

Volume 2 Issue 10

October 2011



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



[www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)



INTERNATIONAL JOURNAL OF  
ADVANCED COMPUTER SCIENCE AND APPLICATIONS



A Publication of  
The Science and Information Organization



## IJACSA Editorial

### *From the Desk of Managing Editor...*

It is a pleasure to present our readers with the October 2011 Issue of International Journal of Advanced Computer Science and Applications (IJACSA).

The renaissance stimulated by the field of Computer Science is generating multiple formats and channels of communication and creativity. IJACSA is one of the most prominent publications in the field and engaging the ubiquitous spread of subject knowledge with effectiveness in all classes of audience. Nevertheless, the promise of increased engagement requires that we consider how this might be accomplished, delivering up-to-date and authoritative coverage of advanced computer science and applications.

The journal has a wide scope ranging from the many facets of methodological foundations to the details of technical issues and the aspects of industrial practice. It includes articles related to research findings, technical evaluations, and reviews. In addition it provides a forum for the exchange of information on all aspects.

The editorial board of the IJACSA consists of individuals who are committed to the search for high-quality research suitable for publication. These individuals, working with the editor to achieve IJACSA objectives, assess the quality, relevance, and readability of individual articles.

The contents include original research and innovative applications from all parts of the world. This interdisciplinary journal has brought together researchers from academia and industry as well as practitioners to share ideas, problems and solutions relating to computer science and application with its convergence strategies, and to disseminate the most innovative research. As a consequence only 28% of the received articles have been finally accepted for publication.

Therefore, IJACSA in general, could serve as a reliable resource for everybody loosely or tightly attached to this field of science.

The published papers are expected to present results of significant value to solve the various problems with application services and other problems which are within the scope of IJACSA. In addition, we expect they will trigger further related research and technological improvements relevant to our future lives.

We hope to continue exploring the always diverse and often astonishing fields in Advanced Computer Science and Applications.

**Thank You for Sharing Wisdom!**

**Managing Editor**

**IJACSA**

**Volume 2 Issue 10, October 2011**

[editorijacsa@thesai.org](mailto:editorijacsa@thesai.org)

**ISSN 2156-5570 (Online)**

**ISSN 2158-107X (Print)**

**©2011 The Science and Information (SAI) Organization**

# Editorial Board

**Dr. Kohei Arai – Editor-in-Chief**

**Saga University**

Domains of Research: Human-Computer Interaction, Networking, Information Retrievals, Optimization Theory, Modeling and Simulation, Satellite Remote Sensing, Computer Vision, Decision Making Methodology

**Dr. Ka Lok Man**

**Xi'an Jiaotong-Liverpool University (XJTLU)**

Domain of Research: Computer Science and Microelectronics

**Dr. Sasan Adibi**

**Research In Motion (RIM)**

Domain of Research: Security of wireless systems, Quality of Service

**Dr. Zuqing Zuh**

**University of Science and Technology of China**

Domains of Research : Optical Communication Systems, Optical network architecture and design, Next generation Internet, Signal processing, Broadband access network, such as cable access (DOCSIS) networks, passive optical networks (PON), fiber to the home (FTTH), Energy-efficient network and green technologies

**Dr. Sikha Bagui**

**University of West Florida**

Domain of Research: Database, database modeling, ER diagrams, XML data, web databases, data mining, association rule mining, data preprocessing

**Dr. T. V. Prasad**

**Lingaya's University**

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

**Dr. Mohd Helmy Abd Wahab**

**Universiti Tun Hussein Onn Malaysia**

Domain of Research: Data Mining, Database, Web-based Application, Mobile Computing

---



## IJACSA Reviewer Board

- **A Kathirvel**  
Karpaga Vinayaka College of Engineering and Technology, India
- **Abbas Karimi**  
I.A.U\_Arak Branch (Faculty Member) & Universiti Putra Malaysia
- **Dr. Abdul Wahid**  
Gautam Buddha University, India
- **Abdul Khader Jilani Saudagar**  
Al-Imam Muhammad Ibn Saud Islamic University
- **Abdur Rashid Khan**  
Gomal University
- **Dr. Ahmed Nabih Zaki Rashed**  
Menoufia University, Egypt
- **Ahmed Sabah AL-Jumaili**  
Ahlia University
- **Md. Akbar Hossain**  
Aalborg University, Denmark and AIT, Greeceas
- **Albert Alexander**  
Kongu Engineering College,India
- **Prof. Alcinea Zita Sampaio**  
Technical University of Lisbon
- **Amit Verma**  
Rayat & Bahra Engineering College, India
- **Ammar Mohammed Ammar**  
Department of Computer Science, University of Koblenz-Landau
- **Arash Habibi Lashakri**  
University Technology Malaysia (UTM), Malaysia
- **Asoke Nath**  
St. Xaviers College, India
- **B R SARATH KUMAR**  
Lenora College of Engineering, India
- **Binod Kumar**  
Lakshmi Narayan College of Technology, India
- **Bremananth Ramachandran**  
School of EEE, Nanyang Technological University
- **Dr.C.Suresh Gnana Dhas**  
Park College of Engineering and Technology, India
- **Mr. Chakresh kumar**  
Manav Rachna International University, India
- **Chandra Mouli P.V.S.S.R**  
VIT University, India
- **Chandrashekhar Meshram**  
Shri Shankaracharya Engineering College, India
- **Constantin POPESCU**  
Department of Mathematics and Computer Science, University of Oradea
- **Prof. D. S. R. Murthy**  
SNIST, India.
- **Deepak Garg**  
Thapar University.
- **Prof. Dhananjay R.Kalbande**  
Sardar Patel Institute of Technology, India
- **Dhirendra Mishra**  
SVKM's NMIMS University, India
- **Divya Prakash Shrivastava**  
EL JABAL AL GARBI UNIVERSITY, ZAWIA
- **Dragana Becejski-Vujaklija**  
University of Belgrade, Faculty of organizational sciences
- **Fokrul Alom Mazarbhuiya**  
King Khalid University
- **G. Sreedhar**  
Rashtriya Sanskrit University
- **Ghalem Belalem**  
University of Oran (Es Senia)
- **Hanumanthappa.J**  
University of Mangalore, India
- **Dr. Himanshu Aggarwal**  
Punjabi University, India
- **Huda K. AL-Jobori**  
Ahlia University
- **Dr. Jamaiah Haji Yahaya**  
Northern University of Malaysia (UUM), Malaysia
- **Jasvir Singh**  
Communication Signal Processing Research Lab
- **Jatinderkumar R. Saini**  
S.P.College of Engineering, Gujarat
- **Prof. Joe-Sam Chou**  
Nanhua University, Taiwan
- **Dr. Juan José Martínez Castillo**  
Yacambu University, Venezuela
- **Dr. Jui-Pin Yang**

- Shih Chien University, Taiwan
- **Dr. K.PRASADH**  
Mets School of Engineering, India
  - **Ka Lok Man**  
Xi'an Jiaotong-Liverpool University (XJTLU)
  - **Dr. Kamal Shah**  
St. Francis Institute of Technology, India
  - **Kodge B. G.**  
S. V. College, India
  - **Kohei Arai**  
Saga University
  - **Kunal Patel**  
Ingenuity Systems, USA
  - **Lai Khin Wee**  
Technischen Universität Ilmenau, Germany
  - **Latha Parthiban**  
SSN College of Engineering, Kalavakkam
  - **Mr. Lijian Sun**  
Chinese Academy of Surveying and Mapping, China
  - **Long Chen**  
Qualcomm Incorporated
  - **M.V.Raghavendra**  
Swathi Institute of Technology & Sciences, India.
  - **Madjid Khalilian**  
Islamic Azad University
  - **Mahesh Chandra**  
B.I.T, India
  - **Mahmoud M. A. Abd Ellatif**  
Mansoura University
  - **Manpreet Singh Manna**  
SLIET University, Govt. of India
  - **Marcellin Julius NKENLIFACK**  
University of Dschang
  - **Md. Masud Rana**  
Khunla University of Engineering & Technology,  
Bangladesh
  - **Md. Zia Ur Rahman**  
Narasaraopeta Engg. College, Narasaraopeta
  - **Messaouda AZZOUZI**  
Ziane AChour University of Djelfa
  - **Dr. Michael Watts**  
University of Adelaide, Australia
  - **Miroslav Baca**  
University of Zagreb, Faculty of organization and  
informatics / Center for biomet
  - **Mohamed Ali Mahjoub**  
Preparatory Institute of Engineer of Monastir
  - **Mohammad Talib**  
University of Botswana, Gaborone
  - **Mohammed Ali Hussain**  
Sri Sai Madhavi Institute of Science & Technology
  - **Mohd Helmy Abd Wahab**  
Universiti Tun Hussein Onn Malaysia
  - **Mohd Nazri Ismail**  
University of Kuala Lumpur (UniKL)
  - **Mueen Uddin**  
Universiti Teknologi Malaysia UTM
  - **Dr. Murugesan N**  
Government Arts College (Autonomous), India
  - **Nitin S. Choubey**  
Mukesh Patel School of Technology Management &  
Eng
  - **Dr. Nitin Surajkishor**  
NMIMS, India
  - **Paresh V Virparia**  
Sardar Patel University
  - **Dr. Poonam Garg**  
Institute of Management Technology, Ghaziabad
  - **Raj Gaurang Tiwari**  
AZAD Institute of Engineering and Technology
  - **Rajesh Kumar**  
National University of Singapore
  - **Rajesh K Shukla**  
Sagar Institute of Research & Technology-  
Excellence, India
  - **Dr. Rajiv Dharaskar**  
GH Raison College of Engineering, India
  - **Prof. Rakesh. L**  
Vijetha Institute of Technology, India
  - **Prof. Rashid Sheikh**  
Acropolis Institute of Technology and Research,  
India
  - **Ravi Prakash**  
University of Mumbai
  - **Rongrong Ji**  
Columbia University
  - **Dr. Ruchika Malhotra**  
Delhi Technological University, India
  - **Dr.Sagarmay Deb**  
University Lecturer, Central Queensland University,  
Australia

- **Saleh Ali K. AlOmari**  
Universiti Sains Malaysia
- **Dr. Sana'a Wafa Al-Sayegh**  
University College of Applied Sciences UCAS-  
Palestine
- **Santosh Kumar**  
Graphic Era University, India
- **Sasan Adibi**  
Research In Motion (RIM)
- **Saurabh Pal**  
VBS Purvanchal University, Jaunpur
- **Seyed Hamidreza Mohades Kasaei**  
University of Isfahan
- **Shahanawaj Ahamad**  
The University of Al-Kharj
- **Shaidah Jusoh**  
University of West Florida
- **Sikha Bagui**  
Zarqa University
- **Dr. Smita Rajpal**  
ITM University
- **Suhas J Manangi**  
Microsoft
- **SUKUMAR SENTHILKUMAR**  
Universiti Sains Malaysia
- **Sunil Taneja**  
Smt. Aruna Asaf Ali Government Post Graduate  
College, India
- **Dr. Suresh Sankaranarayanan**  
University of West Indies, Kingston, Jamaica
- **T C.Manjunath**  
Visvesvaraya Tech. University
- **T V Narayana Rao**  
Hyderabad Institute of Technology and  
Management, India
- **T. V. Prasad**  
Lingaya's University
- **Taiwo Ayodele**  
Lingaya's University
- **Totok R. Biyanto**  
Infonetmedia/University of Portsmouth
- **Varun Kumar**  
Institute of Technology and Management, India
- **Vellanki Uma Kanta Sastry**  
Sreeneedhi
- **Dr. V. U. K. Sastry**  
SreeNidhi Institute of Science and Technology  
(SNIST), Hyderabad, India.
- **Vinayak Bairagi**  
Sinhgad Academy of engineering, India
- **Vitus S.W. Lam**  
The University of Hong Kong
- **Vuda Sreenivasarao**  
St.Mary's college of Engineering & Technology,  
Hyderabad, India
- **Y Srinivas**  
GITAM University
- **Mr.Zhao Zhang**  
City University of Hong Kong, Kowloon, Hong Kong
- **Zhixin Chen**  
ILX Lightwave Corporation
- **Zuqing Zhu**  
University of Science and Technology of China

# CONTENTS

**Paper 1: Data Mining for Engineering Schools**

*Authors: Chady El Moucary*

**PAGE 1 – 9**

**Paper 2: Model of Temperature Dependence Shape of Ytterbium -doped Fiber Amplifier Operating at 915 nm Pumping Configuration**

*Authors: Abdel Hakeim M. Husein, Fady I. EL-Nahal*

**PAGE 10 – 13**

**Paper 3: Healthcare Providers' Perceptions towards Health Information Applications at King Abdul-Aziz Medical City, Saudi Arabia**

*Authors: Abeer Al-Harbi*

**PAGE 14 – 22**

**Paper 4: Route Maintenance Approach For Link Breakage Prediction In Mobile Ad Hoc Networks**

*Authors: Khalid Zahedi, Abdul Samad Ismail*

**PAGE 23 – 30**

**Paper 5: Web Service Architecture for a Meta Search Engine**

*Authors: K.Srinivas, P.V.S. Srinivas, A.Govardhan*

**PAGE 31 – 36**

**Paper 6: A Statistical Approach For Latin Handwritten Digit Recognition**

*Authors: Ihab Zaqout*

**PAGE 37 – 40**

**Paper 7: Plant Leaf Recognition using Shape based Features and Neural Network classifiers**

*Authors: Jyotismita Chaki, Ranjan Parekh*

**PAGE 41 – 47**

**Paper 8: Integrated Information System for reserving rooms in Hotels**

*Authors: Dr. Safarini Osama*

**PAGE 48 – 52**

**Paper 9: Automatic Classification and Segmentation of Brain Tumor in CT Images using Optimal Dominant Gray level Run length Texture Features**

*Authors: A.Padma, R.Sukanesh*

**PAGE 53 – 59**



**Paper 10: An Ontology- and Constraint-based Approach for Dynamic Personalized Planning in Renal Disease Management**

*Authors: Normadiyah Mahiddin, Yu-N Cheah, Fazilah Haron*

**PAGE 60 – 69**

**Paper 11: Asynchronous Checkpointing And Optimistic Message Logging For Mobile Ad Hoc Networks**

*Authors: Ruchi Tuli, Parveen Kumar*

**PAGE 70 – 76**

**Paper 12: An Experimental Improvement Analysis of Loss Tolerant TCP (LT-TCP) For Wireless Network**

*Authors: Md. Abdullah Al Mamun, Momotaz Begum, Sumaya kazary, Md. Rubel*

**PAGE 77 – 80**

**Paper 13: Comparison of Workflow Scheduling Algorithms in Cloud Computing**

*Authors: Navjot Kaur, Taranjit Singh Aulakh, Rajbir Singh Cheema*

**PAGE 81 – 86**

**Paper 14: Study Of Indian Banks Websites For Cyber Crime Safety Mechansim**

*Authors: Susheel Chandra Bhatt, Durgesh Pant*

**PAGE 87 – 90**

**Paper 15: An Empirical Study of the Applications of Web Mining Techniques in Health Care**

*Authors: Dr. Varun Kumar, MD. Ezaz Ahmed*

**PAGE 91 – 94**

**Paper 16: Quality EContent Design using Reusability approach**

*Authors: Senthil Kumar.J, Dr S.K.Srivatsa*

**PAGE 95 – 97**

**Paper 17: Retrieval of Images Using DCT and DCT Wavelet Over Image Blocks**

*Authors: H. B. kekre, Kavita Sonawane*

**PAGE 98 – 106**

**Paper 18: Cryptanalysis of An Advanced Authentication Scheme**

*Authors: Sattar J Aboud, Abid T. Al Ajeeli*

**PAGE 107 – 111**

**Paper 19: Conceptual Level Design of Semi-structured Database System: Graph-semantic Based Approach**

*Authors: Anirban Sarkar*

**PAGE 112 – 121**

# Data Mining for Engineering Schools

## Predicting Students' Performance and Enrollment in Masters Programs

Chady El Moucary

Department of Electrical, Computer and Communication Engineering, Faculty of Engineering  
Notre Dame University – Louaize (NDU)  
North Lebanon Campus – P.O. Box 87, Tripoli – Municipality Street, Barsa – El Koura, Lebanon

**Abstract**— the supervision of the academic performance of engineering students is vital during an early stage of their curricula. Indeed, their grades in specific core/major courses as well as their cumulative General Point Average (GPA) are decisive when pertaining to their ability/condition to pursue Masters' studies or graduate from a five-year Bachelor-of-Engineering program. Furthermore, these compelling strict requirements not only significantly affect the attrition rates in engineering studies (on top of probation and suspension) but also decide of grant management, developing courseware, and scheduling of programs. In this paper, we present a study that has a twofold objective. First, it attempts at correlating the aforementioned issues with the engineering students' performance in some key courses taken at early stages of their curricula, then, a predictive model is presented and refined in order to endow advisors and administrators with a powerful decision-making tool when tackling such highly important issues. Matlab Neural Networks Pattern Recognition tool as well as Classification and Regression Trees (CART) are fully deployed with important cross validation and testing. Simulation and prediction results demonstrated a high level of accuracy and offered efficient analysis and information pertinent to the management of engineering schools and programs in the frame of the aforementioned perspective.

**Keywords-component;** *Educational Data Mining; Classification and Regression Trees (CART); Relieff tool; Neural Networks; Prediction; Engineering Students' Performance; Engineering Students' Enrollment in Masters' Studies.*

### I. INTRODUCTION

Data mining has attracted exceptionally diversified businesses for both the descriptive and predictive capabilities it promises, one of which is Education in its broad fields and organizational hierarchies [15] [36] [50]. In fact, Education, nowadays, not only involves the *ancestral* information- and/or knowledge- communication and transfer, but has also become a standalone and comprehensive *business* with excessive demands in information handling and analysis, as well as the management of a deeply spread tree of positions and interrelated functions [49]. Indeed, Education's features and attributes have dramatically shifted and augmented to a point where the integration of technology became inevitable in the attempt of sustaining a good position and thriving in a highly competitive and merciless market. This technology not only involves new teaching methodologies but also osculates with every single aspect of management of such institutions.

Furthermore, management of large amount of data has become undeniably forbearing to even expert staff; it even requires more powerful computational and specifications requirements when referring to machines and/or algorithms. Diversified challenges face Education and which fortunately attracted researchers from different fields of expertise who keep straining in order to achieve innovative but also *intelligent* techniques to help keep up with the pressure and find astute and reasonable answers to multifaceted questions. Globalization, International Accreditation, and e-learning have only added more threads to the pile.

Data mining, which is the science of digging into databases for information and knowledge retrieval, has recently developed new axes of applications and engendered an emerging discipline, called Educational Data Mining or EDM. This discipline seems to be a lot promising. EDM carries out tasks such as prediction (classification, regression, and density estimation), clustering, relationship mining (association, correlation, sequential mining, and causal data mining), distillation of data for human judgment, and discovery with models [1]. Moreover, by exercising EDM, educators and administrators can tackle both traditional and ad hoc educational issues and benefit from a good decision-making tool when facing challenges and/or exploring new horizons in their specialties. The list is long and requires wide and long tables to fit in. Nevertheless, in a non-exhaustive list, we can enumerate the most frequently inquiring subject matters such as predicting students' performance, developing courseware, students' behavioral modeling, strategic planning and scheduling of programs, supervising attrition rates, and grant management, etc. In other words, EDM aims at enhancing the understanding and supervision of learners', teachers', and administrators' domain representation, pedagogical engagement and behaviors [5] [18] [32] [37] [38] [39].

