and tools that can perform a given workflow task at different level of granularity; in addition to that, the architecture at *the symbol level* describes the capabilities of workflow systems and how to execute the workflow identified tasks [2].

On the other hand, *the knowledge level* describe the scientific tasks that a workflow system expected to accomplish through suggestion of descriptions and capabilities that would affect what can be done. With more knowledge about workflows usage and integration will improve the workflow behavior by solving more tasks and producing new kinds of results [2], [17].

Figure 3, also relate workflow abstraction layers to the knowledge level and the symbol level. In summary having systems that can take workflows requests from users, and then execute the workflow without any details about execution details or resources would decrease the inexperienced user overload who have small amount of knowledge about the workflows tasks selection and execution.

IV. REASONING WITH WORKFLOWS AT THE KNOWLEDGE LEVEL

To receive the accurate knowledge that can improve any system performance requires a system that can determines the user purposes, and then tracks the user's actions and behavior. [18]

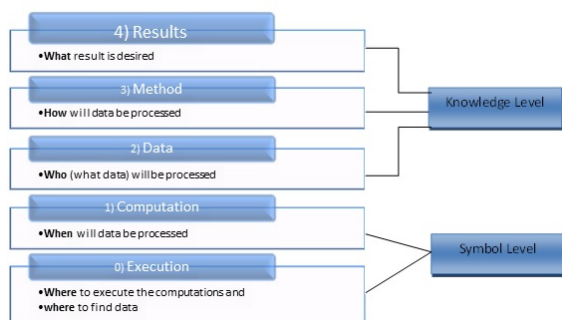


Figure 3: Workflow Abstraction Layers

Semantic system biology (SSB) provides a semantic description of the knowledge about the biological systems on the whole facilitating data integration, knowledge management, reasoning and querying [7]. Figure 4 describe

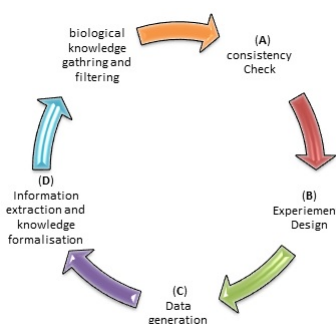


Figure 4: Semantic System Biology Cycle

the semantic systems biology cycle [10]. The cycle begins with gathering and integrating biological knowledge into a semantic knowledge base; then data are checked for consistency, then; (B) This yields criteria about particular functions of biological components that may be used to design experiments; (C) The experiments generates new data and might also verify the design criteria. (D) The new data are then integrated into the knowledge base, thereby enhancing the quality of the knowledge base and allowing a new cycle of hypothesis building and experimentation.

With knowledge of what workflow components do, and experiment design; workflow systems can assist scientists by using those knowledge to make automatically decisions concerned about specific domain [19].

Figure 5 is a schematic representation of the workflow in SSB [13]. Firstly, biological knowledge is extracted from disparate resources and integrated into a knowledge base.

Given the user query and knowledge base about the application domain and input data: the reasoner identifies the strategies to the user to use and run the tools that can execute

his/her request. Executor then run the completed workflow and updates the knowledge base with the results of workflow execution that can be used to make new inference.

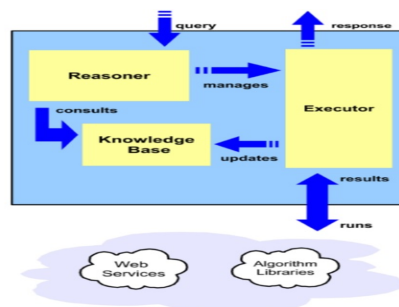


Figure 5: Reasoning in Workflow system at knowledge Level

V. WORKFLOWS AND WORKFLOW SYSTEMS

There have been a large number of workflow design and execution engines that supports *In silico* biological study. Workflow Systems have been used as a useful paradigm to model and manage complex scientific analyses [2] [3]. Some scientific workflows represent how to use and compose a variety of remote services to accomplish an overall task. Other workflows systems have incorporated semantic component to support knowledge representation of workflows execution.

Galaxy in [15] is a workflow web-based platform for analysing genomic sequences. In Galaxy several tools can be merged, ranging from simply manipulating data to complex analysis tasks. Galaxy provide an flexible construction of workflows as it can:

- Combine knowledge of current workflow tasks.
- Can be executed from a single Web interface.
- Share the output of the tool by sending the current results to other tools as input.
- Store the history of all performed actions which facilitate the analyses of any task at any time.
- Galaxy can use users history to build workflow.

- The workflows can be re-used in other systems, like other servers or myExperiment [15].
- The capturing of data provenance and the context of a workflow are automatically tracked and managed.

Wings [20] a workflow system that allow users to describe their desired analyses tasks. After the users describe their goal Wings begin automatically to validate the input goal and data by using a knowledge base (using ontologies and rules) about workflow components and finally map each task to services that Pegasus [16] use to execute that task.

Wings organize all workflow components in hierarchies; components such as workflow tasks, data, properties, and constraints regarding their proper use. In addition Wings allow users to describe a workflow templates that can be reused for different scenarios, and it also can automatically build workflows using data products descriptions of what the user prefer.

In [21], the author presented Sesame a semantic bioinformatics workflow design system with new ontology for bioinformatics tools/services. Sesame allows the biologists to perform their analyses using terms that they are familiar with. After designing the semantic workflow, Sesame have a knowledge repository that associating each analyses entity with the instances of bioinformatics tools/services and data that previously had been used to handle such data and tasks. Sesame free the biologists from the necessity of learning the details of the computational aspects of the bioinformatics tools. Also, Sesame can perform simple instantiation cases and for each analyses entity Sesame ask the user to select one instance of bioinformatics tools/services. Then, the user specifies the parameters and input data for the selected tool.

Taverna [22] is a workflow building platform that facilitate the matching process of users requests with the available workflows templates and services via using of rich knowledge descriptions of workflows components that enable users to specify either the type of service they wish to use or a graph of workflow services and their dataflow. On the other hand Taverna is designed as a do-it-all platform, which can be very complex to be used for biologists with limited computing background.

The authors in [13] have utilized several semantic technologies to identify the scientists intent, and then to facilitate the control of workflow execution and enrichment of workflow provenance of new tasks,

The Magallanes [23] is a library that can develop effective workflow discovery engines that can help to collect web-services which will be used to execute workflow tasks and it's datatypes. The discovery of Web services can be based on *syntax description of services or objects* that is it's name which is often unsatisfactory in bioinformatics because it presumes knowledge of objects names or semantics services discovery by specifying a general descriptions about services

or objects which allow to have a more accurate discovery mechanism. Magallanes uses a syntactic text-based approach and a semantic approach to collect different services that can handle the input and output data types.

In [24] a framework for services selection in the life-sciences is proposed. The solution build workflows by data-type matching methods that provide less time and effort through selection of best services that can handle workflow tasks so that a small set of the available services that can achieve user task are identified.

Kepler [25] is a graphical system for scientific workflows design, execute, reuse, and sharing. Kepler's provide high effective workflow designing process by monitoring data and provenance information during the initial workflow design stage Kepler supports also many advanced features such as automatic workflow validation and editing; by providing a semantic annotation of workflow tasks from a domain ontology. Also Kepler's workflows are created by connecting a chain of workflow components together called *Actors* each Actor has several ports through which input and output ports containing data and data references are sent and received. Each workflow has a Director that determines the model of computation used by the workflow,

The knowledge level of any intelligent workflow system is concerned with the kind of knowledge it can use, and how it response to users requests, or what is the user's goals. [1]

The initiatives proposed comparisons given in Table I demonstrate what is the advantages of the semantic web technologies workflow design systems [10], including;

- **Automatic workflow generation.** During the building of workflows the system can automatically build the workflow without the need to any other tools as it has it's own knowledge about the workflow components, and data.
- **Workflow validation ;** the knowledge of components and data that the system have about the different operations enables the workflow system to validate any workflow task even in a complex composition scenarios.
- **Automatic metadata generation;** the descriptions of new data products that are generated during workflow execution are automatically generated as we have knowledge-level descriptions on each datasets.
- **Guarantee of trusted provenance;** Learning from previous designs that perform a similar task when designing a new workflow is more economic and efficient. That is; if the provenance of new data products obtained through a highly efficient presumed workflow systems it can be a good indicator of high quality process used to obtain those new results.

The system can include knowledge base that integrate its components with *Semantic about provenance* that comprises the experiment with all the other metadata about experiments which help the scientist to learn how to use and compose

Table I: Workflow Systems at Knowledge Level

Feature examined	Galaxy	Wings	Taverna	Sesame	Kepler	Magallanes
Can workflow generated automatically?	Yes	Yes	Yes	Yes	Yes	Yes
Does the user queries is validated?	Yes	Yes	Yes	Yes	Yes	Yes
Can workflows be shared with other users?	Yes	Yes	Yes	Yes	Yes	No
Can results be shared with other users?	Yes	Yes	Yes	Yes	Yes	Yes.
Data provenance available?	Yes	Yes	No	Yes	Yes	No
Can a user add a web service to the tool?	No	Yes	Yes	Yes	No	Yes
The description of new data products?	Limited	Yes	Yes	Yes	Yes	Yes
Contain widely accepted workflows?	No	Yes	Yes	Yes	Yes	No
Automation of workflow execution.	Yes	Yes	Yes	Yes	Yes	Yes

services in a another workflow systems. Also mining prove-nance data of repeatedly executed workflow tasks could help to identify the performance and quality information about those services that can execute a similar tasks or accepting the same inputs data type. This information can assist the scientist to choose between vast amount of alternative services. [7] [14]

VI. CONCLUSIONS AND FUTURE WORK

Knowledge management is a broadly defined concept vary-ing from one domain to the other. For instance, knowledge management in the business domain [26] would mainly deal with management of business activities such as business poli-cies, assets and risk assessments. In comparison, knowledge management in bioinformatics [19] deals with management of what is understood about the various components of a system of interest. Also knowledge representation plays a crucial role in the facilitation of processing and sharing knowledge between people and application systems.

Additionally, knowledge representation languages should adopt a common syntax that is reusable and enables parsing of data in a semantically unambiguous manner [5]. Ontologies in biology (bio-ontologies) and the semantic Web are playing vital role in the integration of data and knowledge by offering an explicit, unambiguous and rich representation mechanism. This increased influence led to the proposal of the seman-tic systems biology paradigm to complement the techniques currently used in systems biology. semantic systems biology provides a semantic description of the knowledge about the biological systems on the whole facilitating data integra-tion, knowledge management, reasoning and querying. These conditions in scientific workflow environment will support intelligent inferencing of facts over a given biological domain and also facilitate processing of information even in complex scenarios that require composition of different sources, or algorithms to be handled.

For future work workflow system could benefit from iden-tifying syntactic patterns [27] which are sets of axioms in an OWL ontology with a regular structure. Detecting these patterns and reporting them in human readable form should help the inexperienced workflow users to understand the style of ontology and is therefore useful in expressing the bioin-formatics experiments knowledge more preciously. However, the detection of such patterns is sensitive to variations in the assertions [27].