Remarkable amount of EDM endeavors have been conducted and published in many journals and conference proceedings related to, but not limited to, Artificial Intelligence, Learning Systems, Education, and others. In July 2011 the International Educational Data Mining Society [2] was founded by the International Working Group on Educational Data Mining with the main objective of capturing contributions from the EDM community and offering a forum for practitioners to impart their labor and exchange their competencies. It has so far organized four international conferences where recognized work can be archived in the

Journal of Educational Data Mining or JEDM (ISSN 2157-2100) [28].

One can find many definitions for data mining in books, journal papers, and e-articles [11] [12] [13] [14]. They all refer to data mining as a young and interdisciplinary field in computer science which is described as an *interactive and iterative* process aiming at *sundering out/revealing hidden/unobvious, but existing, patterns, trends and/or relationships* amidst data using statistical and mathematical procedures with a prime objective of providing decision support systems with information and knowledge. Furthermore, data mining is being recently interchangeably used with what is referred to as Knowledge Discovery in Database or KDD namely when excessively large data repositories is being involved. Fig. 1 shows Data Mining exercise as a step towards KDD [10].

A typical example of inexorable flood of data is the Europe's Very Long Baseline Interferometry (VLBI), which has 16 telescopes, each of which produces one Gigabit/second of astronomical data over a 25-day observation session [6]. An interesting overview of the largest databases in the world can be found in [7] [8]. The top-ten lists include The Library of Congress (LC) with over 130 million items, 530 miles of shelves, 5 million digital documents, and 20 terabytes of text data. The Central Intelligence Agency (CIA) possesses comprehensive statistics on more than 250 countries and entities (unknown number of classified information). Amazon, the world's biggest retail store, maintains over 59 million active customers ending up with over 42 terabytes of data. YouTube, the largest video library, observes more than 65,000 videos added each day and encompasses at least 45 terabytes of videos. ChoicePoint, the business of acquiring information about the American population (addresses, phone numbers, driving records, etc.) possesses a database that extends to the moon and back 77 times and holds over 250 terabytes of personal data. Sprint, one of the world's largest telecommunication companies, offers its services to more than 53 million subscribers with a 2.85 trillion database rows and 70,000 call detail-record insertions per second. Google, the

famous search engine and industry, is subjected to 91 million searches per day (accounting to 50% of all internet search activity) and holds more than 33 trillion database entries, etc. It is spectacularly evident that traditional warehousing techniques and querying algorithms cannot cope with such colossal amount of data, thus, new techniques for information retrieval urged tons of research papers in the data mining/KDD field. Many strains have been deployed to manipulate and handle considerable amount of data [20] [35] [47] [48], but in many applications, the data available only covers parts of the inference chain from evidence to actions. However, a versed miner can extrapolate a small amount of initial knowledge into more knowledge using proficient mining.

In this paper, we will present a study that aims at offering a reliable and predictive tool for academicians and administrators working in engineering schools and universities to monitor students' performance at an early stage of their educational path. The goal is to link this observation/data with students' chances to either finish (succeed) a five-year Bachelor-of-Engineering program (BE) or enroll in Masters' program in a BS/MS track.

In the section to come, Data Mining will be reviewed and presented from different perspectives with emphasis on its various categories, tasks and implementations. In Section III, we will elaborate on the tool developed and underline data preparation and attributes' selection. Furthermore, the use of Neural Networks and Classification and Regression Trees (CART) will be explained and applied with cross-validation and pruning [40]. Error Histogram and ROC curves will also be studied in section III. Finally, section IV will portray a thorough analysis and discussions of the results. The core objective of the paper will be summarized and a conclusion is presented in this section as well.

## II. DATA MINING

### A. A Multifaceted Discipline

Data mining is a twofold discipline in the sense that it subtends two high-level primary objectives: prediction and description [10]. **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest. **Description** focuses on finding human-interpretable patterns describing the data. The relative importance of prediction and description for particular data mining applications can vary considerably. While in the context of KDD description tends to be more important than prediction, prediction is often the primary goal in pattern recognition and machine learning applications.

### B. Learning Approaches and Techniques

Data mining can be described as either supervised or unsupervised [9] [48] as shown in Fig. 2. Unsupervised data mining is rather a bottom-up approach that makes no prior assumptions and aims at discovering relationships in the data. In this sense, data are allowed to speak for themselves; there is no distinction between attributes and targets. In this context, unsupervised data mining is a descriptive approach. Typical methods and applications are clustering, density estimation, data segmentation, smoothing, etc. Supervised data mining, also called direct data mining, aims at explaining those

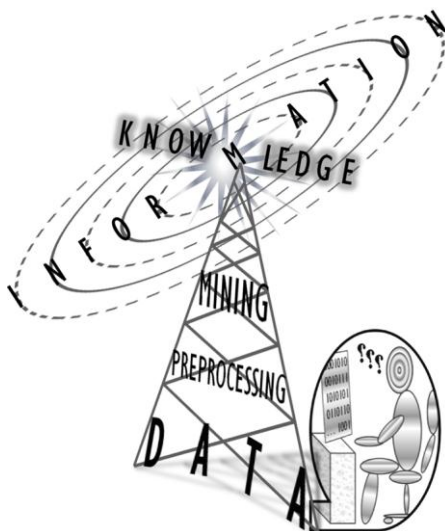


Figure 1 - Data Mining, a Mandatory Step towards KDD

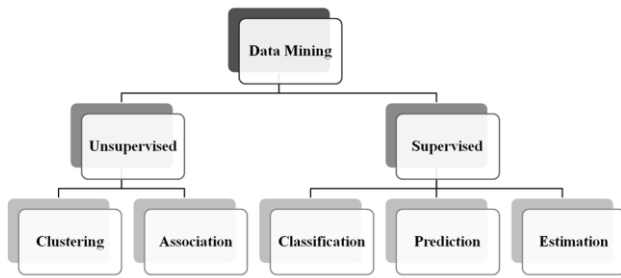


Figure 2 – Data Mining Learning Techniques

relationships once they are found. It is rather a predictive approach where the variables involved are classified as explanatory and dependent ones, and where the main goal is to achieve a liaison between them as in Regression Analysis. Typically, the target variable has to be well specified in advance and three important steps follow before achieving data mining or KDD purposes, as illustrated in Fig. 3.

Specifically, data is to be subdivided for **training, testing, and validation**. The objective of the training step is to construct a provisional model that attempts to subtend and/or engender the hidden relationship between the attributes and the target variable. The validation step plays a key role in reducing the overfitting traits of the model in the sense that it helps reduce the amount of unsatisfactory results or avoid patterns that are not present in the general dataset (sometimes called flat file). This would occur if the model has excessive number of attributes relative to the amount of data collected and available; the model will exaggerate minor fluctuations in the data and thus, have poor predictive capacity.

Another typical hitch could be an acquired *memory* characteristic that downsizes the model to specific cases in the training phase. For instance, assume that in the training data all students who have passed pre-calculus course have succeeded their graduate studies; we do not want the model to remember this liaison and create the pattern “if the student succeeds the pre-calculus course, then he/she will succeed the graduate studies”. Instead, the model should apply *all* patterns found in the training phase to the future data and thus, acquire a generalization characteristic/capability. A number of statistical techniques can be applied to assess the model such as the ROC

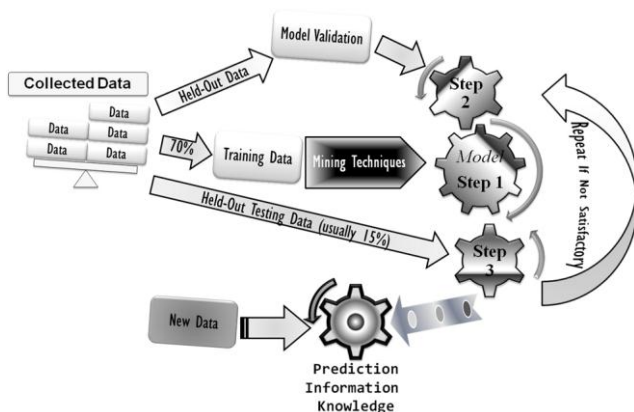


Figure 3 - Roadmap for an Efficient Predictive Tool

(Receiver/Relative Operating Characteristic) curves, which depict the true positive rate vs. the false positive one as the threshold of discrimination is modified in the case of a binary classifier. Finally, testing is used to evaluate, revisit and/or retrain the constructed model by using new *untrained* data that have been held out from the original complete dataset. At this point, the data miner should be able to ascertain whether the learned patterns meet the desired standards. If the outcome is satisfactory, the data miner shall transform the model into information and knowledge, otherwise, re-training, changing the pre-processing and/or the data-mining algorithm is inevitable. This is where the expertise of the data miner plays a decisive role. In fact, not only this expertise is crucial in rectifying the path of the mining process, but also in *converting* the information into knowledge in the sense of filling the gap between the *past* (existing data) and the *future* (prediction) and drawing an efficient roadmap (decision making and strategic planning) using the outcome of the KDD process.

### C. Data Mining Tasks

Data mining objectives can be carried out by means of various procedures, frequently called tasks. Thus a further categorization of data mining is obtained:

- **Classification:** Typical supervised-learning task where information is arranged into predefined classes according to some learnt rules. The learning process aims at developing a model for predicting/assigning the class of a new instance. In other words, it is the generalized application of a known structure to a new data for mapping and classification purposes. The most widely used classifiers are Decision Trees, Bayesian and Neural Networks, etc. This task is applied to diversified fields of expertise such as, Speech and Handwriting Recognition, Web Search Engines, Geostatistics (remote sensing), etc. [42] [44]
- **Regression:** The goal of this task is to achieve a function (regression function) of the independent variables that allows computing the conditional expectation of a dependent variable for prediction and forecasting exercises based on the minimization of a certain type of error via an iterative procedure. Practically, Classification and Regression Trees (CART) summarize these tasks; classification is referred to when the target is nominal, whereas regression is used for continuous values of the target (infinite number of values).
- **Clustering:** It is a descriptive and typical unsupervised machine-learning task common to statistical data analysis wherein a finite set of groups and clusters are identified to describe the data. It is referred to in applications such as Image Analysis, Machine Learning, Biology and Medicine, Pattern Recognition, Education, Crime Analysis, etc. The most reputed approaches related to this task are the K-Means [26] and Fuzzy Clustering techniques [25] [33].
- **Summarization:** Various methods are formulated to describe the set of data and information in a more



compact representation. It also includes report generation.

- *Association Rule Learning or Dependency Modeling:* This task aims at identifying frequent itemsets in a database and deriving association rules. A local model is identified and which describes the important dependencies between variables and datasets. It has been mainly used in applications involving decisions about marketing activities in supermarkets. It is also applied in the fields of Web Mining, Bioinformatics, etc. The most popular and famous algorithm used in this field is the *Apriori* [23] [24] [30] [31].
- *Outlier/Deviation Detection:* Important and significant deviations in the dataset are reported. It finds application in Fraud Detection, Intrusion Detection (Data Security), etc. It is used to increase accuracy by removing anomalous data from the dataset (supervised). Popular algorithms are based on K-Nearest Neighbor, Support Vector Machines, etc. [21] [22] [34].
- *Link Analysis:* Find relationships amongst richly structured databases where hidden patterns are somewhat difficult to be discerned using traditional statistical approaches [20].

Finally, it is noteworthy to mention that *estimation* and *prediction* are sometimes interchangeably used when dealing with classification and regression problems. The reality is that *estimation* is used when a *continuous* value is to be forecasted while *prediction* is referred to when new data is classified into one of predefined *classes*, which are predetermined when building the model.

#### D. Data Miner Role

It should be noted that efficient and plausible mining exercises remain highly dependent from data preprocessing such as gathering, cleaning, representation, etc. Indeed, although data mining tools and tasks can be very appealing, domain-specific skills are required as a prerequisite before embarking on the trip towards KDD. In other words, human interface plays a decisive role in somewhat deploying or *transforming* data from an opaque entity into a transparent one in order to be resourcefully processed [3]. The data miner has a fundamental role in formulating the problem and preparing the suitable and relevant data. Additionally, data mining can turn out non-satisfactory results at first attempts and miners are to integrate their expertise into the model before starting another iteration of the process. Moreover, astutely adjusting some parameters or trying out different algorithms not only reveals necessary but also requires relevant and consistent justifications. Particularly, the choice of apposite and pertinent attributes can grow intractable, intricate and strenuous namely with large and/or complex classes or datasets. In this sense, adept miners would simplify this task at an early stage and at low computational cost by refining and consolidating the raw data. A typical example would be of David Heckerman [4] about Hot-Dogs and Barbecue-Sauce false inference.

### III. PREDICTING ENGINEERING STUDENTS SUCCESS AND/OR ENROLLMENT IN MASTERS PROGRAMS

In this paper, we will deal with a particular concern that considerably affects engineering programs in various types of Higher Education Institutions. Pondering over the engineering students' primordial performance demonstrated imperative for an efficient supervision of the attrition rate (on top of probation and suspension rates) and students' chances to further enroll in Masters' studies or to simply achieve (succeed) their engineering degrees [45] [46] [53] [54]. We will, at a first stage, aim at discovering the relationship between the most affecting factors and the aforementioned issues, then, we will try to construct a predictive model that will endow both advisors and administrators with a powerful decision-making tool. The predictive tool or model, as we will call it here, will help tackle such issues as predicting students' GPA, the attrition rate in the Engineering program, and have an insight of the enrollment rate in Masters' studies. Consequently, it will decisively help in planning the courseware and designating needed faculty members, amongst many other pertinent and resulting matters [16].

Another benefit of this tool is to help advisors and instructors have an insight about their students, namely the weak ones. This would help advisors know the capabilities of their advisees and thus, have a better decision when choosing their courses during registration. It would help instructors pay attention to these students during various in-class activities and team forming. Furthermore, special recommendations could be prescribed such as doing extra work or having office-hours visits, etc.

Engineering degrees are mostly offered in two different curriculum structures. One of them is the 150-credit Bachelor of Engineering program (BE) and the other one is the BS (107cr.)/MS (43 cr.) program. In either case, students are to fulfill strict requirements in order to graduate and hold a degree in the Engineering profession. Generally, the engineering program consists of different categories of courses to be completed by the students to fulfill the graduation requirements. Engineering students at Notre Dame University-Louaize (NDU) in Lebanon accounts for approximately 1,200 students (25% of the total number) repartitioned into three departments and four majors (Electrical, Computer and Communication, Civil and Environmental, and Mechanical). Courses are split into four categories: General Education, Core, Major, and Technical Elective courses. Currently, NDU offers the Bachelor of Engineering degree but it is also studying the prospect of launching the BS/MS program. The study undertaken in this paper applies to both cases since the main objective is to predict the performance of engineering students at an early stage of their residency for it affects various factors related to their academic path such as probation, suspension, attrition, graduation, and enrollment in further more-advanced tracks.

The purpose of displaying Fig. 4 below is to first show that a very strong correlation exists between the performance of a student in Major courses and his/her cumulative GPA.

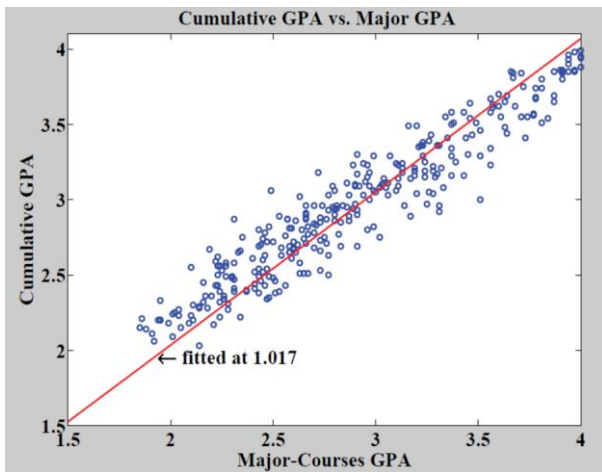


Figure 4 - Relationship between Performance in Major courses and the Cumulative GPA

Certainly, the latter data is unswervingly related to the issues stated at the beginning of this paragraph and therefore, it would be beneficial to use this indicator for it can be obtained, or as we will see later on, is related to attributes obtained at early stages of the engineering curricula.

Most of the core courses are usually taken during the first year. They comprise essentially Math, Physics, and Chemistry courses. *These courses are the prerequisites of almost all major courses* since students are exposed to the fundamental and basic concepts required to pursue specialized theories on a later stage. It is straightforward and safe to assume that if a student has weakly achieved a course, he/she will have fewer chances to *excel* or have superior performance in a higher-level, directly dependent course to which it is a prerequisite. Indeed, the material covered in the higher-level course generally relies on the one covered in the prerequisite and is sometimes a natural/further continuation and advancement in same or similar concepts and topics. Consequently, core courses convincingly play a decisive role in the students' performance in major courses.

As a result, the sought predictive tool will be based on the performance of students in the core-requirement courses, which can be tracked or depicted early. The study will subtend all types of students for the purpose of generalization: weak, average and good performers are included. Table 1 displays the distribution of students with respect to their overall performance.

#### A. Data Preparation

Five hundred Computer and Communication Engineering students' records have been gathered covering a period of almost seven years. These records consist of comprehensive transcripts with the students' grades in all courses taken throughout their academic path. These records also include the number of times students went on probation, those who have been suspended and the cumulative General Point Average (GPA) upon graduation.

The records were preprocessed and cleansed in the sense that records with non-consistent structures were eliminated if adjustment revealed not possible. In fact, over the period of

seven years, the engineering curricula have slightly changed; courses have been deleted, modified and/or replaced by new courses. Additionally, straightforward 4.0 or near 4.0 GPA have also been discarded in order to avoid misleading the prediction procedure. This significantly reduced the outlier presence and noisy data. This part of the data mining process resulted in 305 clean records with no missing data and seamless consistency amongst attributes.

Table 1 below displays the distribution of students in relationship with the cumulative GPA. As mentioned earlier, various types of learners have been enrolled in this study; weak, moderate and superior performers have been chosen as to try enhancing the generalization capacity of the model as well as its accuracy.

Table 2 shows a snapshot of the students' records and transcripts as obtained from the Registrar's Office of NDU.

#### B. Choosing The Most Pertinent Attributes Using Relief Algorithm

The core requirement pool consists of 39 credits comprising the following courses: CEN 201, ENG 201, ENG 202, MAT 211, MAT 213, MAT 215, MAT 224, MAT 235, MAT 326, MAT 335, CHM 211, PHS 212, and PHS 213. For more details regarding a description for each course, please refer to NDU's online catalog cited in [29]. The core requirement pool represents a blend of some chemistry, physics, statics, and mostly math courses, which are mandatory and fundamental. Furthermore, they introduce the students to the most important topics and concepts necessary to pursue courses in Electric Circuits and Electronics, Microprocessor Systems, Electromagnetism, Signal Processing, Communication, Programming, Database, Networking, etc.

In order to produce an effective tool, namely with such database size (not *very* large), we opted for underlying the most influential attributes prior to exercising data mining CART. This helped achieving some sort of pre-pruning before even training the Decision Tree.