Also its a must to differentiate between axioms that are semantically equivalent but syntactically different as in this case it can lead to reducing the effectiveness of the knowledge

presented in any workflow system [18]. So workflow methods could focuses on Semantic regularity analysis that focuses on the knowledge encoded in the ontology, rather than how it is spelled.

REFERENCES

- [1] P. Zhang, L. Zhang, Z. Fan, and X. Qiu, "Knowledge based model-ing method of artificial society oriented to emergency management." Springer Berlin Heidelberg, 2014, vol. 461, pp. 278–287.
- [2] P. Walsh, J. Carroll, and R. D. Sleator, "Accelerating in silico research with workflows: A lesson in simplicity," *Computers in Biology and Medicine*.
- [3] S. Y. Bendoukha, H. and A. Benyettou, "A novel framework for defining and submitting workflows to service-oriented systems." *Journal of Information Processing Systems.*, vol. 10, pp. 365–383, 2014.
- [4] N. R. BP, "Basics for the construction of phylogenetic trees," *Webmed Central BIOLOGY*, p. 12, 2011.
- [5] J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, and G. Dedene, "Formal concept analysis in knowledge processing: A survey on applications." *Expert Systems with Applications*, vol. 40, no. 16, pp. 6538 – 6560, 2013.
- [6] R. Abdullah, H. Ibrahim, R. Atan, S. Napis, M. H. Selamat, N. Haslina, and S. Hernazura, "The development of bioinformatics knowledge man-agement system with collaborative environment," *IJCSNS International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 309–319, 2008.
- [7] E. Antezana, M. Kuiper, and V. Mironov, "Biological knowledge man-agement: the emerging role of the semantic web technologies," *Briefings in Bioinformatics*, 2009.
- [8] A. Fiannaca, M. La Rosa, A. Urso, R. Rizzo, and S. Gaglio, "A knowledge-based decision support system in bioinformatics: an applica-tion to protein complex extraction." *BMC bioinformatics*, vol. 14, 2013.
- [9] R. Maruta, "The creation and management of organizational knowledge," *Knowledge-Based Systems*, vol. 67, pp. 26 –34, 2014.
- [10] H. Chen, T. Yu, and J. Y. Chen, "Semantic web meets integrative biology: a survey," *Briefings in Bioinformatics*, vol. 14, no. 1, pp. 109–125, 2013.
- [11] K. Sutherland, K. McLeod, G. Ferguson, and A. Burger, "Knowledge driven enhancements for task composition in bioinformatics." *BMC bioinformatics*, vol. 10, 2009.
- [12] A. Splendiani, M. Donato, and S. Drghici, "Ontologies for bioinformat-ics," K. Nikola, Ed. Springer Berlin Heidelberg, 2014, pp. 441–461.
- [13] E. Pignotti, P. Edwards, A. Preece, N. Gotts, and G. Polhill, "Enhancing workflow with a semantic description of scientific intent," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, Eds. Springer Berlin Heidelberg, 2008, vol. 5021, pp. 644–658.
- [14] Y. Gil, "From data to knowledge to discoveries: Artificial intelligence and scientific workflows," vol. 17, pp. 231– 246, 2009.
- [15] M. Aranguren, J. Fernandez Breis, C. Mungall, E. Antezana, A. Gonzalez, and M. Wilkinson, "Oppl galaxy, a galaxy tool for enhancing ontology exploitation as part of bioinformatics workflows," *Journal of Biomedical Semantics*, vol. 4, 2013.
- [16] E. Deelman, G. Mehta, G. Singh, M.-H. Su, and K. Vahi, "Pegasus: Mapping large-scale workflows to distributed resources," in *Workflows for e-Science*, I. Taylor, E. Deelman, D. Gannon, and M. Shields, Eds. Springer London, 2007, pp. 376–394.
- [17] A. Belloum, R. Cushing, S. Koulouzis, V. Korkhov, D. Vasunin, V. Guevara-Masis, Z. Zhao, and M. Bubak, "Support for cooperative experiments in e-science: From scientific workflows to knowledge sharing," in *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, ser. Focus on Structural Biology, I. Roterman-Konieczna, Ed. Springer Netherlands, 2013, vol. 8, pp. 135–159.
- [18] V. Chaudhri, P. Clark, A. Overholzer, and A. Spaulding, "Question generation from a knowledge base," in *Knowledge Engineering and Knowledge Management*, ser. Lecture Notes in Computer Science, K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvnen, Eds. Springer International Publishing, 2014, vol. 8876.
- [19] J. Girish, G. Arun, and R. Gintaras, "A novel knowledge management system based on workflows," ser. 24th European Symposium on Com-puter Aided Process Engineering. Elsevier, 2014, vol. 33, pp. 853–858.

- [20] Y. Gil, V. Ratnakar, J. Kim, J. Moody, E. Deelman, P. Gonzalez-Calero, and P. Groth, "Wings: Intelligent workflow-based design of computational experiments," *Intelligent Systems, IEEE*, vol. 26, no. 1, pp. 62–72, 2011.
- [21] L. Zhang, Y. Wang, P. Xuan, A. Duvall, J. Lowe, Y. Wang, A. Subramanian, P. Srimani, F. Luo, and Y. Duan, "Sesame: A new bioinformatics semantic workflow design system," in *Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on*, 2013, pp. 504–508.
- [22] C. Goble, K. Wolstencroft, A. Goderis, D. Hull, J. Zhao, P. Alper, P. Lord, C. Wroe, K. Belhajjame, D. Turi, R. Stevens, T. Oinn, and D. De Roure, "Knowledge discovery for biology with taverna," in *Semantic Web*, C. Baker and K.-H. Cheung, Eds. Springer US, 2007, pp. 355–395.
- [23] J. T. O. Rios, Javier; Karlsson, "Magallanes: a web services discovery and automatic workflow composition tool," *BMC Bioinformatics*, vol. 10, p. 334, 2009.
- [24] M. DiBernardo, R. Pottinger, and M. Wilkinson, "Semi-automatic web service composition for the life sciences using the biomoby semantic web framework," *Journal of Biomedical Informatics*, vol. 41, pp. 837–847, 2014.
- [25] D. Barseghian, I. Altintas, M. B. Jones, D. Crawl, N. Potter, J. Gallagher, P. Cornillon, M. Schildhauer, E. T. Borer, E. W. Seabloom, and P. R. Hosseini, "Workflows and extensions to the kepler scientific workflow system to support environmental sensor data access and analysis," *Ecological Informatics*, vol. 5, no. 1, pp. 42 – 50, 2010.
- [26] M. Sonntag, K. G?rlach, D. Karastoyanova, F. Leymann, P. Malets, and D. Schumm, "Views on scientific workflows," in *Perspectives in Business Informatics Research*, ser. Lecture Notes in Business Information Processing, J. Grabis and M. Kirikova, Eds. Springer Berlin Heidelberg, 2011, vol. 90, pp. 321–335.
- [27] E. Mikroyannidi, M. Quesada-Martnez, D. Tsarkov, J. Fernndez Breis, R. Stevens, and I. Palmisano, "A quality assurance workflow for ontologies based on semantic regularities," in *Knowledge Engineering and Knowledge Management*, ser. Lecture Notes in Computer Science, K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvnen, Eds. Springer International Publishing, 2014, vol. 8876, pp. 288–303.

Public Transportation Management System based on GPS/WiFi and Open Street Maps

Saed Tarapiah
Telecommunication Engineering Dept.
An-Najah National University
Nablus, Palestine

Shadi Atalla
Lavoro Autonomo (LA)
Torino
Italy

Abstract—Information technology (IT) has transformed many industries, from education to health care to government, and is now in the early stages of transforming transportation systems. Transportation faces many issues like high accidents rate in general, and much more rate in developing countries due to the lack of proper infrastructure for roads, is one of the reasons for these crashes. In this project we focus on public transportation vehicles - such as buses, and mini-buses -, where the goal of the project is to design and deploy a smart/intelligent unit attached to public vehicles by using embedded microcontroller and sensors and empowering them to communicate with each other through wireless technologies. The proposed Offline Intelligent Public Transportation Management System will play a major role in reducing risks and high accidents rate, whereas it can increase the traveler satisfactions and convenience. Here, we propose a method, software as well as a framework as enabling technologies to for evaluation, planning and future improvement the public transportation system. Our system even though can be as whole or parts can be applied all over the world we mostly target developing countries. This limitation mostly appear by consider off-shelf technologies such as WiFi, GPS and Open Street Maps (OSM).

Keywords— ITS; GPS; WiFi; Transportation; OSM

I. INTRODUCTION

Intelligent Transportation Systems (ITS) have received much attention in recent years in academia, industry and standardization entities due to their wide impact on people's life as their scope to provide vital applications and services to improve transportation safety and mobility and to optimize the usage of available transportation resources and time. ITS applications and services rely on advanced technologies to be deployed and distributed among the intelligent infrastructure systems and vehicles system. Mainly, these technologies include but not limited to physical world perceive technologies that able to perform real world measurements and convert them into the digital world, processing and storage capabilities that operate on the digital measurements by storing, analyzing them and communication technologies include wired and wireless technologies to exchange the collected data among the vehicles them self and also to their infrastructure and vice versa.

GPS tracking devices stand at the core of the enabling perceive technologies for ITS applications and services. Indeed, the number of vehicles' GPS-enabled On-board Unit has sharply increased due to their vital and beneficial rules for both the vehicles and the drivers. GPS-based services include in-vehicle satellite navigation, vehicle security system, accident notification and tracking along with monitoring to name few.

Although most of GPS-based application depend on real time information collection, the historic GPS collected data intrinsically has great potential for further offline based application such as compute the journey speed, congestion monitoring, accidents deep analysis such as accidents reason and driver's behavior. Preliminary results of our study are reported in [1] This project aims to build an open framework that focuses on traffic and vehicular data for enhancing Public Transportation Management System (PTMS) efficiency in terms of analysis and planning. The proposed framework consists of four main phases namely user data collection, transmission, data analysis and decision making.

The first phase, data collection, uses off-the-shelf hardware components in order to build Smart on-Board Unit (SBU) that is fitted in to the public Transport Vehicles such as buses and mini-buses. SBU endowed with limited processing capabilities, temporary and persistent memory such as EPROM, GPS sensor and WiFi module to transfer the collected information to database storage. GPS tracking devices collect information regarding the vehicle such as the vehicles geographical location (i.e. longitude and latitude), speed and the driving direction at regular intervals of time [1]. The design of this phase relies on integrating multi-sensor capabilities together, in order to increase of the range of possible application that may serve the public transportation system such as collecting the pollution level along the road segments.

In the second phase, transmission, since we are storing the data from the earlier phase onto SBU, here we are interested to transfer the information to a back-end server. We are using a WiFi-based throwbox to the access point located in the main bus station which act as a gateway connected through the internet to the back-end server. The these data is in central database server based on a trip ID that is unique which is consist of triplet the vehicle ID, trip start time and end time.