Matlab *Relieff* algorithm computes the ranks and weights of attribute (predictors) for an input data matrix and response vector for classification and regression with K-Nearest neighbors. When applied to the 305-record data matrix, the importance of the Math courses outperformed the one of Physics courses and finally, ENG 201, ENG 202, CHM 211, and CEN 201 came at the bottom of the list. Particularly, the following courses were retained for our classification and regression study based on the outcome of the *Relieff* algorithm: MAT 213, MAT 224, MAT 235, MAT 335, PHS 212, and PHS 213. Furthermore, weights of the latter Math courses were very close to each other. A similar observation was

Table 1 - Students Distribution vs. GPA

Overall Performance	Number of Students	Cumulative %	Relative %
GPA $\leq$ 2.0	0	0.00%	0.00%
2.0 < GPA $\leq$ 2.7	104	34.10%	34.10%
2.7 < GPA $\leq$ 3.3	128	76.07%	41.97%
3.3 < GPA $\leq$ 4.0	73	100.00%	23.93%



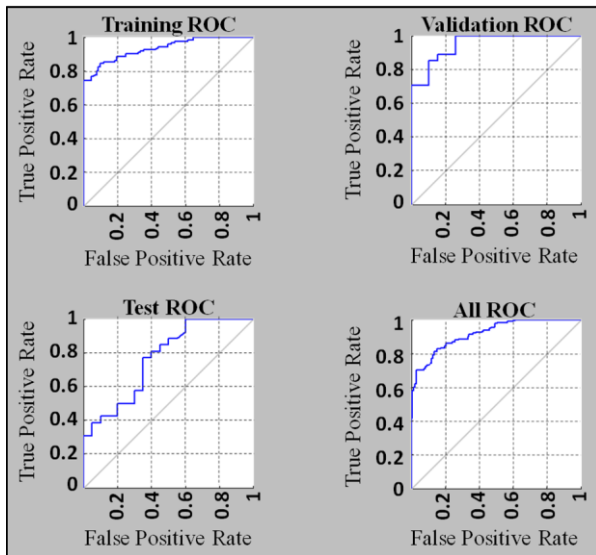


Figure 6 - ROC Curves

Fig. 7 shows how the cumulative GPA is estimated based on the performance of students in Math and Physics courses from the core-requirement pool with a best pruning level reached at 72. It is obvious that the cumulative GPA significantly worsens when students do not highly perform in these courses. Nonetheless, it is indubitably **not reliable** for accurate prediction because of both the size and type of the data. On the other hand, it constitutes another indicator that confirms our postulations and hypotheses.

To use the decision tree effectively and make use of the results from the Neural Networks Error-Histogram and ROC curves, a binary classification tree will be deployed and which will *accurately* predict the ability of students to either succeed their bachelor of engineering or enroll in Masters' programs. As shown in Fig. 8, the classification tree designates one class, "YES" or "NO", to a record based on the attributes. The former class infers achievement of a BE or enrollment in Masters, while the latter one signifies that the student would most probably fail or not be allowed to enroll in Masters, as previously detailed.

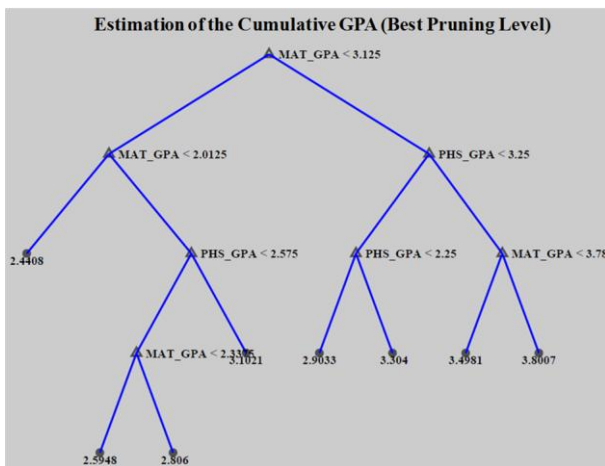


Figure 7 - Regression Tree for Estimation of the Cumulative GPA

Fig. 8 portrays two different pruning levels: Fig. 8-b exhibits the best-pruning level while Fig. 8-a exhibits a pruning level that is less by one than the best level. The purpose is to have a deeper and wider bifurcation when a student presents close attributes with respect to some deciding node values and has to be evaluated by the model.

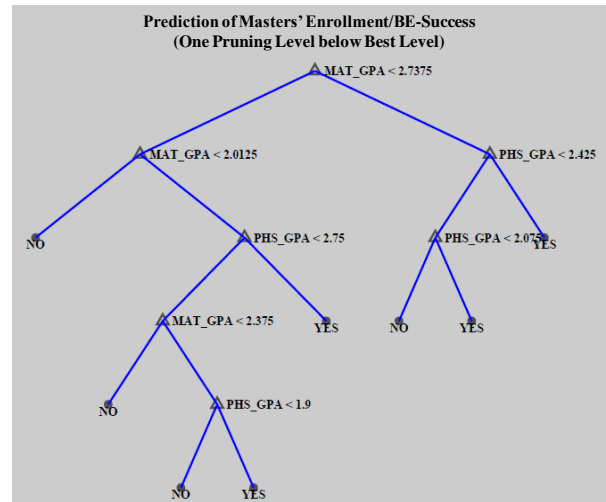


Figure 8.a - Classification Tree (Best Level minus one)

The induced if-then structure as given by Matlab is shown below. This structure is related to the decision tree obtained for the best pruning level.

```

Decision tree for classification
1 if MAT_GPA<2.7375 then node 2 else node 3
2 if MAT_GPA<2.0125 then node 4 else node 5
3 class = YES
4 class = NO
5 if PHS_GPA<2.75 then node 6 else node 7
6 class = NO
7 class = YES
    
```

After examining the classification tree, we can summarize the following results:

- If a student's GPA in Math courses is above 2.7375, then he/she is most likely to succeed and/or enroll in Masters.
- If a student's GPA in Math courses is above 2.0125, then the condition for a positive outcome is to have a GPA on the Physics courses of at least 2.75.
- Otherwise, the chances for students who have a GPA in Math courses less than 2.0125 are scarce to enroll in Masters or accomplish a Bachelor of Engineering.

#### IV. RESULTS ANALYSIS, CONCLUSION

In this paper, we carried out a study to find a reasonably accurate and reliable predictive tool that enables academicians (instructors and advisors) and administrators to decide about the enrollment of engineering students in Masters' studies or to succeed a Bachelor-of-Engineering program.

The study has been conducted in different stages and on different levels. The strong correlation between students' performance in major courses, which are usually taken during the last three years of a Bachelor-of-Engineering program or during the Masters' curriculum, and their cumulative GPA was



demonstrated. The objective of this part of the study was to enable linking core-requirement courses with the GPA and thus, allow predicting students' overall performance at an early stage of their studies.

At first, Matlab *Neural Networks Pattern Recognition* tool was applied in order to examine and decide of the accuracy of the predictive tool, which was to be eventually developed. The Error Histogram as well as the ROC curves demonstrated high level of satisfaction and reliability, namely given the size of the data.

*Preprocessing* the data was of high importance because it helped achieve a low level of outliers and present the data in a more efficient way to classification and regression trees. Particularly, the number of attributes was relatively high when compared to the number of records and thus, preliminary studies have been conducted and which efficiently, and without loss of information, reduced the size of attributes to two highly decisive ones. Matlab *Relieff* function was used to reveal the most influential attributes to be taken into account.

At the last stage, Matlab *classregtree* tool was used in two steps. The first one was to create a regression tree that estimates the cumulative GPA based on the attributes. This gave us another confirmation of the postulation we started from and then, at a later step, a binary classification tree was achieved after cross-validation and appropriate pruning. This decision tree created an if-then rule structure that enables the engineering staff to ponder over students' chances of succeeding their engineering studies.

The results revealed promising especially when discussed with Math and Physics courses' instructors and engineering advisors who all agreed on the *discovered/learned* liaison and patterns. It confirmed that students, who are weak performers particularly in this pool of courses, exhibit difficulties in most of the cases in comprehending advanced engineering concepts and in achieving high performance in major courses. Furthermore, the study gave specific numbers and thresholds in

courses taken at early stages that would alarm academicians about the situation of the concerned students.

Finally, this study will indubitably help in predicting the number of students who will reach the end of the engineering program and thus, constitutes a performant tool for decision-making and forecasting enrollment and courseware planning as well as pondering over attrition in engineering studies. It would also endow advisors and courses' instructors an anticipated estimation of their advisees and students capabilities during registration and in-class activities and a reliable platform for special recommendations.

## REFERENCES

- [1] R.S.J.d. Baker, "Data Mining for Education," In International Encyclopedia of Education, vol. 7, B. McGaw, P. Peterson, E. Baker (Eds.), 3e, Oxford, UK: Elsevier, 2010, pp. 112-118.
- [2] <http://www.educationaldatamining.org/>
- [3] J. Teresko, "Information Rich, Knowledge Poor?," Data Warehouses Transform Information into Competitive Intelligence, February 3, 1999. ([http://www.industryweek.com/articles/information\\_rich\\_knowledge\\_po\\_or\\_245.aspx](http://www.industryweek.com/articles/information_rich_knowledge_po_or_245.aspx))
- [4] C. Silverstein, S. Brin, R. Motwani, and J. Ullman, "Scalable Techniques for Mining Causal Structures," Data Mining and Knowledge Discovery, volume 4, numbers 2-3, pp. 163-192, 2000, DOI: 10.1023/A:1009891813863
- [5] P. Domingos, "Toward knowledge-rich data mining," In Proceedings of Data Min. Knowl. Discov., 2007, volume 15, issue 1, pp. 21-28, DOI: 10.1007/s10618-007-0069-7
- [6] [http://www.kdnuggets.com/data\\_mining\\_course/x1-intro-to-data-mining-notes.html](http://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html) (retrieved in September 2011)
- [7] <http://www.focus.com/fyi/10-largest-databases-in-the-world/>
- [8] <http://beyondrelational.com/justlearned/posts/290/top-10-largest-databases-in-the-world.aspx>
- [9] M. Berry and G. Linoff, "Data Mining Technique: For Marketing, Sales, and Customer Support," New York: Wiley Computer Publishing, 1997
- [10] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," In Advances In Knowledge Discovery and Data Mining, AAAI/MIT press, Cambridge mass, 1996.
- [11] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann, 2011.
- [12] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann, 2011.
- [13] Krzysztof J. Cios, Witold Pedrycz, and Roman W. Swiniarski, Data Mining: A Knowledge Discovery Approach, Springer 2007.
- [14] D. J. Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, Massachusetts Institute of Technology, 2001.
- [15] N. Delavari, S. Phon-Amnuaisuk, and M. R. Beikzadeh, "Data Mining Application in Higher Learning Institutions," Informatics in Education, vo. 7, no. 1, pp. 31-54, 2008.
- [16] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance by an Application of Data Mining Techniques," International Conference on Management and Artificial Intelligence, IACSIT Press, IPEDR vol. 6, pp. 110-114, 2011.
- [17] S. A. Kumar and M. N. Vijayalakshmi, "Efficiency of Decision Trees in Predicting Student's Academic Performance," First International Conference on Computer Science, Engineering and Applications, CS and IT 02, pp. 335-343, 2011.
- [18] C. Ho Yu, S. DiGangi, A. Jannasch-Pennell and C. Kaprolet, "A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year," Journal of Data Science, vol. 8, pp. 307-325, 2010. (neural networks, cross validation)
- [19] Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New Jersey, USA.

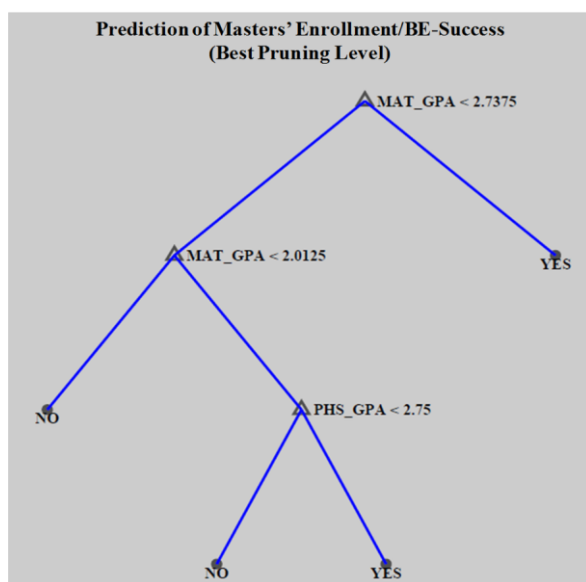


Figure 8.b - Classification Tree (Best Pruning Level)

- [20] L. Getoor, "Linking Mining: A New Data Mining Challenge," ACM SIGKDD Explorations, vol. 5, no. 1, pp. 84-89, 2003. (large data)
- [21] D. Denning, "An Intrusion Detection Model," Proceedings of the Seventh IEEE Symposium on Security and Privacy, pp. 119-131, May 1986.
- [22] M. R. Smith and T. Martinez, "Improving Classification Accuracy by Identifying and Removing Instances that Should Be Misclassified," Proceedings of International Joint Conference on Neural Networks, pp. 2690-2697, 2011.
- [23] 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, pp. 207-216, May 1993.
- [24] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in G. Piatetsky-Shapiro and W. J. Frawley, eds, Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA, 1991.
- [25] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1-13, August 2006.
- [26] J.H.Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," Journal of the American Statistical Association, vol. 48, pp. 236-244, 1963.
- [27] J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, Kluwer Academic Publishers, Boston, pp. 81-106, 1986.
- [28] <http://www.educationaldatamining.org/JEDM/>
- [29] <http://www.ndu.edu.lb/administration/registrar/catalogs.htm>
- [30] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in VLDB'94 Proceedings of the 20<sup>th</sup> International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1994. (a priori)
- [31] P. Fournier-Viger, U. Faghihi, R. Nkambou, and E. M. Nguifo, "CMRULES: An Efficient Algorithm for Mining Sequential Rules Common to Several Sequences," in Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), AAAI Publications, pp. 410-415, 2010. (association task)
- [32] A. Nandeshwar, T. Menzies, and A. Nelson, "Learning Patterns of University Student Retention," Expert Systems with Applications, vol. 38, issue 12, pp. 14984-14996, Nov.-Dec. 2011. (attrition)
- [33] R. Nock and F. Nielsen, "On Weighting Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1-13, August 2006.
- [34] M. A. Maalouf, Machine Learning and Data Mining for Computer Security: Methods and Applications, Springer-Verlag, Limited 2006. (outlier)
- [35] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in OSDI'04 Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation, USENIX Association Berkeley, CA, USA, vol. 6, 2004. (large data)
- [36] J. Luan, "An Exploratory Approach to Data Mining in Higher Education: A Primer and a Case Study," Paper presented at the AIR Forum, Seattle, Wash., 2000. (Higher Education)
- [37] R.S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, "Off-Task Behavior in the Cognitive Tutor Classroom: When Students Game The System," in Proceedings of ACM CHI: Computer-Human Interaction, pp. 383-390, 2004. (higher education)
- [38] R.S.J.d. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments," International Journal of Human-Computer Studies, vol. 68, no. 4, pp. 223-241, 2010. (higher education)
- [39] R.S.J.d. Baker, and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, issue 1, pp. 3-17, 2009.
- [40] F. Thabtah, "Pruning Techniques in Associative Classification: Survey and Comparison," Journal of Digital Information Management, vol. 4, no. 3, pp. 197-202, September 2006. (pruning)
- [41] Sarle, Warren S. (1994), "Neural Networks and Statistical Models," Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, pp 1-13.
- [42] J. Quinlan, "Simplifying Decision Trees," International Journal of Man-Machine Studies, vol. 27, no. 3, pp. 221-248, 1987. (Decision Trees)
- [43] L.-X. Zhang, J.-X. Wang, Y.-N. Zhao, and Z.-H. Yang, "A Novel Hybrid Feature Selection Algorithm: Using Relief Estimation for GA-Wrapper Search," in Proceedings of the Second International Conference on Machine Learning and Cybernetics, pp. 380-384, 2-5 November 2003.
- [44] P. Shekhawat and S. Dhande, "A Classification Technique using Associative Classification," International Journal of Computer Applications, vol. 20, no.5, pp. 20-28, April 2011. (AC)
- [45] G. P. Adanez and A. D. Velasco, "Predicting Academic Success of Engineering Students in Technical Drawing from Visualization Test Scores," Journal for Geometry and Graphics, vol. 6, no. 1, pp. 99-109, 2002. (engineering)
- [46] V.O. Oladokun, A.T. Adebajo, and O.E. Charles-Owaba, "Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course," The Pacific Journal of Science and Technology, vol. 9, no. 1, pp. 72-79, 2008. (engineering)
- [47] F. Thabtah, P. Cowling, and Y. Peng, "Multiple Label Classification Rules Approach," Journal of Knowledge and Information System, vol. 9, pp. 109-129, Springer-Verlag, 2006. (large data)
- [48] T.-M. Huang, V. Kecman, and I. Kopriya, Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning, Springer-Verlag Berlin Heidelberg, 2006. (large data + supervised)
- [49] J. Meenakumari and R. Krishnaveni, "Transforming Higher educational institution administration through ICT," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 2, no. 8, pp. 51-54, 2011. (higher education as a business)
- [50] V. Kumar and A. Chadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84, March 2011.
- [51] Y. Wang and F. Makedon, "Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification Using Microarray Data," in Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference, IEEE Computer Society Washington, DC, USA, 2004 <http://dx.doi.org/10.1109/CSB.2004.35> (relieff)
- [52] I. Kononenko, E. Simec, and M. R. Sikonja, "Overcoming the Myopia of Inductive Learning Algorithms with Relieff," Journal of Applied Intelligence, vol. 7, no. 1, pp. 39-55, 1997. (relieff)
- [53] A. H. Basha, A. Govardhan, S. V. Raju, and N. Sultana, "A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University," European Journal of Scientific Research, vo. 46, no. 2, pp. 186-193, 2010.
- [54] V. Ramesh, P. Parkavi, and P. Yasodha, "Performance Analysis of Data Mining Techniques for Placement Chance Prediction," International Journal of Scientific and Engineering Research, vol. 2, issue 8, pp. 1-6, August 2011.

#### AUTHOR'S PROFILE

Dr. Chady El Moucary graduated from the Lebanese University – Faculty of Engineering with a diploma in Electrical and Electronics Engineering. He pursued his postgraduate studies in Paris (France) with a scholarship from the French National Center for Scientific Research (CNRS) and received his PhD degree in Electrical Engineering from the University of Paris XI and the SUPELEC in 2000. Currently, Dr. El Moucary is a full-time Assistant Professor and Researcher at Notre Dame University –Louaize (NDU, Lebanon) and the Coordinator of the Faculty of Engineering in NDU's North Lebanon Campus. His research interests and publications are in Electric Machine Control, Digital Watermarking, and Data Mining. He is also a member of the Scientific Committee of the *Order of Engineers* in Tripoli (Lebanon) and a reviewer for diverse reputable worldwide conferences as well as an active member in many other educational/academic committees.

# Model of Temperature Dependence Shape of Ytterbium -doped Fiber Amplifier Operating at 915 nm Pumping Configuration

Abdel Hakeim M. Husein<sup>a\*</sup>,

<sup>a\*</sup> Department of Physics, Al-Aqsa University, P.O. Box 4051, Gaza, Gaza Strip, Palestine

Fady I. EL-Nahal<sup>b</sup>

<sup>b</sup> Department of Elec. Eng., Islamic University of Gaza, Gaza, Gaza Strip, Palestine

**Abstract**— We numerically analyze the temperature dependence of an ytterbium-doped fiber amplifier (YDFA) operating at 915 nm, investigating its gain and Noise Figure properties variation with temperature. The temperature-dependent gain and noise figure variation with YDFA length are numerically obtained for the temperature range of +20 °C to +70 °C. The results show that good intrinsic output stability against temperature change can be achieved in ytterbium doped fiber amplifiers even when operating at high gain regime with small signal input. This result demonstrates the great potential for stable high power laser communication systems based on ytterbium system.