The third phase, data analyses, the core of this phase is to inject the bus geographical location at a given timestamp along the trip on a digitized map such as open street maps (OSM) each street segment has different attributes such as the street category (pedestrian, highway or motorway) each category has a maximum allowed speed attribute. Here, we record the number of times a vehicle violates these speed limits and the corresponding violation time durations. The resulted information is compared to predefined threshold and limits which allows the system to decide whether the vehicle violated the traffic regulations or not at a given trip. After running our model for long enough periods, we expect that tracking

and ticketing system can be fed by authorized department which can be utilized to get clear view about infrastructure which can be used for developing and planning to improve the infrastructure on some field or apply some regulations which will aim to reduce traffic accidents.

OSM, which provides free geographical information, is sometimes referred to as a map version of Wikipedia. Its data additions, updates, and corrections are made available by its participants.

II. RELATED WORK

This section explores a selected tracking fleet of vehicles solutions appeared on academic and research works so far. This work considers only solutions span a whole system for tracking a group of vehicles. In general tracking systems composed of two parts, the first part is on-Board Unit attached to moving vehicles, whilst the second part is a central application to collect, to process and to visualize useful reports. This section attempts to classify the considered related work based in different criteria. Each criteria requires intrinsic requirements to build the system.

Table I presents and compares selected related works based on the following criteria.

The first criteria (namely Type) differentiates between on-line and offline tracking systems. While online systems require the on-Board Unit to have permanent (available everytime and everywhere) connection with the central application. In the contrary Offline systems aggregate the collected data on a local storage unit and communicates with the central application only when the communication link available such as WiFi. The Offline systems focus on historical data processing and visualizing the generated reports(such as track a vehicle over a digital map). The offline system data transmission can be handled manually such as removing the on-Board Unit from the vehicle and connected it to a PC and transfer the data. Finally, some systems can be considered as online, offline or combined (online and offline) solutions.

The second criteria (namely Smart Unit Type) here we differentiate between different types of hardware used for build on-Board Unit for the tracking system. Three main units hardware are considered, first Commercial unit available in the market, second option could be available smart Phone with GPS and wireless links, the last option is the customized unit(where authors provide design and implementations in the considered work).

The third criteria (namely connection Type) which states the communication channel used by the on-Board Unit to transfer the collected data. Such connections could be any cellular connection or combination of them such as GPRS, SMS and 3G for simplicity we call it cellular connection.

Last criteria (namely Features) this part we make comparison between different tracking systems such as : features:

- 1) Visual Vehicles' Tracking : the ability to project the vehicle trip onto geographic digital map.
- 2) Instantaneous Vehicles Speed : Reporting the vehicle speed along the travelling track.

Data Element	Type	Size[Bytes]
Vehicle ID	String	8
longitude	float	4
latitude	float	4
speed	float	4
direction	byte	1
Timestamp	integer	4

TABLE II: Single GPS Data Record

- 3) Alerting: Send alerts to the driver such as breaking maximum speed on specific road segment while driving.
- 4) Geo-fencing: Identify geometric shapes over the digital map where the driver have to avoid while moving.
- 5) Geo-Casting: Sending Alerting information to specific central office and the vehicles in proximity of the vehicle generating accident or alarm situations.

III. SYSTEM MODEL PHASES AND ENABLING TECHNOLOGIES

A. GPS Data Collection

GPS tracker data used in this work is supplied from SBU fitted to public transportation vehicles. This data consist of one record for each instance a vehicle reported its position. Each record includes Vehicle ID, vehicle type, position coordinates(longitude and latitude), speed, date and time, direction.

Table II illustrates the GPS data record of interest. The Data type is described by C programming language notations. SBU stores the while GPS data record but the Vehicle ID. The total size of each locally stored GPS recoded equals 17 Bytes. With second-to-second data logging the SBU requires $17 \times 60 \times 60 = 61200$ Bytes of local storage to accommodate one hour of GPS recording.

The Vehicle ID is an unique identifier but is anonymous and does not include information about the driver identity.

If the vehicle moves away from the source bus station (the wireless connection with access point will not be available) then SBU will perceive and store the GPS recodes data into a local storage for off-line data logging. If the vehicle arrives close to the destination bus station, (the wireless connection reestablished again) the SBU will send the collected trip data to the gateway storage in the bus-station and will delete it is local copy for the sake on disk space.

SBU is equipped with motion detection sensory device. Which means that SBU can detect anonymously and automatically if it is moving or in stationary state. Through this way the SBU will log GPS data if it is moving and it will go to sleep mode and stop collecting GPS data if it detect that it is in stationary state for the last 5 minutes and when the SBU moves again to will resume collecting the GPS data.

B. GPS Data Transmission

This includes the tools and mechanisms to transfer the perviously collected GPS data to the back-end server for permanent storage in database system. Furthermore, it also considers the intermediates transfer from one place to another till arrive it final back-end server.

Reference(s)	Type	Smart Unit Type	Connection Type	Features				
				Tracking	Speed	Alerting	Geo-fencing	Geo-Casting
[2], [3]	Online	Custom Unit	GPRS	Yes	Yes	Yes	No	No
[4]	Online	Custom Unit	GPRS and SMS	Yes	No	Yes	No	Yes
[5]	Online	Custom Unit	GPRS	Yes	Yes	No	No	No
[6]	Online	Custom Unit	SMS	Yes	Yes	No	No	No
[7]	Combined	Custom Unit	GPRS	Yes	Yes	Yes	No	No
[8]	Combined	Custom Unit	SMS	Yes	No	No	No	No
[9]	Combined	Commercial Unit	GPRS and SMS	Yes	Yes	Yes	Yes	No
[10]	Combined	Smart Phone	GPRS, 3G and SMS	Yes	Yes	Yes	No	No
[11]	Online	Smart Phone	Cellular	Yes	Yes	No	No	No

TABLE I: Features of Others Vehicle Monitoring & Tracking Systems

The transmission technologies vary among many communication option depending the intended application for instance real time tracking application will require an instance and permanent mobile connection to the back-end server or any intermediate stage which always have direct connection to the back-end server. Cellular connections (2G, 3G, and LTE) are the conventional shapes of communication to this purposes. Whilst for non real time application, in particular this work, this ease a lot the communication challenges. Thus, the communication options include the aforementioned ones plus other options like vehicle mobility which is well known as Delay Tolerant Networks (DTN). DTN exploits the vehicle mobility to transfer the GPS collected data from one point to another. [12].

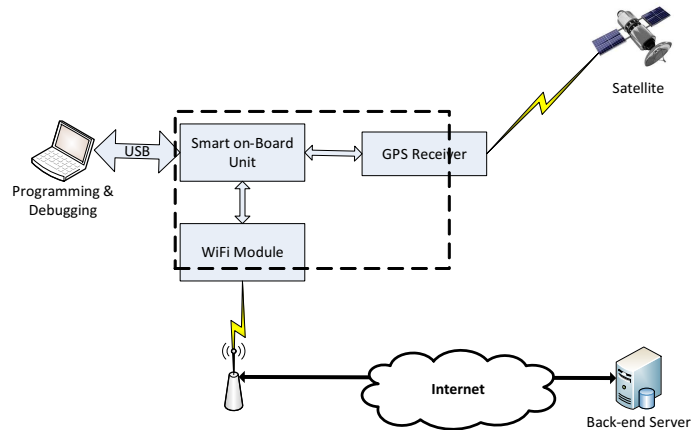


Fig. 1: System High Level Architecture.

C. GPS Data Analysis

1) *Injecting GPS Data into Map:* We utilized Open Street Map [13] to inject the travel path along the trip by using the javascript library Open Layer [14].

2) *Measurements and Statistics :* Different statistics can be carried out on the collected data with different criteria it cloud be statistics on a given trip, a give route, a give time period, or on a given vehicle behaviour. Driver's Average Speed distribution - String -For each trip we calculate the average speed along the trip after having number of runs/trips we cloud identify the best fit distribution that describes the average along given path per driver/vehicle moreover since we consider a give road followed by every public transportation vehicle(mandatory) so the total trip distance in kilome ter is fixed (i.e. 25 km for we consider path Toulkarem/Nablus). But the trip traveling duration directly proportional to the average speed such information cloud be used to estimate the vehicle arrival at the given bus stop and cloud be used for enhancing the travel scheduling to be more realistic. pathway's Average Speed distribution - String - For each trip we calculate the average speed along the trip after having number of runs/trips we cloud identify the best fit distribution that describes the average along given path per route/pathway.

IV. ARCHITECTURE AND IMPLEMENTATIONS

This section depicts the high level architecture(as shown in Figure 1) of the proposed system by identifying the main building blocks which consists the SBU which is mounted and attached to the vehicles. In addition, the Web based application running on the back-end server is introduced.

A. Smart on-Board Unit (SBU) MAIN COMPONENTS

Here we provide detailed description about the used hardware modules by the *SBU*:

- 1) The Arduino Uno (DEV-11021)Figure 2:which is an open hardware source controller, which is recently being used in many applications, due to its high performance, and easy to deal with. the Arduino Microcontroller board is based on the ATmega328, which has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, a 16 MHz ceramic resonator,32k Flash Memory, a USB connection, a power jack, and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable or power it with an AC to DC adapter or battery to get started [15],
- 2) Arduino GPS Shield (GPS-10710) Figure 3:is a high accuracy GPS receiver, which is used in our system due to its great characteristics and features such as this module can be easily integrated to the Arduino board, GPS-10710 is able to give the vehicle location within a few meters, this GPS module also gives accurate time reading which is an important feature to provide a good distributed synchronization mechanism to our system and all the control messages between the GPS receiver and the Arduino microcontroller are performed using the well-known AT commands standard [15],
- 3) Arduino Wi-Fi Shield (DEV-11287) Figure 4: this

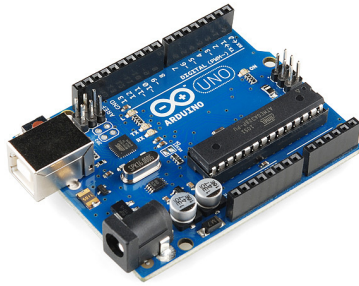


Fig. 2: Arduino Uno Microcontroller [15].

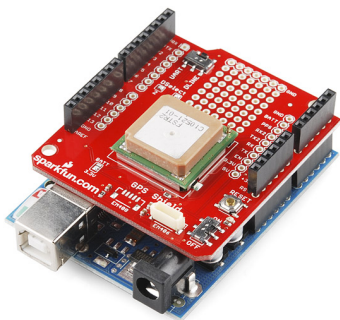


Fig. 3: GPS-10710 Shield Module [15].

Shield allows an Arduino board to connect to the internet using the 802.11b/g wireless specification (WiFi).the shield has an Atmega 32UC3 which provides a network (IP) stack capable of both TCP and UDP. In addition, the shield has an onboard micro-SD card slot, which is mainly used to locally store the user data on SBU along the trip, before being transferred to the central server via WiFi link. [15].

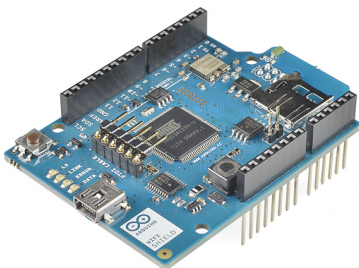


Fig. 4: Wi-Fi Shield Module [15].