**Keywords**- YDFA; Gain; Noise Figure.

## I. INTRODUCTION

Fiber lasers and amplifiers have attracted great interest recently, because they offer the advantages of compact size, high gain, guided mode propagation, better stability and their outstanding thermo-optical properties [1-4]. Ytterbium (Yb<sup>3+</sup>) -doped fiber Amplifier (YDFA) has a great potential because it does not have some of the drawbacks associated with erbium-doped amplifier: excited state absorption phenomenon that can reduce the pump efficiency and concentration quenching by interionic energy transfer do not occur, and high doping levels are possible. Thus, it offers high output power (or gain) with a smaller fiber length. YDFA's have a simple energy level structure and provide amplification over a broad wavelength range from 975 to 1200 nm. Moreover, YDFA's can offer high output power and excellent power conversion efficiency [1,5-12].

YDFA's have great potential in many applications, including power amplification, sensing applications, free-space laser communications, and chirped-pulse amplification of ultra-short pulses [1,12-14].

This paper explores the effects of temperature on amplifier performance. For an amplifier, the temperature affects relative extraction by the signal and ASE. Thermal management is a critical issue and cannot be ignored. Some papers about temperature effect of Er<sup>3+</sup>-doped fiber laser and amplifier were reported [15-17]. Temperature can affect the absorption and emission cross sections [2]. This paper demonstrates output characteristics of Yb<sup>3+</sup>-doped fiber laser at different

temperatures degrees. So from this research the temperature can affect the gain and noise figure (NF) at different length.

Established methods of modeling erbium amplifiers can be used to model ytterbium system [11,18-19]. However, modeling temperature sensitivity is completely different. The small energy gaps in the relevant stark levels in erbium systems, makes the determination of distinct sub-transition characteristics quite difficult [17], while this is not true in ytterbium system. Accurate characterization of Yb<sup>3+</sup> absorption and emission cross sections is crucial [20,21].

## II. THEORETICAL MODEL

We used standard rate equations for two-level systems to describe the gain and propagation characteristics of the Yb-doped fiber amplifier operating at 975 nm because the ASE power is negligible for a high power amplifier with sufficient input signal (about 1 mW). After the overlap factors are introduced and the fiber loss ignored, the simplified two-level rate equations and propagation equations are given as follows [12]:

$$\frac{dN_2(z,t)}{dt} = \frac{\Gamma_s P_s}{Ah\nu_s} [N_1 \sigma_{sa} - N_2 \sigma_{se}] + \frac{\Gamma_p P_p}{Ah\nu_p} [N_1 \sigma_{pa} - N_2 \sigma_{pe}] - \frac{N_2}{\tau} \quad (1)$$

$$N_1(z,t) - N_2(z,t) = N_0 \quad (2)$$

$$\frac{dP_s(z)}{dz} = P_s \Gamma_s [N_2(z,t) \sigma_{se} - N_1(z,t) \sigma_{sa}] \quad (3)$$

$$\frac{dP_p(z)}{dz} = P_p \Gamma_p [N_2(z,t) \sigma_{pe} - N_1(z,t) \sigma_{pa}] \quad (4)$$

Here,  $N_0$  is the Yb-dopant concentration,  $N_1$  and  $N_2$  are the ground and upper-level populations.  $\Gamma_s$  and  $\Gamma_p$  are the overlapping factor between the pump (signal) and the fiber-doped area.  $P_s(z,t)$ ,  $P_p(z,t)$  are the signal and pump power respectively.  $\sigma_{sa}$  and  $\sigma_{se}$  are the signal absorption and emission cross sections.  $\sigma_{pa}$  and  $\sigma_{pe}$  are the signal absorption and emission cross sections.  $\nu_s$  and  $\nu_p$  are the

frequencies of signal and pump light, respectively.  $A$  is the doped area of the fiber.  $\tau$  is the upper state lifetime.

Under the condition of the steady state regime, where all of the level population are time invariant i.e.,

$$\frac{dN_i(z,t)}{dt} = 0 \quad (i = 1, 2),$$

$$\frac{N_2}{\tau} = \frac{\Gamma_s P_s}{Ah\nu_s} [N_1 \sigma_{sa} - N_2 \sigma_{se}] + \frac{\Gamma_p P_p}{Ah\nu_p} [N_1 \sigma_{pa} - N_2 \sigma_{pe}] \quad (5)$$

$$N_1(z,t) = N_0(z,t) - N_2(z,t) \quad (6)$$

### A. Spectroscopy of ytterbium in silica

Ytterbium in silica is a simple, two level system having four Stark levels in the lower manifold  $F_{7/2}$  and three Stark levels in upper manifold  $F_{5/2}$ . An energy level diagram specific to Nufern 5/125 fiber is shown in Figure 1 [22]. Because the splitting of the levels depends on the glass composition, concentration of dopants and co-dopants, and the degree of structure disorder of the glass network, the energy level diagram for Yb in silica may vary with each individual fiber. The absorption and emission cross-sections for Yb in silica are related to the temperature and the energy of the levels by the following relationships:

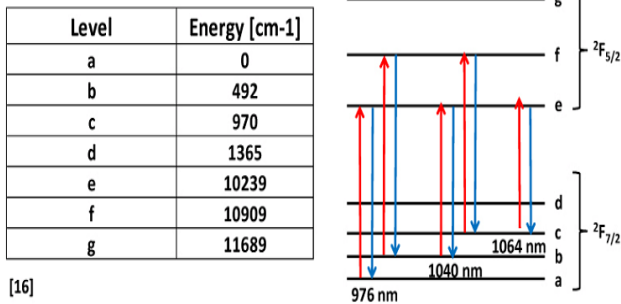


Figure 1: Energy level diagram of Yb in silica with 976 nm, 1040 nm, and 1064 nm transitions labeled [22].

$$\sigma_{sa}(\nu, T) = \sum_{x=a}^d \sum_{y=e}^g \frac{e^{-\frac{E_x}{KT}}}{\sum_{x=a}^d e^{-\frac{E_x}{KT}}} \sigma_{xy}^{sa}(\nu) \quad (7)$$

$$\sigma_{se}(\nu, T) = \sum_{x=e}^g \sum_{y=a}^d \frac{e^{-\frac{E_x}{KT}}}{\sum_{x=e}^g e^{-\frac{E_x}{KT}}} \sigma_{xy}^{se}(\nu) \quad (8)$$

where  $E_x$  are the energies of the level difference between level  $x$  and  $a$ , when  $x \in (a, d)$  where  $a$  is the ground state in lower manifold and the energies of the level difference between levels  $x$  and  $e$ , in the ground state in upper manifold when  $x \in (e, g)$ ,  $T$  is the temperature,  $K=1.3806 \times 10^{-23} J/k$  is Boltzmann's constant, and  $\sigma_{xy}^{sa}(\nu)$  and  $\sigma_{xy}^{se}(\nu)$  are the absorption and emission cross-sections of the sub-transitions [17]. To be able to calculate  $\sigma_{se}$ , it needed to use McCumber theory [23] in the Yb system, this expressed as

$$\sigma_{se}(\nu, T) = \sigma_{sa} e^{(\varepsilon - h\nu)/KT} \quad (9)$$

$$\text{where } e^{\varepsilon/KT} = e^{\frac{E_{ea}/KT}{1 + e^{-E_{fa}/KT} + e^{\frac{E_{ca}/KT}} + e^{\frac{E_{da}/KT}}}}, E_{xy} = E_x - E_y \quad (10)$$

And  $h=6.626 \times 10^{-34} J.s$  is Planck constant and  $\varepsilon$  is an active energy and given by

$$\frac{N_1}{N_2} = \exp(\varepsilon / KT) \quad (11)$$

Where  $N_1$  and  $N_2$  are Yb doping concentration at the upper and lower energy levels, respectively. Using the absorption and emission cross section from Eqs.(7-11), and the propagation equations (3) and (4) can even be solved analytically in this case [23].

The small signal gain  $G$  for active length of the fiber  $L$  can be given by

$$G(\lambda) = \exp[\Gamma_s (N_2(z,t) \sigma_{se} - N_1(z,t) \sigma_{sa}) L] \quad (12)$$

Figure 2 and 3 show the variation of Yb absorption and emission cross sections at various temperatures respectively [21]. It is clear from figure 3 that the signal absorption cross section declines with increasing wavelength and it is almost 0 at 1064 nm. The term  $N_1(z,t) \sigma_{sa}$  can be ignored as it is a decreasing function of signal wavelength, then

$$G(\text{dB}) = 10 \log_{10} \exp[\Gamma_s N_2 \sigma_{se} L] \quad (13)$$

The amplified spontaneous emission (ASE) noise spectrum uses the noise figure (NF) given or as input parameter. In the practical case the ASE is presented at input of the doped fiber, therefore the amplified input ASE  $P_{ASE}^i$  spectral density can be added to the amplified output ASE spectral density  $P_{ASE}^o$ , so

$$P_{ASE}^o = P_{amp} + P_{ASE}^i G \quad (14)$$

Where  $G$  is the gain and  $P_{amp}$  is the spectral density of ASE generated by the doped fiber. For input ASE gives the signal spontaneous beat noise ( $1/G$ ) limited noise figure as a function of the signal gain and input and output ASE spectral densities therefore the NF can be expressed as:

$$NF = \frac{1}{G} + \frac{P_{ASE}^o(\lambda_s)}{G h \nu_s} - \frac{P_{ASE}^i(\lambda_s)}{h \nu_s} \quad (15)$$

Where  $1/G$  is the beat noise,  $P_{ASE}^o(\lambda_s)$  is the output ASE spectral density (Watt/Hertz) at signal wavelength,  $P_{ASE}^i(\lambda_s)$  is the input ASE spectral density at signal wavelength, and  $\nu_s$  is the frequency of the signal wavelength. For lower gains, a much simpler model which ignores the effect of ASE on the level populations can be used, so the  $P_{ASE}^i(\lambda_s) = 0$ .

For each signal wavelength the noise figure can be calculated in decibel (dB) and is given by:

$$NF(dB) = 10 \log_{10} \left( \frac{1}{G} + \frac{P_{ASE}^o(\lambda_s)}{Gh\nu_s} \right) \quad (16)$$

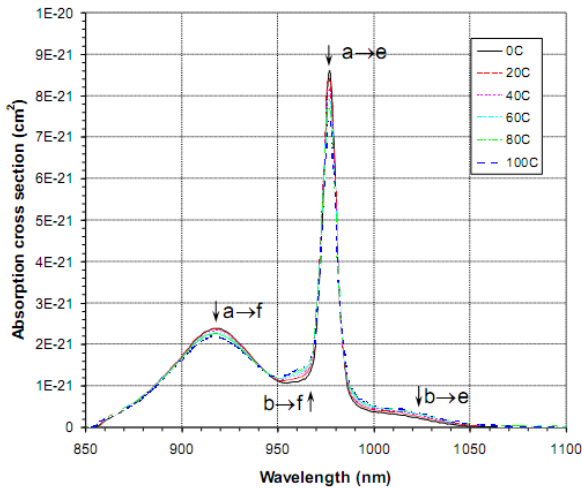


Figure 2: Yb absorption cross sections at various temperatures. [18]

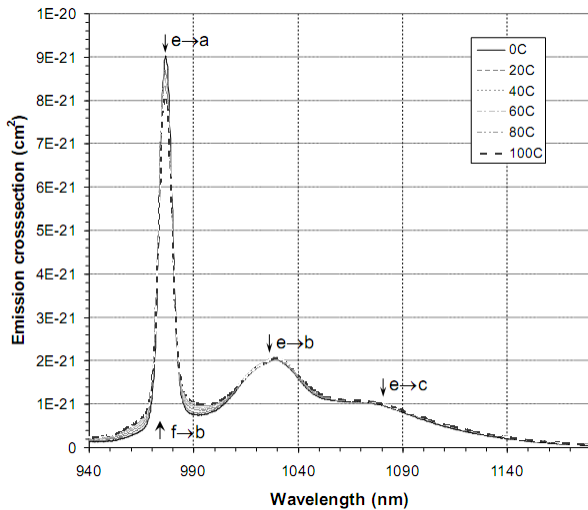


Figure 3: Yb emission cross sections at various temperatures [21].

### III. RESULTS AND DISCUSSIONS

For numerical calculation, the fiber parameters for YDFA amplifier are shown in Table 1. We studied the variation of gain and NF with the length of the amplifier over the temperature range from 20 °C to 70 °C at signal wavelength of 1064 nm. The results are shown in Figures 4 and 5, respectively. It is clear from the results that the signal gain raises with increases in the length, at the same time the gain declines when the temperature increases. However the NF increases when the temperature rises. Furthermore the NF increases with increasing the length. It is clear from the results that the variation of the NF with temperature is minimal for lengths over 5 m.

### IV. CONCLUSION

A YDFA model has been introduced including the temperature effects for gain and noise figure of a length of the YDFA amplifier. The temperature dependence of the gain and noise figure on various temperatures was taken into consideration which shows that the performance of YDFA exhibit excellent low temperature sensitivity. The analytical solution of the propagation equations has also been derived for the temperature range from 20 °C, to +70 °C for finding the gain. These results demonstrate that highly stable ytterbium doped fiber amplifiers can potentially be achieved.

Table 1: The fiber parameters and symbols used in the numerical calculations [18]

Symbol	Definitions	Value
$\sigma_{sa}$	Signal absorption cross sections	$2.3 \times 10^{-27} m^2$
$\sigma_{se}$	Signal emission cross sections	$1.09 \times 10^{-27} m^2$
$\nu_s$	Frequency of input signal	$3.279 \times 10^{14} Hz$
$\nu_p$	Frequency of pump signal	$2.819 \times 10^{14} Hz$
$\lambda_s$	Signal Wavelength	1064nm
$\lambda_p$	Pump wavelength	915nm
$\tau$	The upper state lifetime	0.84 ms
$N_0$	Yb-dopant concentration	$3.35 \times 10^{25} m^{-3}$
$A$	Doped area of the fiber	$7.8 \times 10^{-11} m^2$
$\Gamma_s$	Oveperlaping factor between the signal and the fiber-doped area	0.6
$\Gamma_p$	Oveperlaping factor between the pump and the fiber-doped area	0.01
$P_p$	Input Pump power of 20 °C and 70 °C	10 $\mu W$ to 10 W
$P_{ASE}^i (L = 0)$	Amplified input power of ASE	0
$L$	Length of the fiber	0 to 7m



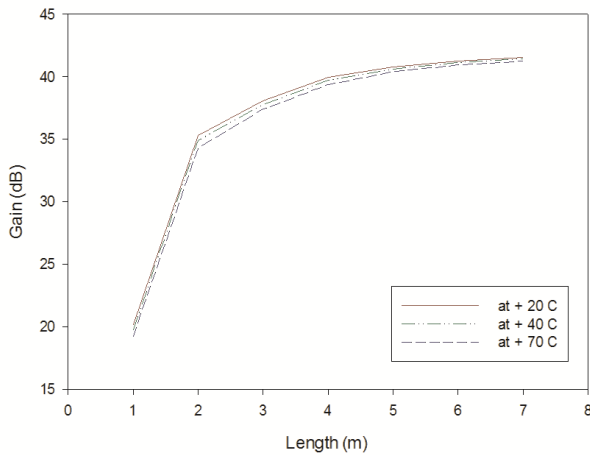


Figure 4: The change of the signal gain with temperature and length.

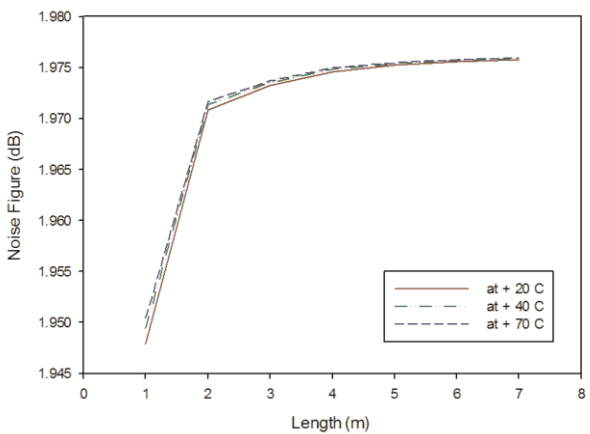


Figure 5: The change of the Noise Figure with temperature and length.

#### REFERENCES

- [1] Rüdiger Paschotta, Johan Nilsson, Anne C. Tropper, and David C. Hanna, Ytterbium-Doped Fiber Amplifiers, IEEE JOURNAL OF QUANTUM ELECTRONICS, vol. 33, no. 7, JULY 1997.
- [2] J. Chen, Z. Sui, F. Chen and J. Wang, Output characteristics of Yb<sup>3+</sup>-doped fiber laser at different temperatures, vol. 4, no. 3, Chin. Opt. Lett. (2006).
- [3] D. Xue, Q. Lou, J. Zhou, L. Kony, J. Li, Chin. Opt. Lett. 3,345 (2005).
- [4] M.J.F. Digonnet, "Rare-Earth-Doped Fiber Lasers and Amplifiers model for rare-earth-doped fiber amplifiers and lasers," CRC Press; 2nd edition (2001).
- [5] H. M. Pask, R.J. Carman, D. C. Hanna, A. C. Tropper, C. J. Mackechinc, P. R. Barber, and J. M. Dawes, IEEE Sel. Top. Quantum Electron. 1, 2 (1995).
- [6] D. C. Hanna, R. M. Percival, I. R. Perry, R. G. Smart, P. J. Suni, J. E. Townsend, and A. C. Tropper, "Continuous-wave oscillation of a monomode ytterbium-doped fiber laser," *Electron. Lett.*, vol. 24, pp.1111-1113, (1988).
- [7] D. C. Hanna, R. M. Percival, I. R. Perry, R. G. Smart, P. J. Suni, and A. C. Tropper, "An ytterbium-doped monomode fiber laser: broadband tunable operation from 1.010  $\mu$ m to 1.162  $\mu$ m and three-level operation at 974 nm," *J. Mod. Opt.*, vol. 37, pp. 517-525, (1990).
- [8] J. C. Mackechnie, W. L. Barnes, D. C. Hanna, and J. E. Townsend, "High power ytterbium (Yb<sup>3+</sup>)-doped fiber laser operating in the 1.12  $\mu$ m region," *Electron. Lett.*, vol. 29, pp. 52-53, (1993).
- [9] J. Y. Allain, M. Monerie, H. Poignant, and T. Georges, "High-efficiency ytterbium-doped fluoride fiber laser," *J. Non-Crystalline Solids*, vol. 161, pp. 270-273, (1993).
- [10] S. Magne, M. Druetta, J. P. Goure, J. C. Thevenin, P. Ferdinand, and G. Monnom, "An ytterbium-doped monomode fiber laser: amplified spontaneous emission, modeling of the gain and tunability in an external cavity," *J. Lumin.*, vol. 60, pp. 647-650, (1994).
- [11] H. M. Pask, R. J. Carman, D. C. Hanna, A. C. Tropper, C. J. Mackechnie, P. R. Barber, and J. M. Dawes, "Ytterbium-doped silica fiber lasers: versatile sources for the 1-1.2  $\mu$ m region," *IEEE J. Select. Topics Quantum Electron.*, vol. 1, pp. 2-13, (1995).
- [12] Liu Yan, Wang Chunyu, Lu Yutian, A four-passed ytterbium-doped fiber amplifier, *Optics & Laser Technology*, pp. 1111-1114 39 (2007).
- [13] R. Paschotta, D. C. Hanna, P. DeNatale, G. Modugno, M. Inguscio, and P. Laporta, "Power amplifier for 1083 nm using ytterbium doped fiber," *Opt. Commun.*, vol. 136, pp. 243-246, (1997).
- [14] V. Cautaearts, D. J. Richardson, R. Paschotta, and D. C. Hanna, "Stretched pulse Yb<sup>3+</sup>: silica fiber laser," *Opt. Lett.*, vol. 22, no. 5, pp. 316-318 (1997).
- [15] N. Kagi, A. Oyobe, and K. Nakamura, *J. Lightwave Technol.* 9, 261 (1991).
- [16] H. Tobbon, *Electron. Lett.* 29, 667 (1993).
- [17] M. Bolshtyansky, P. Wysocki and N. Conti, "Model of temperature dependence for gain shape of erbium-doped fiber amplifier", *Journal of Lightwave Technology*, vol.18, 1533-1540 (2000).
- [18] C.R. Giles, E. Desurvire, "Modeling erbium-doped fiber amplifiers," *J. Lightwave Technol.*, vol. 19, no. 7, pp. 271 - 283, (1991).
- [19] Yu. A. Varaksa, G. V. Sinitsyn, and M. A. Khodasevich, "Modeling the gain and amplified spontaneous emission spectra of erbium-doped fiber amplifiers," *J. Appl. Spectrosc.*, vol. 73, no. 2, pp. 309-312, (2006).
- [20] Xiang Peng, Joseph McLaughlin and Liang Dong, "Temperature Dependence of Ytterbium Doped Fiber Amplifiers", Conference Paper Optical Amplifiers and Their Applications (OAA) Budapest, Hungary, August 7, (2005).
- [21] Xiang Peng and Liang Dong, "Temperature dependence of ytterbium-doped fiber amplifiers", *J.Opt.Soc. Am. B*, Vol. 25, Issue 1, pp. 126-130 (2008).
- [22] Leanne J. Henry, Thomas M. Shay, Dane W. Hult and Ken B. Rowland Jr "Thermal effects in narrow linewidth single and two tone fiber lasers", vol. 19, no. 7 *Optics express* 6165 (2011)
- [23] C. Barnard, P. Myslinski, J. Chrostowski, and M. Kavehrad, "Analytical model for rare-earth-doped fiber amplifiers and lasers," *IEEE J. Quantum Electron.*, vol. 30, pp. 1817-1830, (1994).