B. Web Based Application

In order to make our Web Application flexible and extendable, we have adapted the REST (RESTful) architecture. And our implemented system has used the three-tier architecture [28]:

- 1) A front-end which relate to the client side. The user interface is based on a web-browsers application. It contains a responsive web page developed using Hypertext Transfer Markup Language v.5 HTML5 , Javascript , JQuery library and Cascading Style Sheet (CSS) whose application is tested on both desktop and smartphone web browsers. This web page uses Asynchronous JavaScript AJAX in order to build bidirectional data flow with middle layer.
- 2) A middle layer which includes a dynamic PHP program running on top of Apache web server. This program exposes its internal functionality through a RESTful interface towards the front-end and it uses the MySQL native driver for PHP for storing and retrieving data.
- 3) A back-end containing MySQL database server used to store all known roads in the region, system users, users profiles and user alerts. This component is a relational database that is used to store and retrieve the data. Note that the positioning and speed data are time-stamped according to the UTC time reference.

In RESTful vocabulary things are resources. Each resource is a uniquely addressable entity by a Universal Unique Identifier (URI) attached to it. Moreover, each resource has a representation which can be transferred and manipulated by means of four verbs. These verbs are create, read, update and delete (CRUD).

V. FUNCTIONAL TESTING AND USER EXPERIENCE

In order to test the system prototype, we attached our system box to a public transportation vehicle (mini-bus) traveling on the same route/path between two cities; namely, from Tulkarem to Nablus.

we have collocated and locally recorded the trips information (i.e. long., lat., Speed, and timestamp),based on the system model, these information will be transferred via WiFi link to the gateway allocated in the final bus station, by turn, such data will be inserted in the corresponding DataBase. in order to analyze drivers behavior, we are interested in two kind of plots, first plot Figure 5 depicts for each trip, the vehicle speed along the route, while in the second plot as in Figure 6 we show the traveled path along the trip, notice that, the corresponding plot was generated using Open Street Maps (OSM) API [13]. we have considered for analysis Five different trips of the same driver with the same vehicle, it is worth to mention that, these data are collected almost in the same time during a normal working day with almost same weather conditions.

in order to analyze the speed violation, we know that the maximum allowed speed limit on the outside city roads in Palestine is configured to be 90 Km/h, while it is 60 Km/h for inside city road segments. by looking to Figure 5, we notice that, the driver in first trip does not exceed the maximum allowed speed limit; the dotted horizontal line,

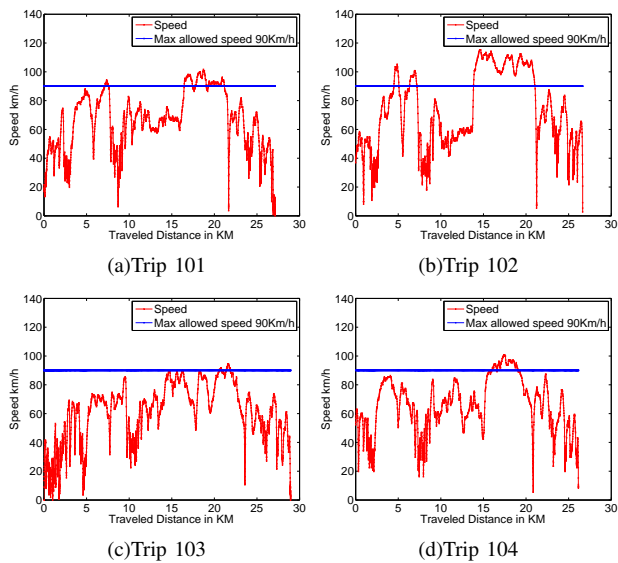


Fig. 5: Driving speed along the route for Four different trips: (a) Trip 101; (b) Trip 102; (c) Trip 103; and, (d) Trip 104.

indicates the 90Km/h speed limit, while the connected dotted line is related to 60Km/h speed limit. moreover, we could notice that the driver in the second trip 5(b) violates the speed limit frequently in comparison to other trips. on the contrary, the driver in the third trip 5(c) almost did not violates the maximum allowed speed. for more analysis. some related statistics are shown in the table III. it shows for each of the four trips the corresponding, ID, starting date and time, traveled distance in Kilometers unit, traveling trip duration in minutes, average vehicle speed along the whole trip, vehicles maximum reached traveling speed, the maximum legal allowed speed on the outside city road segments, the exceeding max speed duration in Minutes, this metric accumulates the time duration when driver exceeds the speed limit, while the last metric, indicates the percentage of violation duration to the total trip duration. from the statistics in the earlier table, it is easy to notice that the driver violates the speed during the second trip, where the violation percentage is almost 20% of the time.

Furthermore, Figure 6 shows the followed path for the first four trips, we found that all trips follow the same path along the route, in addition, we indicates on the track (red points) the segments where the driver violates the maximum speed limit, in fact, by referring to the authorized national transportation system, we found that recently, many traffic accidents happened on that segment.