# Healthcare Providers' Perceptions towards Health Information Applications at King Abdul-Aziz Medical City, Saudi Arabia

Abeer Al-Harbi

Health & Hospital Administration department, College of Business Administration  
King Saud University (KSU)  
Riyadh, Saudi Arabia

## Abstract-

The purpose of this study was to assess the perceptions of healthcare providers towards health information technology applications in King Abdul-Aziz Medical City in terms of benefits, barriers, and motivations.

The study population consists of all healthcare providers working at KAMC. A sample size of 623 was drawn from a population of 7493 healthcare providers using convenience random sampling method. 377 were returned, giving a response rate of 60.5%.

A self-administered questionnaire was developed based on extended literature review and comprised 25 statements on a five-point Likert-scale.

Results indicate that the majority of healthcare providers use KAMC health information applications. The majority of healthcare providers perceived that the applications are valuable and beneficial. However, healthcare providers were split over the barriers to HIT use in KAMC. As for drivers, healthcare providers generally would be motivated to use the IT applications by provision of new applications and training, contribution in change hospital's work procedures, and provision of technical support. Also, there were many barriers identified by healthcare providers. These were insufficient number of computers, frequent system down, and the use of computerized systems is time consuming. Finally, there were significant differences in the perceptions with respect to gender, occupation, and training.

**Keywords-** *Healthcare providers; Health Information Technology; Computerized Patient Record; King Abdul-Aziz Medical City.*

## I. INTRODUCTION

### a. Health Information Technology (HIT):

Healthcare information technology (HIT) has become a key preoccupation of healthcare systems worldwide [1, 2]. A review of the literature reveals that there is significant consensus that the implementation of electronic health records (EHRs) and HIT systems is considered among the highest priorities of modern healthcare systems [3].

Clinical practices rely heavily on the collection and analysis of medical data for decision-making abilities when caring for patients [4]. Thus, health information systems are capable of having a significant, positive impact on patient care within healthcare settings [5].

Health information technology is in general increasingly viewed as the most promising tool for improving the overall quality, safety and efficiency of the health delivery system [6,7, 8]. The institute of Medicine (IOM) identified information technology as one the critical forces that could significantly improve healthcare quality and safety [9].

One of the most challenging areas of health information technology is integrating it into the workflow of the healthcare providers [10]. Despite the increasing availability of health information technology applications, anecdotal evidence suggests that its use has not been well accepted by healthcare providers [11,12]. Acceptance of information technologies has occupied a central role in information technology research. There have been many studies investigating IT acceptance in different settings at both individual and organizational levels of analysis and different theoretical models have been used [13,14]. The literature provides evidence of failed clinical system implementations, due to lack of adoption by users [15]. However, with few significant exceptions, information system research is scarce regarding information technology acceptance in a healthcare environment [16,17,18,19].

In Saudi Arabia, the government strives to improve quality and safety of healthcare services through adoption health information technology [20]. However, most Saudi health organizations have no electronic health records (HER) systems implemented in their facilities, and they are totally dependent either on manual paper work or on very basic software tools to do their day to day tasks such as patient admissions [21]. KAMC is one of the few hospitals that have a basic EHR system which was later replaced by a computerized Patient Record (CPR) system. CPR is a single integrated system with a comprehensive suite of modules that provides depth and breadth of patient-care support and workflow management. CPR system streamlines administrative functions and eliminates paperwork to get caregivers back in the business of quality patient care. CPR system provides for an array of technological imperatives, including Computerized Physician Order Entry (CPOE), Clinical Decision Support (CDS), automated nursing documentation, integrated pharmacy and automated medication administration. A CPOE system, for example, makes prescription orders legible, identifies the correct medication and dose as well as signals alerts for



potential medication interactions or allergic reactions [20]. According to Dr. David Brailer, cited in Harrison and Daly [22], CPOE reduces medication errors by 20 percent .

Despite the importance of HIT in improving healthcare efficiency, there were few studies carried on use, barriers and drivers to HIT in Saudi health organizations. Therefore, there is a need for investigating the perceptions of healthcare providers towards the health information technology applications. This research is an attempt to understand the perceptions of healthcare providers towards health information technology applications in King Abdul-Aziz Medical City in terms of benefits, barriers, and motivation toward the use of health information applications. In addition, the research will investigate the effect of demographic and organizational variables on the perceptions of the healthcare providers towards the health information technology applications.

*b. King Abdul-Aziz Medical City (KAMC):*

King Abdul-Aziz Medical City commenced its operations in 1983 in Riyadh under National Guard Health Affairs (NGHA). NGHA has passed the requirements for accreditation under the (JCI) Joint Commission International standards with excellent performance in December 2009. The total bed capacity of the hospital is 847 beds. The average length of stay is 4.6 days, and the average number of outpatient visits per day is 3,145 patients. Total number of physicians is 1564, total number of nurses is 3921, and the total number of clinical/paramedical staff is 2008.

II. METHODS

*a. Survey Instrument:*

In this study, quantitative research method approach was used. To collect the data, a questionnaire form was designed to achieve the research objectives. Based on extended literature review, appropriate research constructs which had been validated in prior studies were developed. These include benefits, barriers, and motivation to use health information systems. In addition, the questionnaire included a section of general information about the respondents' demographics and organizational variables which were considered as moderators to the perceptions towards the health information applications. The second section included 25 statements regarding the benefits, barriers, and motivation of the health information applications using five-point Likert-scale (1=Strongly Disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree). Thereafter, the questionnaire was validated through evaluation by two faculty members of King Saud University, and a pilot study. Cronbach's alpha values for the three dimensions (benefits, barriers, and drivers) were strictly above 0.74; meeting the recommended alpha threshold values of at least 0.7 [23;24]. Therefore, all the three dimensions were internally consistent.

*b. Population and Sample*

The study population consists of all healthcare providers working at KAMC. The healthcare providers include physicians, nurses, and clinical/paramedical personnel. The population size is 7493. A sample size of 623 was drawn from the population using convenience random sampling method. Convenience sampling is a non-probability sampling technique

where subjects are selected because of their convenient accessibility and accessibility to the researcher [25]. The questionnaires with cover letters that explained the purpose of the study were distributed during April/May 2011. Of 623 questionnaires distributed, 377 questionnaires were returned, giving a response rate of 60.5 percent.

*c. Analysis:*

Descriptive statistics were used to analyze the demographic and organization variables and the respondents' perceptions towards benefits, barriers, and motivation to use health information systems. One-sample t-test was conducted to determine whether the mean score of each item of the three dimensions (benefits, barriers, motives) is significantly higher than a score 3; this being the mid-point on the Likert scale for "Neither agree nor disagree" response to the item. Two-sample t-test was used to test whether there are differences in respondents' perceptions towards IT benefits, barriers to using IT, and motives to using IT variables with respect to gender. One-way analysis of variance (ANOVA) was used to test whether there are difference in respondents' perceptions towards IT benefits, barriers to using IT, and motives to using IT variables with respect to education and occupation.

III. RESULTS:

*a. Respondents' Characteristics:*

Table I shows the profile of respondents by age, gender, education, work experience, and occupation. The average respondent's age was 36.2 years associated with a relatively high standard deviation of 9.6 years. This shows the medical workforce at KAMC is young. With respect to gender, the vast majority of the respondents were female, 86.2 percent, while the remaining 13.8 percent were males. The sex disproportion is due to the fact that most of the sample were nurses, 55.2 percent; as nurses are usually females.

In terms of educational background, most of the respondents, 79.3 percent, hold bachelor's degree, followed by 17.2 percent who had postgraduate degree and 3.4 percent had high school education or less.

The work experience of the respondents ranged from one year to twenty-six years. About 43 percent of the respondents had less than five years of work experience; followed by 27.3 percent who had between five to nine years, 20.2 percent between ten to fourteen years, 5.6 percent between fifteen to nineteen years, and lastly 4.2 per cent had work experience of more than twenty years. The average working experience was 7.1 years with a relatively high standard deviation of 5.5 years.

As for occupation, the table shows that about two thirds of the sample were nurses, followed by 17.2 percent were physicians and the same percentage were other medical staff. The results show that the nurse-physician ratio is relatively high; 3.8 nurses per physicians in the sample compared to 2.5 for the Ministry of Health [26].

TABLE I. PROFILE OF RESPONDENTS

Variables	Frequency	Percent
Age (Years):		

Variables	Frequency	Percent
20 - 29	107	28.4
30 - 39	157	41.6
40 - 49	63	16.7
50 and above	50	13.2
<b>(Mean =36.2, Std deviation= 9.6)</b>		
<b>Gender:</b>		
Male	52	13.8
Female	325	86.2
<b>Education:</b>		
High school or less	13	3.4
Bachelors' degree	299	79.3
Postgraduate degree	65	17.2
<b>Experience (years):</b>		
< 5	161	42.7
5 - 9	103	27.3
10 - 14	76	20.2
15 - 19	21	5.6
20 and above	16	4.2
<b>(Mean =7.1, Std deviation=5.5)</b>		
<b>Profession:</b>		
Physician	65	17.2
Nurse	247	65.5
Other	65	17.2

*b. Literacy and Use of Information Technology*

Table II presents the levels of IT applications' knowledge and training and frequency of IT use. The results show that about two-thirds of the respondents attended training courses in information applications, while the remaining respondents, 34.5 percent, stated that they didn't attend any training course in this field. With regards to information technology literacy, 82.8 percent of the respondents stated that they had good knowledge and skills in the use of information applications, whereas 17.2 percent had poor skills in the use of information applications. However, the results show that most respondents who had training in IT had also good IT applications skills. The Chi-squared test confirmed there were significant relationship between training and IT knowledge at 0.01 level of significance. As can be seen from the table, 62.1 percent had training in IT field had also good IT skills compared to 20.7 percent of the respondents who had no training and had good IT knowledge. These results indicated that training has positive effect on health providers' IT knowledge and skills.

Table II shows also that, the vast majority of the respondents, 86.2 percent, reported that they always use systems' applications, while 10.3 percent stated that they

sometimes systems' applications. The remaining 3.4 percent of the respondents expressed that they rarely use systems' applications in KAMC. However, the results show there was significant relationship between frequency of systems' application use and IT knowledge at 0.01 level of significance. The results show that most respondents with good IT applications skills use always KAMC computerized systems.

TABLE II. RELATIONSHIPS BETWEEN RESPONDENTS' IT KNOWLEDGE WITH TRAINING AND FREQUENCY IT USE

	Knowledge of IT applications		Total	Chi-square Test
	Good	Poor		
<b>Training:</b>				
Had training in IT	234	13	247	
	62.1%	3.4%	65.5%	
Did not have training in IT	78	52	130	
	20.7%	13.8%	34.5%	
All respondents	312	65	377	
	82.8%	17.2%	100%	$\chi^2 = 75.7;$ (sig.=0.000)
<b>Frequency of IT use:</b>				
Always	286	39	325	
	75.9%	10.3%	86.2%	
sometimes	13	26	39	
	3.4%	6.9%	10.3%	
Rarely	13	0	13	
	3.4%	0.0%	3.4%	
All respondents	312	65	377	
	82.8%	17.2%	100%	$\chi^2 = 72.0;$ (sig.=0.000)

*c. Perceptions of healthcare providers towards the benefits, barriers, and motives to use information technology applications in KAMC*

Table III presents the perceptions of healthcare providers towards benefits, barriers, and motives to use IT applications. The high mean scores of the respondents' responses on benefits of IT applications, ranged between 3.6 to 4.4, reveal that all the respondents perceive that the information technology applications in KAMC are valuable. Therefore, healthcare providers believe that all information technology applications are important and beneficial to both patients and KAMC

With regard to barriers to IT use, the mean scores of the respondents' responses ranged between 2.6 to 3.4. This explains that the respondents were split over the barriers to IT use in KAMC. The results of the t-test show that the following represent barriers to IT use in KAMC (items with p-values less than or equal 0.05):

- Insufficient number of computers
- Time consuming
- Low system performance
- System being down frequently

The results of the t-test show that the following do not represent barriers to IT use in KAMC (items with p-values greater than 0.05):

- Lack of training for the hospital staff
- Lack of technical support
- Incapability of the system
- Lack of management support

As for drivers to IT use, the respondents' mean score on items measuring the motives of IT use ranged from 3.58 to 3.89, implying the respondents agree with four motives shown in the table. Therefore, it can be concluded that healthcare providers generally would be motivated to use IT applications in KAMC by provision of new applications and training, contribution in change hospital's work procedures, and provision of technical support.

d. The effect of gender, occupation, and training on respondents' perceptions towards IT benefits, barriers to using IT, and motives to using IT variables:

1) Gender:

Two-sample t-test was used to test whether there are differences in respondents' perceptions towards IT benefits, barriers to using IT, and motives to using IT variables with respect to gender. As for benefits of IT, Table IV shows that there were significant differences in respondents' perceptions on items 2, 4, 6, 7, 8, 9, and 13 with respect to gender at 0.05 significance level. The mean score of these items by gender show that the female respondents rated items "provides speed to accomplish work", "easier to find investigation results", "facilitates coordination among departments", and "improves quality of patients' care" significantly higher than did male respondents. Whereas, male respondents were more likely to agree on items "prevent loss of patients' data", "helps in preparing hospital reports", and "improves decisions making process" compared to female respondents.

TABLE III. RESPONDENTS' PERCEPTIONS TOWARDS BENEFITS, BARRIERS, AND MOTIVES TO USE INFORMATION TECHNOLOGY APPLICATIONS IN KAMC (N=377)

Item	Mean score	Std deviation	t-value	P-value
<b>Benefits of IT:</b>				
1. Easier to access patient records	4.4	0.61	43.3	0.000
2. Easier to find investigation results	4.4	0.62	44.9	0.000

Item	Mean score	Std deviation	t-value	P-value
3. Prevent loss of patients' data	4.3	0.71	36.7	0.000
4. Helps in preparing hospital reports	4.3	0.81	30.8	0.000
5. Helps in managing patients	4.2	0.81	28.5	0.000
6. Provides speed to accomplish work	4.1	0.77	28.3	0.000
7. Saving paper work	4.0	1.01	19.4	0.000
8. Facilitates coordination among departments	4.0	0.99	18.9	0.000
9. Improves decisions making process	4.0	0.86	21.8	0.000
10. Ensures patients' privacy	4.0	1.00	19.7	0.000
11. Reduces medical errors	3.9	0.70	24.1	0.000
12. Improves quality of patients' care	3.9	0.91	19.3	0.000
13. Decreases work load	3.6	1.27	8.5	0.000
<b>Barriers to IT use:</b>				
1. Time consuming	3.4	1.18	5.9	0.000
2. Insufficient number of computers	3.2	1.26	3.0	0.001
3. System being down frequently	3.1	1.12	2.5	0.007
4. Low system performance	3.1	1.08	1.6	0.053
5. Lack of training for the hospital staff	2.9	1.07	-1.2	0.880
6. Lack of technical support	2.7	1.04	-5.9	1.000
7. Incapability of the system	2.6	0.87	-9.9	1.000
8. Lack of management support	2.6	0.98	-7.9	1.000
<b>Motives to IT use:</b>				
1. Provide new / durable applications	3.8	0.65	23.3	0.000
2. Provide training to staff	3.8	0.79	20.0	0.000
3. Change hospital's work procedures	3.6	0.87	13.1	0.000
4. Provide technical support	3.9	0.62	28.1	0.000

The two-sample test's results show there were no significant differences between male and female respondents in their perceptions towards items 1, 3, 5, 10, 11, and 12 at 0.05 level of significance.

With respect to barriers to IT use, the results of two-sample t-test show that there were significant differences in perceptions of respondents on all items except for items 5, and 6 by gender. Male respondents indicated a higher agreement with the first two barriers (insufficient number of computers and time consuming) than did female respondents. While, females were likely to agree on four barriers, low system performance, system being down frequently, incapability of the system, and lack of management support, than male respondents. Furthermore, the results show there were no significant differences in the perceptions of male and female

respondents towards "lack of training for the hospital staff" and "lack of technical support" at 0.05 level of significance.

As for motives, the results show there were significant differences between male and female respondents in their attitudes towards the item which states "IT provides motives new / durable applications" at 0.05 level of significance. Female respondents indicated a higher agreement with the statement compared with male respondents. Whereas, the results show no other significant perceptions differences between male and female respondents on the remaining items, "provide training to staff", "change hospital's work procedures" and "provide technical support", at 0.05 level of significance. This means all health providers, regardless their gender, agreed that these three items represent motives to IT applications in KAMC

1.	Provide new / durable applications	3.5	3.8	-3.2	0.002
2.	Provide training to staff	3.8	3.8	-0.3	0.800
3.	Change hospital's work procedures	3.8	3.6	1.7	0.096
4.	Provide technical support	4.0	3.9	1.9	0.063

TABLE IV. RESULTS OF TWO-SAMPLE T-TEST OF IT BENEFITS, BARRIERS, AND MOTIVES WITH RESPECT TO GENDER

Item	Mean Score		Two-sample t-test	
	Male	Female	t-value	P-value
<b>Benefits:</b>				
1. Easier to access patient records	4.3	4.4	-1.0	0.300
2. Provides speed to accomplish work	3.8	4.2	-3.4	0.001
3. Saving paper work	3.8	4.0	-1.8	0.068
4. Easier to find investigation results	4.3	4.5	-2.1	0.041
5. Helps in managing patients	4.3	4.2	1.0	0.310
6. Facilitates coordination among departments	3.3	4.1	-5.5	0.000
7. Prevent loss of patients` data	4.5	4.3	2.0	0.049
8. Helps in preparing hospital reports	4.5	4.2	2.1	0.032
9. Improves decisions making process	4.3	3.9	2.6	0.011
10. Reduces medical errors	3.8	3.9	-0.7	0.474
11. Ensures patients` privacy	4.0	4.0	0.2	0.812
12. Decreases work load	3.8	3.5	1.3	0.196
13. Improves quality of patients` care	3.5	4.0	-3.3	0.001
<b>Barriers:</b>				
c. Insufficient number of computers	3.8	3.1	3.6	0.000
d. Time consuming	3.8	3.3	2.7	0.007
e. Low system performance	2.5	3.2	-4.5	0.000
f. System being down frequently	2.3	3.3	-6.5	0.000
g. Lack of training for the hospital staff	3.0	2.9	0.6	0.548
h. Lack of technical support	2.5	2.7	-1.4	0.170
i. Incapability of the system	2.3	2.6	-2.7	0.006
j. Lack of management support	2.0	2.7	-4.9	0.000
<b>Motives:</b>				

2) Occupation:

One-way Analysis of Variance (ANOVA) was used to determine whether there was a significant mean difference in the respondent's perceptions on benefits of IT applications, barriers to using IT applications, and motives to use IT applications with respect to occupation. Table V demonstrates the results of ANOVA test of IT benefits with occupation.

As for benefits, the results of the ANOVA tests showed that there were significant differences between physicians, nurses and other staff in their perceptions of all items measuring the benefits of IT applications at 0.05 significance level. Nurses had a higher positive perception than physicians and other staff in the following items:

- Provides speed to accomplish work.
- Easier to find investigation results.
- Helps in preparing hospital reports.
- Improves decisions making process.
- Decreases work load.

Whereas, other staff indicated higher agreement with the following statements when compared to physicians and nurse:

- Easier to access patient records
- Saving paper work
- Helps in managing patients
- Facilitates coordination among departments
- Prevent loss of patients` data
- Reduces medical errors
- Ensures patients` privacy
- Improves quality of patients` care

Interestingly, the physician's respondents indicated the lowest agreement with all statements that measure the benefits of IT applications in KAMC. These results are sensible since all these benefits affect the performance of nurses and other staff more than physicians. For example, from physician's point of view IT does not decrease their workload, which is why it was rated by them as low as 1.6 compared to 4 and 3.8 for nurses and other staff respectively.

As for barriers, the results of the ANOVA tests showed that there were significant differences between physicians, nurses and other staff in their perceptions of all items measuring barriers to use IT applications at 0.05 significance level; all p-values were strictly less than 0.02. Physicians indicated higher agreement with the following seven statements:

- Insufficient number of computers
- Time consuming
- Low system performance
- System being down frequently
- Lack of training for the hospital staff
- Lack of technical support
- Incapability of the system
- Lack of management support

Whereas, the results show that the higher mean score of respondents responses to item "System being down frequently" was for other staff, followed by nurses and physicians. It is also noted that nurse respondents were less likely to agree with these stated barriers.

Table V presents the results of ANOVA test of motives IT applications use with occupation as the factor.

Regarding motives, the ANOVA results showed that there were significant differences between physicians, nurses and other staff in their perceptions of all items measuring motives to IT use at 0.01 significance level. Other staff respondents indicated higher agreement with the statements "Provide new/durable applications ", "Provide training to staff", and "Provide technical support" compared to physicians and nurses:

TABLE V. ONE-WAY ANALYSIS OF VARIANCE TEST OF IT BENEFITS AND OCCUPATION

Item	Mean score				ANOVA	
	Physician	Nurse	Other	Total	F	Sig.
<b>Benefits:</b>						
Easier to access patient records	4.2	4.3	4.6	4.4	7.3	0.001
Provides speed to accomplish work	3.6	4.2	4.2	4.1	17.3	0.000
Saving paper work	2.6	4.3	4.4	4.0	121.5	0.000
Easier to find investigation	4.2	4.5	4.5	4.4	4.7	0.010
Helps in managing	4.0	4.2	4.4	4.2	4.2	0.015
Facilitates coordination	3.0	4.1	4.2	4.0	48.1	0.000
Prevent loss of patients` data	4.0	4.4	4.6	4.3	12.6	0.000
Helps in preparing	4.0	4.4	4.0	4.3	10.5	0.000
Improves decisions making	3.2	4.2	4.0	4.0	37.9	0.000
Reduces medical errors	3.6	3.9	4.0	3.9	5.1	0.007
Ensures patients` privacy	3.6	4.0	4.4	4.0	10.2	0.000
Decreases work load	1.6	4.0	3.8	3.6	178.3	0.000

Item	Mean score				ANOVA	
	Physician	Nurse	Other	Total	F	Sig.
Improves quality of	3.0	4.0	4.6	3.9	72.4	0.000
<b>Barriers:</b>						
Insufficient number of computers	3.6	3.1	3.2	3.2	4.3	0.015
Time consuming	4.4	3.1	3.4	3.4	39.7	0.000
Low system performance	3.8	2.9	3.2	3.1	21.6	0.000
System being down frequently	2.8	3.1	3.8	3.1	17.5	0.000
Lack of training for the hospital	3.4	2.9	2.6	2.9	10.6	0.000
Lack of technical support	3.4	2.5	2.6	2.7	23.0	0.000
Incapability of the system	3.0	2.4	2.6	2.6	12.6	0.000
Lack of management support	3.0	2.6	2.4	2.6	6.8	0.001
<b>Motives:</b>						
Provide new/durable applications	3.0	3.9	4.0	3.8	81.2	0.000
Provide training to staff	2.8	4.0	4.2	3.8	109.2	0.000
Change hospital`s work	3.2	3.7	3.4	3.6	13.6	0.000
Provide technical support	3.4	3.9	4.2	3.9	33.3	0.000

### 3) Training:

A two-sample t-test was performed to test whether there were differences in respondents' perceptions towards IT benefits, barriers and motives to using IT with respect to training (Table VI). As for IT benefits, the results show that there were significant differences (p-value < 0.05) in perceptions of respondents who had training in IT and those who had no training on all items except items "Ensures patients' privacy" and "Improves quality of patients' care". It is worth noting that the mean scores of the respondents who had training on these items were higher than the mean scores of the respondents who did not attend training in IT field. This shows that the staff who attended training courses in IT perceive the benefits of IT more than those who did not attend training courses in this field. As shown in the table, the two-sample test's results show that there were no significant differences in the perceptions of staff who had training and those did not attend training on items: "Ensures patients' privacy" and "Improves quality of patients' care"

As for barriers, the results show there were significant differences between respondents who attended training in IT and those who did not attend training in their perceptions

towards barriers to using health information applications in KAMC in the following items:

- Time consuming
- Low system performance
- Lack of training for the hospital staff
- Lack of technical support

The mean scores of the respondents who did not attend training on IT on these items were higher than the mean scores of the respondents who had training except for the item "low system performance". This indicates that the staff who attended training courses in IT perceive less obstacles to IT use in KAMC compared to staff who had no training in IT. Conversely, the respondents who had training in IT perceive that the system performance was low more than those who had no training in IT. Moreover, the results show there were no significant differences between respondents who had training in IT and those who had no training in their perceptions towards barriers to using health information applications in KAMC in the following items:

- Insufficient number of computers
- System being down frequently
- Incapability of the system
- Lack of management support

With regards to drivers of IT use, the results show there were significant differences between respondents who attended training in IT and those who did not attend training in their perceptions towards two items; "IT provides new/durable applications" and "IT provides technical support at 0.05 level of significance. The mean scores of the respondents who did not attend training on IT on these items were higher than the mean scores of the respondents who had training except for the item "Provide new / durable applications". This reveals that the staff who attended training courses in IT perceive less motivation to IT use in KAMC compared to those who had no training in IT. On the contrary, the respondents who had training in IT perceive that IT provides technical support more than those who had no training in IT. Furthermore, there were no significant differences between respondents who had training in IT and those who had no training in their perceptions towards two items; "IT provides training to staff" and "IT changes hospital's work procedures".

TABLE VI. RESULTS OF TWO-SAMPLE T-TEST OF IT BENEFITS, BARRIERS, AND MOTIVES WITH RESPECT TO ATTENDANCE OF TRAINING IN IT

Items	Mean Score		Two-sample t-test	
	Attended	Not attended	t-value	P-value
<b>Benefits:</b>				
Easier to access patient records	4.5	4.0	8.0	0.000
Provides speed to accomplish work	4.2	3.9	3.8	0.000
Saving paper work	4.2	3.7	3.9	0.000

Items	Mean Score		Two-sample t-test	
	Attended	Not attended	t-value	P-value
Easier to find investigation results	4.5	4.3	3.2	0.001
Helps in managing patients	4.3	3.9	4.5	0.000
Facilitates coordination among departments	4.1	3.8	2.4	0.016
Prevent loss of patients` data	4.4	4.2	3.4	0.001
Helps in preparing hospital reports	4.4	4.1	2.7	0.006
Improves decisions making process	4.2	3.6	6.2	0.000
Reduces medical errors	4.0	3.6	5.9	0.000
Ensures patients` privacy	4.0	4.0	-0.3	0.747
Decreases work load	3.7	3.3	3.1	0.002
Improves quality of patients` care	4.0	3.8	1.5	0.134
<b>Barriers:</b>				
Insufficient number of computers	3.2	3.2	0.0	0.965
Time consuming	3.2	3.7	-4.2	0.000
Low system performance	3.2	2.9	2.6	0.010
System being down frequently	3.1	3.2	-0.9	0.365
Lack of training for the hospital staff	2.8	3.2	-3.6	0.000
Lack of technical support	2.5	3.0	-4.6	0.000
Incapability of the system	2.6	2.5	0.8	0.451
Lack of management support	2.6	2.7	-1.3	0.197
<b>Motives:</b>				
Provide new / durable applications	3.7	3.9	-2.9	0.004
Provide training to staff	3.9	3.7	1.9	0.057
Change hospital`s work procedures	3.6	3.6	-0.3	0.789
Provide technical support	4.0	3.7	3.7	0.000

#### IV. DISCUSSION

The results show that the majority of healthcare providers use KAMC health information systems when the survey was conducted. This result somewhat conflicts with Anazy [19] who found about 26 percent of healthcare providers use electronic health records in six hospitals in Riyadh. Despite the high HIT use, KAMC healthcare providers with good IT skills used KAMC computerized systems more than those with poor skills. This finding is consistent with that of Alam and Noor [27, 28] who found significant effects of IT skills on adoption of ICT.

The high mean scores of the respondents' responses on benefits of HIT applications reveal that healthcare providers perceive that the information technology applications in KAMC are valuable and beneficial to both patients and



KAMC. This is consistent with findings of many researches carried in USA which found that the healthcare providers perceive the benefits of HIT in improving healthcare [29,30]. With regard to barriers, the healthcare providers were split over the barriers to HIT use in KAMC. The healthcare providers agreed that insufficient number of computers, time consuming, low system performance, the system being down frequently as barriers to HIT use in KAMC. Whereas, they didn't perceive that lack of training for the hospital staff, lack of technical support, incapability of the system, and lack of management support were barriers to HIT use. These results are somewhat consistent with Houser and Johnson [29]. As for drivers, the results showed that healthcare providers generally would be motivated to use IT applications in KAMC by provision of new applications and training, contribution in change hospital's work procedures, and provision of technical support.

The results showed the perceptions of healthcare providers on benefits, barriers and motives were influenced by gender. However, the gender effect on perceptions of healthcare providers is not consistent, as some items of the three dimensions (benefits, barriers, and motives) were higher by males and others were rated higher by females. However, there were no significant differences in perceptions of some items between male and female health providers. These results are to some extent consistent with other research findings [31,32].

With respect to the effect occupation, the results show that there were significant differences between physicians, nurses and other staff in their perceptions towards all items measuring benefits, barriers, and motives. However, the effect of occupation is also inconsistent; as some healthcare providers had a higher positive perceptions than others. These results conform with those of other research findings [33].

As regards the effect of training, the results show that healthcare providers who attended training courses in IT perceive the benefits of HIT more than those who did not attend any training courses in this field. Similarly, the results indicate that healthcare providers who attended training courses in IT perceive less barriers to HIT use in KAMC compared to those who had no training in IT. As for drivers of IT use, the results show that the effect of training on motives to HIT use were inconsistent as there were significant differences between healthcare providers who attended training in IT and those who did not attend training in their perceptions towards some items. These results are consistent with previous research findings which acknowledged the positive impact of training on IT adoption [32,34].

The major research limitation of this study was the use of convenience sample for data collection which might not represented the target population accurately. Despite this limitation and due to the lack of research in this area, the study provides important information on the perceptions of healthcare providers towards benefits, barriers and drivers of health information technology in KAMC.

## V. CONCLUSION , RECOMMENDATIONS AND FUTURE SCOPE

The purpose of this study was to assess the perceptions of healthcare providers towards health information technology applications in King Abdul-Aziz Medical City in terms of benefits, barriers, and motives to use these applications. This study also contributes in investigating the effects of gender, occupation, and training on the perceptions of the healthcare providers towards the health information applications.

Despite the perceived benefits and motives of health information technology use, there were many barriers identified by healthcare providers. The barriers include insufficient number of computers, frequent system down, and that using computerized systems is time consuming. Furthermore, there were significant differences in the perceptions of healthcare providers towards benefits, barriers, and motives to health information technology with respect gender, occupation, and training. Based on these results, the study recommends that KAMC to provide easy access to health information applications, continuous training to all healthcare providers on health information technology, technical support services and change hospital's work procedures. Further, the study also recommends that KAMC administration to engage healthcare providers in planning and promotion of health information applications.

As a future scope, more research on the adoption of health information technology applications can be carried out. The scope can also be widened by considering the effect of additional demographic and organizational variables on the adoption HIT. Moreover, similar research can be carried in other KAMC braches to trace geographic variations in HIT adoption.

## REFERENCES

- [1] Kaye R, Kokia E, Shalev V, darD I, Chinitz D, "Barriers and success factors in health information technology: A practitioner's perspective", *Journal of Management & Marketing in Healthcare*. 2010; 3 (2): 163-175.
- [2] Jones, TM. National Infrastructure for eHealth: Considerations for Decision Support. *Studies in. Health Technology and Informatics*, 2004; 100: 28-34
- [3] Castells, M., Lupiáñez, F., Saigí, F. and Sánchez, J. E-Health and Society: An Empirical Study of Catalonia — Summary of the Final Research Report, Catalan Internet Project, UOC and Generalitat de Catalunya, Barcelona. 2007
- [4] Hersh WR. *Medical Informatics, "Improving Healthcare through Information"*. *Journal of the American Medical Association*. 2002; 288(16):1955-1958.
- [5] Burton LC, Anderson GF, Kues IW. "Using Health Records to Help Coordinate Care". *The Milbank Quarterly*. 2004; 82(3):457-481.
- [6] Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton S, Shekelle P. Systematic review: "Impact of health information technology on quality, efficiency, and costs of medical care", *Annals of Internal Medicine*. 2006; 144 (10):742-752.
- [7] McCullough J, Casey M, Moscovice I, Prasad S. "The Effect of Health Information Technology on Quality in US Hospitals". *Health Affairs*. 2010; 29 (4):647-654.
- [8] Hillestad R, Bigelow J, Bower Ay, Girosi F, Meili R, Scoville R and Taylor R. Can Electronic Medical Record Systems Transform Health Care? Potential Health Benefits, Savings, And Costs. *Health Affairs*, 2005;24(5):1103-1117.
- [9] Institute of Medicine. *To err is human: Building a safer health system*. ed. L. Kohn, J. Corrigan, and M. Donaldson. Washington, DC: National Academy Press. 2000.

- [10] Johnson KB, Ravich WJ, Cowan JA Jr. "Brainstorming about next-generation computer-based documentation". *International Journal of Medical Information*. 2004; 73(9-10):665-674.
- [11] Anderson, JG., *Computer-based ambulatory information systems: recent development*, 2000
- [12] Shekelle PG, Morton SC, Keeler EB. "Costs and Benefits of Health Information Technology". Evidence Report/Technology Assessment No. 132. AHRQ Publication No.06-E006. Rockville, MD: Agency for Healthcare Research and Quality. April 2006.
- [13] Agarwal R, and Prasad J. "The Antecedents and Consequents of User Perceptions in Information Technology Adoption". *Decision Support Systems*. 1998; 22:15-29.
- [14] Venkatesh V, Morris M, Davis GB, Davis FD. "User Acceptance of Information Technology: Toward a Unified View". *MIS Quarterly*. 2003; 27(3):425-78.
- [15] Morton ME. "Use and Acceptance of an Electronic Health Record: Factors Affecting Physician Attitudes". Unpublished dissertation, Drexel University. 2008.
- [16] Kim KK, Michelman J. "An examination of factors for the strategic use of information systems in the health care industry". *MIS Q*. 1990;14(2): 201-215.
- [17] Chau P, Hu P. "Examining a model of information technology acceptance by individual professionals: An exploratory study". *Journal of Management Information Systems*. 2001; 18:191-229.
- [18] Devaraj S, Kohli R. "Performance Impacts of Information Technology: Is Actual Usage the Missing Link?" *Management Science*. 2003;49(3):273-289.
- [19] Kohli R, Kettinger WJ. "Informing the Clan: Controlling Physician Costs and Outcomes". *MIS Quarterly*. 2004; 28(3): 363-394.
- [20] Ministry of Economy and Planning. Ninth development plan; 2010-2014. Ministry of Economy and Planning, Kingdom of Saudi Arabia. 2010.
- [21] Alanazy, S. "Factors Associated With Implementation of Electronic Health Records in Saudi Arabia". Unpublished PhD dissertation submitted to the University of Medicine and Dentistry of New Jersey. 2006.
- [22] Harrison JP, Daly MA. "Leveraging Health Information Technology to Improve Patient Safety". *Public Administration & Management*. 2009; 13(3):218-237.
- [23] De Vellis RE *Scale development: Theory and application* (2nd edn) Thousand Oaks, California: Sage. 2003.
- [24] Morgan GA, Gliner JA, Harmon RJ. *Understanding and evaluating research in applied clinical settings*. Publisher: Lawrence Erlbaum Associates; 2006.
- [25] Roberts-Lombard M. *Marketing Research – A South African Perspective*. (3rd ed.). Cape Town: Oxford University Press. 2002.
- [26] Ministry of Health. *Health Statistical Year Book*. Ministry of Health. Kingdom of Saudi Arabia. 2009.
- [27] Alam, SS, Noor, MKM. "ICT Adoption in Small and Medium Enterprises: an Empirical Evidence of Service Sectors in Malaysia". *International Journal of Business and Management*. 2009; 4(2):112-125.
- [28] Hashim J. *Information Communication Technology (ICT) Adoption Among SME Owners in Malaysia*. *International Journal of Business and Information*; 2007; 2(2):221-240.
- [29] Houser SH, Johnson L. "Perceptions Regarding Electronic Health Record Implementation among Health Information Management Professionals in Alabama: A Statewide Survey and Analysis". *Perspectives in Health Information Management*. 2008; 5: 6.
- [30] Thakkar M, Davis DC. *Risks, Barriers, and Benefits of EHR Systems: A Comparative Study Based on Size of Hospital*. *Perspectives in Health Information Management*. 2006; 3(5):1-10.
- [31] MacGregor R, Hyland P, Harvie C. (The Effect of Gender on Perceived Benefits of and Drivers for ICT Adoption in Australian Medical Practices. *International Journal of E-Politics*. 2011; 2(1):68-85.
- [32] Yu P, Li H, Gagnon M. "Health IT acceptance factors in long-term care facilities: A cross-sectional survey". *International Journal of Medical Informatics*. 2009; 78:219-229.
- [33] McDonald CJ, Overhage JM, Tierney WM, Dexter PR, Martin DK, Suico JG, et al. *The Regenstrief Medical Record System: a quarter century experience*. *International Journal of Medical Informatics* 1999;54:225-53.
- [34] Shachak A Fine S. *The effect of training on biologists acceptance of bioinformatics tools: A field experiment*. *Journal of American Society for Information Science*. 2008;59(5):719-730.