VI. CONCLUSIONS

Recently, the demand for developing public transportation management systems (PTMS) using GPS technologies have sharply increased due to the fact that, a well designed PTMS will save human life by monitoring the driver behaviour which in turn will reduce number of times when the driver violate traffic regulations. This paper introduces system composition structure and explains the system software and hardware design. Experiments show that our system is practicable and reliable of data transmissions using WiFi links, with compar-

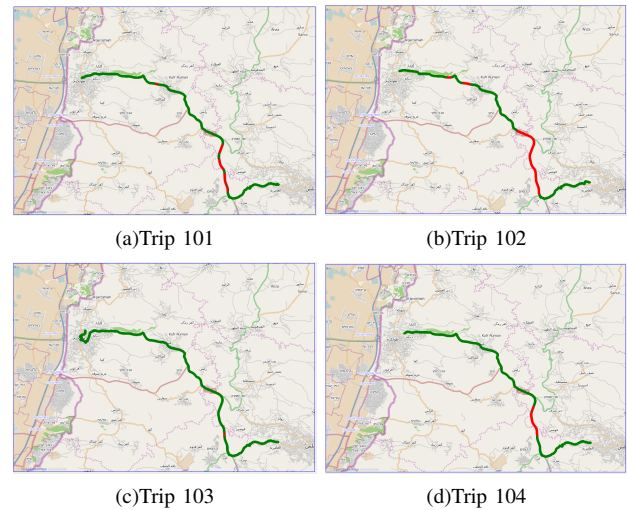


Fig. 6: Traveled Path along the route for Four different trips: (a) Trip 101; (b) Trip 102; (c) Trip 103; and, (d) Trip 104.

ison with the others management systems based on GPRS and GPS technology, it is greatly reduced the operating price. In our system, the Open Street Maps plays a major rule during monitoring, visualization and identifying the maximum allowed speed for each road segment along with the traveled route. Ongoing work is to enable the system for supporting online tracking and monitoring services, by using mobile data networks.

REFERENCES

- [1] S. Tarapiah, S. Atalla, N. Muala, and S. Tarabeh, "Offline public transportation management system based on GPS/WiFi and open street maps," in *CICSyN2014, 6th Int Conference on Computational Intelligence, Communication Systems and Networks (CICSyN2014)*, Tetovo, Macedonia, the former Yugoslav Republic of, May 2014.
- [2] S. Tarapiah, R. AbuHania, and D. J. Islam Hindi, "Applying web based gps/gprs ticketing and tracking mechanism to reduce traffic violation in developing countries," in *The International Conference on Digital Information Processing, E-Business and Cloud Computing (DIPECC2013)*. The Society of Digital Information and Wireless Communication, 2013, pp. 102–106.
- [3] S. Tarapiah, S. Atalla, and R. AbuHania, "Smart on-board transportation management system using gps/gsm/gprs technologies to reduce traffic violation in developing countries," *International Journal of Digital Information and Wireless Communications (IJDWC)*, vol. 3, no. 4, pp. 96–105, 2013.
- [4] S. Tarapiah, S. Atalla, and B. Alsaid, "Smart On-Board transportation management system Geo-Casting featured," in *International Conference on Computer Information Systems 2014 (ICCIS-2014)*, Hammamet, Tunisia, Jan. 2014.
- [5] G. A. Giannopoulos, "The application of information and communication technologies in transport," *European Journal of Operational Research*, vol. 152, no. 2, pp. 302–320, 2004.
- [6] M. Al-Rousan, A. Al-Ali, and K. Darwish, "Gsm-based mobile tele-monitoring and management system for inter-cities public transportations," in *Industrial Technology, 2004. IEEE ICIT'04. 2004 IEEE International Conference on*, vol. 2. IEEE, 2004, pp. 859–862.
- [7] M. Popa and B. Suta, "A solution for tracking a fleet of vehicles," in *Telecommunications Forum (TELFOR), 2011 19th*. IEEE, 2011, pp. 1558–1561.
- [8] C. Koukourlis, S. Spyridakis, and N. Kokkalis, "On the design of a fleet monitoring system with reduced power consumption," *Electrical*

Trip ID	1	2	3	4	5
Starting Time	1_19_2014 9_22am	1_20_2014 8_16am	1_21_2014 9_37am	1_22_2014 8_28am	1_23_2014 9_34am
Distance (Km)	27.122	26.65	28.95	26.15	25.03
Duration (minutes)	28.62	25.4	34.48	26.36	26.8
Average Speed (Km/H)	56.86	62.96	50.37	59.5	56.03
Maximum Speed	101.7	115.2	94.5	100.8	109.8
Maximum Allowed Speed(km/h)	90	90	90	90	90
Exceeding Max allowed Speed Duration	2.85	4.95	0.37	1.9	3.43
Violating Speed (%)	09.69	19.49	01.06	07.20	12.81

TABLE III: Trips Related Statistics.

- Engineering*, vol. 84, no. 4, pp. 203–210, 2002. [Online]. Available: <http://dx.doi.org/10.1007/s00202-002-0120-z>
- [9] I. M. Almomani, N. Y. Alkhalil, E. M. Ahmad, and R. M. Jodeh, “Ubiquitous gps vehicle tracking and management system,” in *Applied Electrical Engineering and Computing Technologies (AEECT), 2011 IEEE Jordan Conference on*. IEEE, 2011, pp. 1–6.
- [10] I. A. H. Eltoun and M. Bouhorma, “Velocity based tracking and localization system using smartphones with gps and gprs/3g,” *International Journal of Computer Applications*, vol. 76, 2013.
- [11] J. Biagioni, T. Gerlich, T. Merrifield, and J. Eriksson, “Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones,” in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2011, pp. 68–81.
- [12] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, “Map-matching for low-sampling-rate gps trajectories,” in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2009, pp. 352–361.
- [13] “Open Street Map,” 2014. [Online]. Available: <http://www.openstreetmap.org/>
- [14] “Javascript Library Open Layer,” 2014. [Online]. Available: <http://openlayers.org/>
- [15] “SparkFun Electronics,” 2013. [Online]. Available: www.sparkfun.com