# Route Maintenance Approach for Link Breakage Prediction in Mobile Ad Hoc Networks

Khalid Zahedi

Faculty of Computer Science and Information Systems  
Universiti Teknologi Malaysia (UTM)  
Johor, Malaysia

Abdul Samad Ismail

Faculty of Computer Science and Information Systems  
Universiti Teknologi Malaysia (UTM)  
Johor, Malaysia

**Abstract**— Mobile Ad hoc Network (MANET) consists of a group of mobile nodes that can communicate with each other without the need of infrastructure. The movement of nodes in MANET is random; therefore MANETs have a dynamic topology. Because of this dynamic topology, the link breakages in these networks are something common. This problem causes high data loss and delay. In order to decrease these problems, the idea of link breakage prediction has appeared. In link breakage prediction, the availability of a link is evaluated, and a warning is issued if there is a possibility of soon link breakage. In this paper a new approach of link breakage prediction in MANETs is proposed. This approach has been implemented on the well-known Dynamic Source Routing protocol (DSR). This new mechanism was able to decrease the packet loss and delay that occur in the original protocol.

**Keywords**- MANET; link breakage prediction; DSR.

## I. INTRODUCTION

Mobile Ad hoc Network (MANET) consists of a group of mobile nodes that can be communicated with each other wirelessly without the need to any existed infrastructure. MANETs in general are known with its dynamic topology. The nodes are mobile and their movement is random. MANET's dynamic topology makes link breakages a frequent habit. This habit causes many problems such as data loss, delay, and others which degrade the performance of the MANETs protocols. In order to reduce the damage size of this phenomenon, the idea of link breakage prediction has appeared.

In link breakage prediction, a link breakage can be predicted before its real occurring so route maintenance can start before the occurring of the problem avoiding the problems that come with a link breakage. In the link breakage prediction, a node in an active route can predict if the link between it and its previous hop will break soon. In this case it can inform the source node about the problem and the source node, if still needs the route, will be able to construct a new route which avoids this soon to be broken link. It has been found that this procedure has made a good improvement in the performance of the mobile ad hoc network's protocols, but the problem is that the focusing during constructing a new route was only on excluding the link that was predicted to have a link breakage. This mechanism may cause constructing a new

route with some or all bad links from the current used route which are weak but did not predicted to be broken yet. These links may break during or directly after the constructing of the new route which will cause a high decrease in the packet delivery ratio and a high increase in the packet loss and delay. In order to improve the idea of link breakage prediction, this paper has proposed a new approach for link breakage prediction in MANETs. In this new approach, the source node of an active route, after being informed about a link breakage in its current used route, will construct a new route which avoids the use of any link from the current used route. That means excluding all the links in the current route, or in other words, excluding the whole current used route not just the soon to be broken link. So, the new constructed route will be completely different from the current used one. This approach is novel and it has been implemented on the well-known reactive routing protocol Dynamic Source routing Protocol (DSR).

This paper is organized in seven sections: Section I is an introduction. Section II gives some examples of the works that have been done in this area. Section III gives a description about the Dynamic Source Routing protocol (DSR). Section IV illustrates the proposed idea. Section V discusses the simulation environment. Section VI detailed the results that have obtained. Section VII concludes this paper, and section VIII provides some future works.

## II. LITERATURE REVIEW

Several researchers have investigated the area of link breakage prediction in mobile ad hoc networks. In this section, some examples of their works are discussed.

Ramesh et al. [1] have studied the problem of link breakage prediction in the DSR routing protocol. Their idea is that during the route discovery process, the source node builds two routes which are the source route and another route can be used as a backup. The backup route can be used if the primary route (source route) was predicted to have a link breakage soon.

Li et al. [2] have studied the link prediction in the AODV routing protocol by establishing a signal intensity threshold which is Pr-THRESHOLD. If the received signal intensity is lower than the threshold, the upstream node will calculate the distance between it and the sending node through the intensity of the received packet signal, and estimate the relative velocity between it and the sending node through the time difference of

the neighboring received data and the intensity of the packet signal. Then, according to the relative position and the relative velocity with the sending node, a node can estimate when to send a RRER to the sending node to warn it about a link failure. When the source node received this RRER message, it will start its restored process searching its routing table and find another route to the destination.

Qin & Kunz [3] have dealt with the problem of link failure prediction by proposing an equation to calculate the exact time that a link breakage can occur. They named their method the link breakage prediction algorithm. In their idea, each node maintains a table that contains the previous hop node address, the value of the received packet signal power, and the time which this data packet has been received. After receiving three data packets, a node will calculate the link breakage time and compare it with a fixed threshold. If the node predicted that the link with its previous neighbor will have a link breakage soon, it will send a warning message to the source node of the active route to warn it about the link breakage probability. If the source still needs the route it will perform a route discovery process to establish a new route to the destination. Their idea has been implemented using DSR routing protocol.

Zhu [4] has studied the problem of link breakage prediction by using the same equation that have been proposed by Qin & Kunz [3] which is the link breakage prediction algorithm, but she has implemented this algorithm using the AODV and MAODV routing protocols

Choi et al. [5] has dealt with the problem of link breakage prediction in vehicular ad hoc network. They proposed an algorithm to predict a link breakage possibility using the value of the RSSI (Received Signal Strength Indicator). Each vehicle in the network periodically scans the received signals from its neighbors and uses the collected value to calculate the distance, the velocity, and the acceleration of its next hop which it receives data packets from. By calculating these three values, the node can predict if a link breakage will occur, and can determine if the effected link can be maintained or a new link is needed to be constructed. If the effected vehicle found that a link breakage in the link with its next hop will occur, it will use one of its neighbors which has the highest value of RSSI with (that means the one which is the nearest to it) to build a new link with before the previous link with its other neighbor becomes broken.

Goff et al. [6] have studied the link breakage problem in the DSR routing protocol. They defined a region they named it the preemptive region, and they also defined a threshold which they named it the preemptive threshold, they defined this threshold as the signal power of the received packets at the edge of the preemptive region. When a node enters the preemptive region it will send a warning message to the source node of the active route in order to inform it that a link breakage will soon occur. So if the source is still interesting with the route, it will generate a route discovery process to establish a new route without that soon to be broken link.

Ouni et al. [7] studied the problem of link breakage prediction in the DSR routing protocol and tried to propose a solution by proposing a check model composed of two

modules. The first module includes performing different simulations to have an idea about the nodes behavior and by this allowing determining the suitable routes to use, while the second model checks the path availability and the deadline delay satisfaction. This check model was also used to predict the validity periods of the selected path and the satisfaction of the delay constrains.

Lu et al. [8] have worked on the DSR routing protocol and proposed a mechanism for switching to a new route if the current route is found to have a link breakage soon. Their mechanism which is named DSR-link switch (DSR-LS) first detects a link breakage between a node and its next hop to the source by measuring the power of the received packets. If a link failure is detected to occur soon, the node, using this mechanism, will send a link switch request (LSRE) in one hop range to search appropriate nodes that act as relaying stations or bridge nodes. This LSRE request will be sent by including it in the RTS/CTS packets of the MAC layer during the current communication. After finding a new strong links, the current route will be shift to a more stable path.

### III. DYNAMIC SOURCE ROUTING (DSR)

The Dynamic Source Routing (DSR) is a simple and efficient routing protocol designed to be used specifically in mobile ad hoc networks. Through using DSR, the network is completely self-organizing and self-configuring. Network nodes cooperate to forward packets for each other to allow communication over multiple hops between the nodes that are not located within the transmission range of each other. As nodes in the mobile ad hoc network move about, join or leave the network, and as wireless transmission conditions such as types of interference change, all routing is automatically determined and maintained by the DSR routing protocol.

The DSR routing protocol applies the idea of source routing, this idea can be summarized by sending the whole route from the source node to the destination node in each transmitted IP packet, so the intermediate nodes will have to only forward these packets without taking any routing decision. In order to implement the idea of source routing, DSR makes use of special header for carrying control information which can be included in any IP packet. This header is named DSR options header [9].

The DSR options header is a header existed in any sent IP packet by a node implements DSR routing algorithm. This header must immediately follow the IP header in the sent packet. It consists of two fields, fixed length field and variable length field. The fixed length field is a 4-octet portion that has four fields (Next Header, F, Reserved, Payload Length) while the variable length field is called the options field, which has zero or more pieces of optional information which are called DSR options. In DSR routing protocol there are eight types of options, each one of them must be included in a DSR options header in order to be transmitted along the network.

DSR options header is located in an IP packet directly after the IP header and before any other header in the packet. It can contain one or more of the following options:

- 1) *Route Request option.*
- 2) *Route Reply option.*
- 3) *Route Error option.*
- 4) *Acknowledgement request option.*
- 5) *Acknowledgement option.*
- 6) *DSR source route option.*
- 7) *Pad1 option.*
- 8) *PadN option.*

The DSR protocol composes of two basic mechanisms which work together to allow the discovery and maintenance of the source routes in mobile ad hoc networks. These two basic mechanisms are:

- 1) *Route discovery*
- 2) *Route maintenance*

Route discovery is the mechanism that is used by a source node which wishes to send data packets to a destination node which has no route to it in its route cache. Using this mechanism the source node can obtain a source route to the destination.

Route maintenance is the mechanism that is used by a source node to detect a link breakage along its source route to a destination node. Using this mechanism the source node can know if it can still use the route or not. When the source node indicates the existence of a broken link in the source route, it can use another route or trigger a new route discovery process. Route maintenance is used only with active routes.

Route discovery and route maintenance mechanisms each operates entirely on demand. Unlike other protocols, DSR does not require periodic packets of any kind at any level within the network. For example, DSR does not use any periodic routing advertisement and does not use neighbor detection messages. This is a full on demand behavior.

It is possible that a link may not work equally well in both directions because of antenna, or propagation patterns, or sources interference. These types of links are called unidirectional links. The routes that compose of such type of links are called asymmetric routes or paths. DSR allows unidirectional links to be used when necessary; this improves the overall performance and the network connectivity.

DSR also supports the internetworking between different types of wireless networks allowing a source route to be composed of hops over a combination of any types of networks available [10]. As an example, some nodes in the ad hoc network may have only short-range radios, while other nodes have both short-range and long-range radios; the combination of these nodes together can be considered by DSR as a single ad hoc network.

#### IV. THE PROPOSED IDEA

In this section a new approach for the link breakage prediction in the mobile ad hoc networks will be introduced. The idea is to construct a new route which is completely different from the current used route by excluding all the links exist in the current used one. So during the phase of constructing the new route if another link or other links have

been predicted to be broken, there will be no need for trying to avoid this link or these links, because from the beginning, the new constructed route has excluded all the links in the previous route. The approach's idea is as follows:

Each node along an active source route scans the received data packets signals from its previous hop node. When a node found that the Received Signal Strength Indicator (RSSI) value of the received data packets from its previous hop is still decreasing after three successive measurements, the node will realize that the link between it and its previous hop will have a link breakage soon. In this case it will generate a packet and initialize a new option which will be named Soon Link Breakage warning (SLBW). This option will be inserted in the options field of the DSR options header of the packet. Then, this packet which can be named SLBW message will be unicast to the source node of this active route to indicate to it that a link breakage along this route will occur. The SLBW option is similar to the RERR option of the DSR routing protocol with some modifications, the error type in the SLBW will be set to (4) in order to indicate the link breakage probability. SLBW will include the source node's address in order to reach the source of the affected route in case more than one route share some of the links of the affected route, and will also include the addresses of both, the node that predicted the link breakage and its previous hop node's address. By sending the addresses of the nodes at the end of the soon to be broken link, the source node will be able to determine which route will have a link breakage. When the source node receives the SLBW message, if it still needs the route, it will set the route that has a soon to be broken link with the state of Route with a Breakage Prediction (RBP) in its route cache. Then it will check its route cache to see if it has another route to the destination. If it has one, it will make a match between the intermediate node addresses of the cached route and the node addresses in the current used route which has the state (RBP). If there was no match, the source starts sending data packets using this new source route. Otherwise, it will trigger a route discovery process by broadcasting to its neighbors a Modified Route Request (MRREQ) message. The MRREQ message is an IP packet generated by the source node which its DSR options header contains two options, the RREQ option and the source route option.

In the source route option, the source node will append the route with the (RBP) state. This step is made by the source node in order to discover a new route which has no any relationship with the current used route which has the state (RBP), because the current route may have other weak links. Each node receives this MRREQ message will check first if it is the destination of this MRREQ. If it is the destination, it will initialize a RREP option similar to the one in the original DSR routing protocol. Else, it will check if it has received this message before, so if the RREQ option in the received MRREQ message has the same source address and REQUEST ID of a previous received one, or if the receiving node found its address appended in the RECORD of the received option, it will discard this message. Otherwise, the node will check if its address is appended in the source route option of the MRREQ message. If it found its address appended, it will discard the MRREQ message. Else, it will append its address



in the RECORD of the RREQ option in the MRREQ message and rebroadcast the message to its neighbors.

In Fig. 1, in order to construct a route which has no any relationship with the current used one, when node 1 receives the MRREQ message it will make a match between its address and the addresses in the source route option of the MRREQ message. So when it found its address appended, it will discard the message and not forward it any more. The same situation will repeat with the other nodes of the route.

### V. SIMULATION

In this section the parameters that have been manipulated, the metrics that have been used for comparison, and the environment that has been used to make the experiments will be discussed in detail.

#### THE USED PARAMETERS

From our literature review, we found that most of the other papers have used three parameters for making their comparisons; these parameters are (number of nodes in the network, simulation time, and pause time). In order to make new and unique comparisons, we used in this paper three other parameters which we found that no other paper in the link prediction area has used before. These three parameters are:

- 1) *Number of nodes per route.*
- 2) *Node mobility speed.*
- 3) *Node transmission range.*

#### THE USED METRICS

In this paper three metrics have been used in order to make the comparisons between the two protocols. These metrics are:

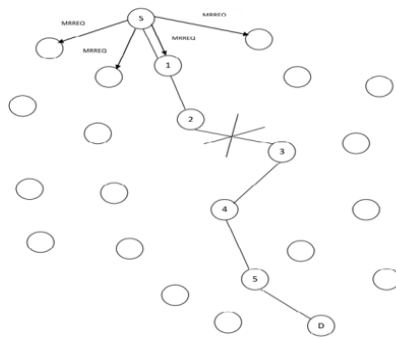


Figure1. A clarification to the idea

- 1) *Packet Delivery Ratio.*
- 2) *Number of dropped data packets.*
- 3) *Average End to End Delay.*

The following is the definition of each metric:

- **Packet Delivery Ratio:** It is the ratio between the number of received data packets by the destination and the number of generated data packets by the source.

- **Number of dropped data packets:** It is the number of data packets that have failed to arrive successfully to the destination.
- **Average End to End Delay:** It is the time that is taken by a packet in order to transfer from a source node to a destination node.

#### THE USED ENVIRONMENT

As we mentioned, the simulations in this paper have been carried out by varying three parameters. When any parameter (of the three used parameters) is manipulated, all the other parameters will be fixed. See Table I.

TABLE I. SIMULATION PARAMETERS

Parameter	Value
Number of nodes in the network	1000
Number of nodes per route	10-100
Mobility speed	10-100 m/s
Transmission range	750-10 m
Simulation time	80 sec
Pause time	5 ms
Terrain size	1000x1000
Traffic mode	CBR
Packet size	1300 byte
Packet sending rate	5 p/s
MAC protocol	IEEE 802.11
Mobility model	Random Waypoint
Antenna type	Omni-directional
Simulator	NCTUns 6.0

### VI. RESULTS AND DISCUSSION

In this section, the achieved results will be discussed in detail. In figure 2, we can see that the (Packet delivery ratio) is decreasing for both protocols as the number of nodes per route is increasing, but the decreasing in the case of (DSR modified) is much less than the decreasing in the (DSR original). The reason of decreasing in the PDR is that when the number of nodes in the route increases this means that the number of links in that route also increases, so the probability of link breakages occurrence also increases. Also, we can notice that the difference in PDR between the two protocols is big when the number of nodes per route is low (as it is clear when there is 10 nodes), but this difference is reduced gradually as the number of nodes per route increases (as it is clear when there is 100 nodes). The reason behind this is that the increase in the number of nodes per route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 3, we can see that the (Number of dropped data packets) is increasing for both protocols as the number of nodes per route is increasing, but the increasing in the case of (DSR modified) is much less than the increasing in the (DSR original). The reason of increasing in the number of dropped data packets is that when the number of nodes in the route increases this means that the number of links in that route also increases, so the probability of link breakages occurrence also increases. Also, we can notice that the difference in the number of dropped data packets between the two protocols is

big when the number of nodes per route is low (as it is clear when there is 10 nodes), but this difference is reduced gradually as the number of nodes per route increases (as it is clear when there is 100 nodes). The reason behind this is that the increase in the number of nodes per route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 4, we can see that the (Average End to End Delay) is increasing for both protocols as the number of nodes per route is increasing, but the increasing in the case of (DSR modified) is much less than the increasing in the (DSR original). The reason of increasing in the average end to end delay is that when the number of nodes in the route increases this means that the number of links in that route also increases, so the probability of link breakages occurrence also increases. Also, we can notice that the difference in the average end to end delay between the two protocols is big when the number of nodes per route is low (as it is clear when there is 10 nodes), but this difference is reduced gradually as the number of nodes per route increases (as it is clear when there is 100 nodes). The reason behind this is that the increase in the number of nodes per route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 5, we can see that the (Packet delivery ratio) is decreasing for both protocols as the mobility speed of nodes is increasing, but the decreasing in the case of (DSR modified) is much less than the decreasing in the (DSR original). The reason of decreasing in the PDR is that the increase in the mobility speed of nodes forming a route means the increase in the link breakages in the links between those nodes. Also, we can notice that the difference in PDR between the two protocols is big when the mobility speed of nodes is low (as it is clear when it is 10 m/s), but this difference is reduced gradually as the mobility speed increases (as it is clear when it is 100 m/s).

In figure 6, we can see that the (Number of dropped data packets) is increasing for both protocols as the mobility speed of nodes is increasing, but the increasing in the case of (DSR modified) is much less than the increasing in the (DSR original). The reason of increasing in the number of dropped data packets is that the increase in the mobility speed of nodes forming a route means the increase in the link breakages in the links between those nodes. Also, we can notice that the difference in the number of dropped data packets between the two protocols is big when the mobility speed of nodes is low (as it is clear when it is 10 m/s), but this difference is reduced gradually as the mobility speed increases (as it is clear when it is 100 m/s). The reason behind this is that the increase in the mobility speed of nodes of the route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 7, we can see that the (Average End to End Delay) is increasing for both protocols as the mobility speed of nodes is increasing, but the increasing in the case of (DSR modified) is much less than the increasing in the (DSR original). The reason of increasing in the average end to end delay is that the increase in the mobility speed of nodes forming a route means the increase in the link breakages in the

links between those nodes. Also, we can notice that the difference in the average end to end delay between the two protocols is big when the mobility speed of nodes is low (as it is clear when it is 10 m/s), but this difference is reduced gradually as the mobility speed increases (as it is clear when it is 100 m/s). The reason behind this is that the increase in the mobility speed of nodes of the route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 8, we can see that the (Packet delivery ratio) is decreasing for both protocols as the transmission range of nodes is decreasing, but the decreasing in the case of (DSR modified) is much less than the decreasing in the (DSR original). The reason behind the decreasing in the PDR is that the decrease in the transmission range of nodes in a route means that the links between those nodes will be weaker, so the probability of breaking such links will be higher. Also, we can notice that the difference in PDR between the two protocols is big when the transmission range of nodes is high (as it is clear when it is 750 m), but this difference is reduced gradually as the transmission range decreases (as it is clear when it is 10 m). The reason behind this is that the decrease in the transmission range of nodes of the route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 9, we can see that the (Number of dropped data packets) is increasing for both protocols as the transmission range of nodes is decreasing, but the increasing in the case of (DSR modified) is much less than the increasing in the (DSR original). The reason behind the increasing in the number of dropped data packets is that the decrease in the transmission range of nodes in a route means that the links between those nodes will be weaker, so the probability of breaking such links will be higher. Also, we can notice that the difference in the number of dropped data packets between the two protocols is big when the transmission range of nodes is high (as it is clear when it is 750 m), but this difference is reduced gradually as the transmission range decreases (as it is clear when it is 10 m). The reason behind this is that the decrease in the transmission range of nodes of the route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

In figure 10, we can see that the (Average End to End Delay) is increasing for both protocols as the transmission range of nodes is decreasing, but the increasing in the case of (DSR modified) is much less than the increasing in the (DSR original). The reason behind the increasing in the average end to end delay is that the decrease in the transmission range of nodes in a route means that the links between those nodes will be weaker, so the probability of breaking such links will be higher. Also, we can notice that the difference in the average end to end delay between the two protocols is big when the transmission range of nodes is high (as it is clear when it is 750 m), but this difference is reduced gradually as the transmission range decreases (as it is clear when it is 10 m). The reason behind this is that the decrease in the transmission range of nodes of the route reduces the efficiency of the new mechanism where link breakages will so frequently occur.

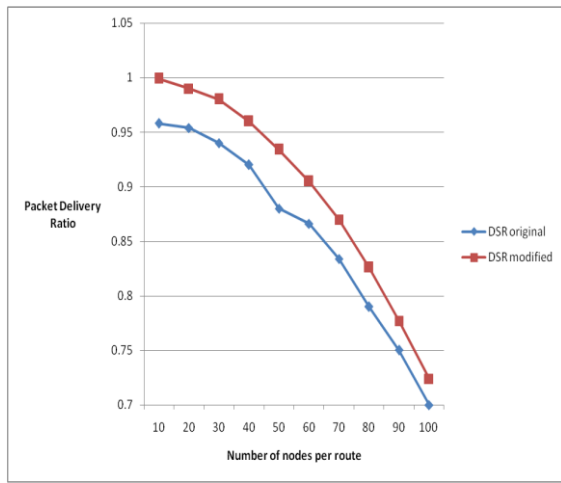


Figure 2 PDR and No. of nodes per route

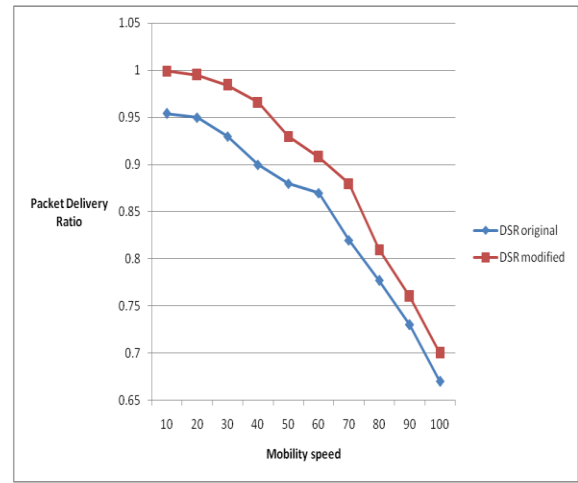


Figure 5 PDR and mobility speed

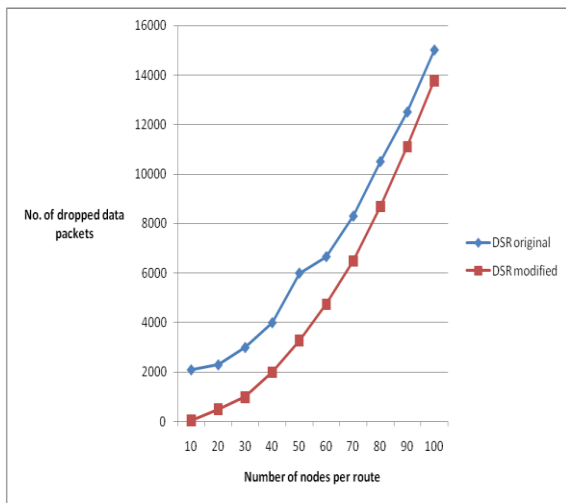


Figure 3 No. of dropped packets and No. of nodes per route

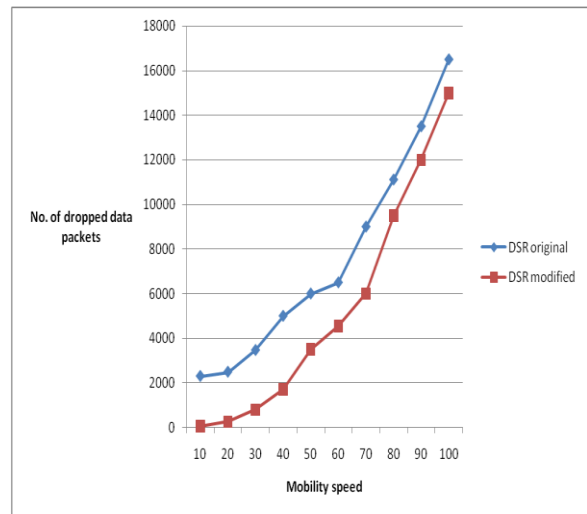


Figure 6 No. of dropped packets and mobility speed

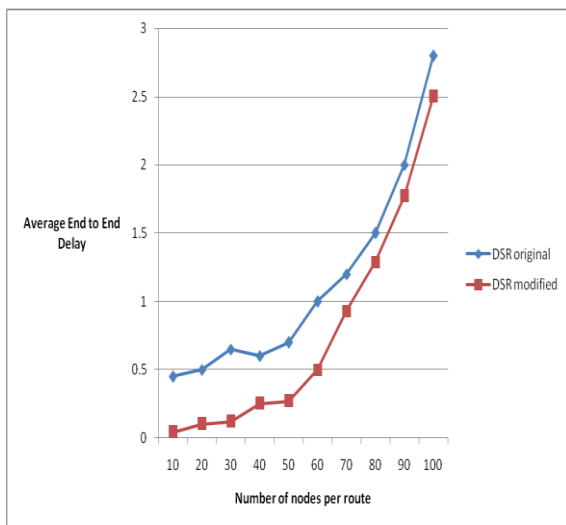


Figure 4 Delay and No. of nodes per route

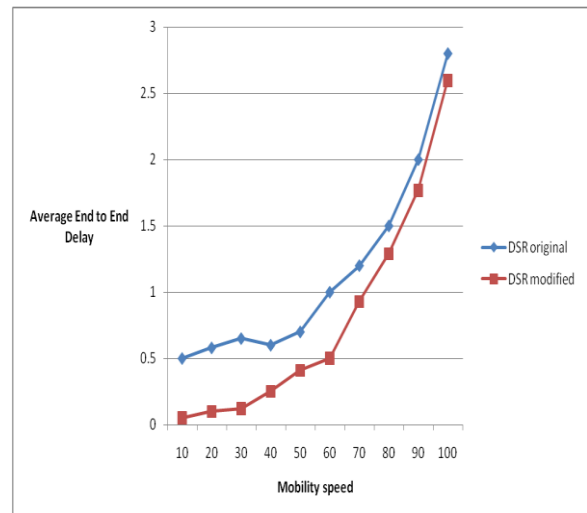


Figure 7 Delay and mobility speed

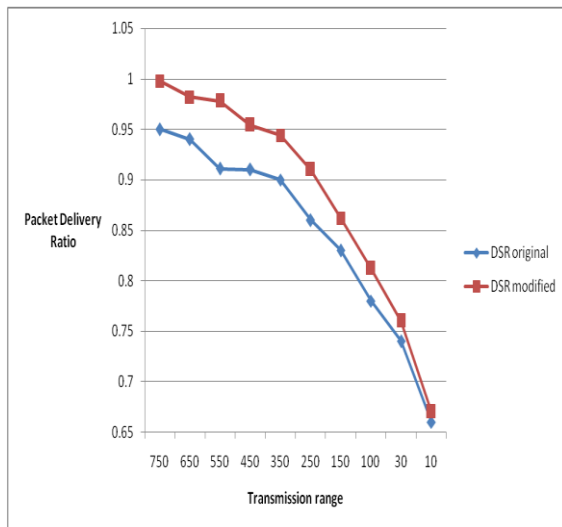


Figure 8 PDR and transmission range

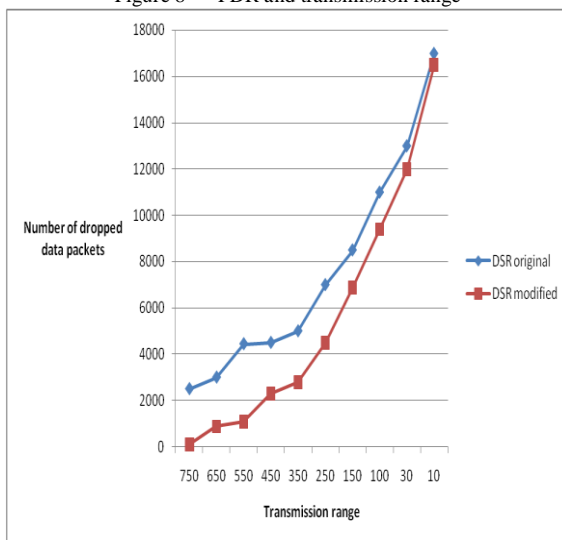


Figure 9 No. of dropped packets and transmission range

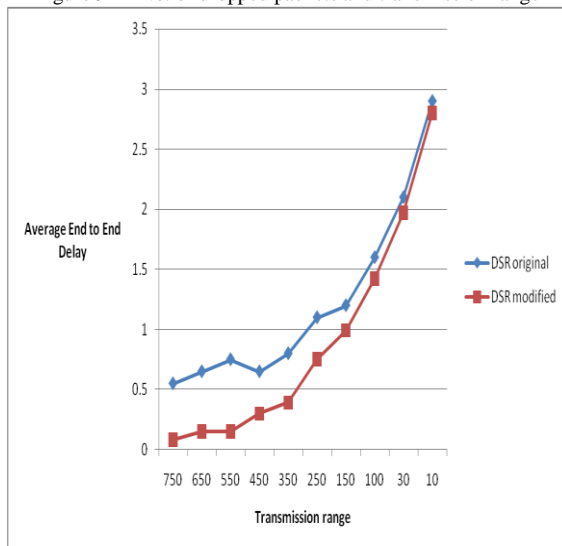


Figure 10 Delay and transmission range

## VII. CONCLUSION

Many approaches have been proposed to deal with the idea of link breakage prediction, but the problem is that all the previous approaches were building a new route that avoids using only the same soon to be broken link, but no one of these approaches was able to build a new route which avoids all the other links in the old route. In this paper, a new approach for solving the problem of link breakages in MANET has been proposed and implemented on the Dynamic Source Routing (DSR) routing protocol. In this approach, the Received Signal Strength Indicator (RSSI) value will be used by a node along an active route to predict a link breakage in its link with its next hop to the source node of this active route. The node will warn the source node, and the source (if it still needs the route) will discover a new route without using any link from the current route which has a soon to be broken link. The idea behind this is to reduce the probability of constructing a route with bad links which can break during or directly after the constructing of a new route. It has been found that this approach was able to increase the packet delivery ratio and decrease both the packet loss and the end to end delay comparing to the DSR routing protocol. So, this approach was able to improve the performance of the protocol.

## VIII. FUTURE WORK

As a future work, this work can be extended by using other metrics for making the comparisons between the original and modified DSR routing protocols such as the terrain size, packet size, packet sending rate, and others. Also, the traffic mode can be changed from CBR to VBR and find the difference. Another change can be made to the mobility model. In this work the mobility model that has been used is the random way point mobility model, so another research can be done by using other mobility models such as the random walk mobility model, or the random direction mobility model, and see the difference.

## REFERENCES

- [1] Ramesh, V., Subbaiah, P., and Supriya, K. (2010). Modified DSR (preemptive) to reduce link breakage and routing overhead for MANET using proactive route maintenance (PRM). *Global Journal of Computer Science and Technology*. Vol. 9. Issue 5, 124-129.
- [2] Li, Q., Liu, c., and Jiang, H. (2008). The routing protocol AODV based on link failure prediction. *ICSP IEEE*. 2008.
- [3] Qin, L., and Kunz, T. (2002). Increasing packet delivery ratio in DSR by link prediction. *HICSS 03. IEEE*. 2002. Hawaii.
- [4] Zhu, Y. (2002). Proactive connection maintenance in AODV and MAODV. Master of Science. Carleton University, Canada.
- [5] Hoi, W., Nam, J., and Choi, S. (2008). Hop state prediction method using distance differential of RSSI on VANET. *NCM 2008. IEEE*. 426-431.
- [6] Goff, T., Abu-Ghazaleh, N., Phatak, D., and Kahvecioglu, R. (2003). Preemptive routing in ad hoc networks. *Journal of Parallel and Distributed Computing*. 63 (2003), 123-140.
- [7] Ouni, S., Bokri, J., and Kamoun, F. (2009). DSR based routing algorithm with delay guarantee for ad hoc networks. *Journal of Networks*. Vol. 4(5), 359-369.
- [8] Lu, H., Zhang, J., and Luo, X. (2008). Link switch mechanism based on DSR route protocol. *ICINIS IEEE*. 2008.
- [9] Internet Engineering Task Force (IETF), Dynamic Source Routing (DSR) protocol, 2007. <http://www.ietf.org>

- [10] Sarkar, S. K., Basavaraju, T., Puttamadappa, C. (2008). Ad hoc mobile wireless networks. (First ed.). New York: Auerbach Publications.
- [11] Rajabhushanam, C., & Kathirvel, A. (2011). Survey of Wireless MANET Application in Battlefield Operations. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(1).
- [12] Multicasting over Overlay Networks – A Critical Review. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(3), 54-61.

AUTHORS PROFILE

**Khalid Zahedi** is a Master's student in the faculty of Computer Science and Information Systems in UTM. His research interests include: Mobile Ad hoc Networks (MANETs) and Vehicular Ad hoc Networks (VANETs).

**Abdul Samad Ismail** is an associate professor in the faculty of Computer Science and Information Systems in UTM. His research interests include: Mobile Ad hoc Networks (MANETs), Vehicular Ad hoc Networks (VANETs), Wireless Sensor Networks (WSNs), and Network Security.

# Web Service Architecture for a Meta Search Engine

K.Srinivas

Associate Professor, Department of  
IT, Geethanjali College of  
Engineering & Technology,  
Cheeryal(V), Keesara(M), Ranga  
Reddy, Andhra Pradesh-501301,  
India.

P.V.S. Srinivas

Professor & Head, Department of  
CSE, Geethanjali College of  
Engineering & Technology,  
Cheeryal(V), Keesara(M), Ranga  
Reddy, Andhra Pradesh-  
501301,India.

A.Govardhan

Professor, Department of CSE,  
School Of Information Technology,  
Jawaharlal Nehru Technological  
University Hyderabad, Kukapally,  
Hyderabad,  
Andhra Pradesh-500085, India.

**Abstract**—With the rapid advancements in Information Technology, Information Retrieval on Internet is gaining its importance day by day. Nowadays there are millions of Websites and billions of homepages available on the Internet. Search Engines are the essential tools for the purpose of retrieving the required information from the Web. But the existing search engines have many problems such as not having wide scope, imbalance in accessing the sites etc. So, the effectiveness of a search engine plays a vital role. Meta search engines are such systems that can provide effective information by accessing multiple existing search engines such as Dog Pile, Meta Crawler etc, but most of them cannot successfully operate on heterogeneous and fully dynamic web environment. In this paper we propose a Web Service Architecture for Meta Search Engine to cater the need of heterogeneous and dynamic web environment. The objective of our proposal is to exploit most of the features offered by Web Services through the implementation of a Web Service Meta Search Engine.

**Keywords**-Meta Search Engine; Search engine; Web Services.

## I. INTRODUCTION

A Meta Search Engine is a search tool that sends user requests to several other search engines and/or databases and aggregates the results into a single list or displays them according to their source. Meta Search Engines enable users to enter search criteria once and access several search engines simultaneously [18]. More comprehensive search results can be obtained by combining the results from several search engines. This may save the user to use multiple search engines separately.

This paper explores the web services and also the Architecture of a Meta Search Engine. In section 2, we propose the existing background and architecture of a Meta Search Engine. Section 3 describes the architecture of a Web Service framework. Section 4 describes about the Meta Search Engine architecture, components and function of the each component. In section 5, we propose the Web Service Meta Search Engine architecture and Section 7 provides the advantages of Web Service Metasearch Engines. Section 8 finally presents the conclusions and the future work.

## II. BACKGROUND

There are many methods for finding information, but one of the leading ways is through search engines. Now a day,

everyone uses search engines for research, school, business, shopping, or entertainment. So, the traffic on the Web is growing exponentially [2]. To understand the working strategy of a Search Engine, one should have an overview and clear understanding of Information Retrieval System. Information retrieval deals with the representation, storage, organization and access to information items in order to give the user desired information. A distinction between traditional information retrieval and web information retrieval is that in traditional or classic information retrieval, the process of search is simple[6] and these document collections are stored in physical form. An example would be looking for information in books of a public library. Nevertheless, nowadays, most of the documents are computerized that can be retrieved with the help of a computer. Web information retrieval is on the other hand, not like the traditional IR, where, searching is performed within the globally largest collection of documents that are linked, such as the well known search services on the internet like Google or Yahoo [4].

Intense stress on user requirements recommended the architecture of Meta Search Engine. A few Meta Search Engines has been already proposed that provides quick response with re ranked results after extracting user preference. It uses Naïve Bayesian Classifier for re ranking. An enhanced version of open source Helios Meta Search Engine takes input keywords along with specified context or language and gives refined results as per user's need [8]. All the proposed solutions refine search-results up to some extent but they have a serious drawback, which is that the user profile is not stationary. The idea behind metasearch is to use multiple "helper" search engines to do the search, then to combine the results from these engines. Engines that use metasearch include Metacrawler, SavvySearch, MSN Search and Altavista, among others [11].

## III. WEB SERVICES

Web service protocol is designed for providing service via web. Web Services are emerging as the fundamental building blocks for creating distributed, integrated and interoperable solutions across the Internet. They represent a new paradigm in distributed computing that allow applications to be created from multiple Web Services dispersed across the web originating from various sources regardless of where they reside or how they were implemented [19]. Unlike services in



general, Web services are based on specifications for data transfer, method invocation and publishing. This is often misunderstood and when a Web service is mentioned it sometimes refers to a general service provided on the Web, like the weather forecast on a Web page for example. The weather forecast is a service and provides its functionality for a variety of users but unless it comprises an interface to communicate with other applications via SOAP [16]. Web services can be seen as software components with an interface to communicate with other software components. They have a certain functionality that is available through a special kind of Remote Procedure Call. In fact they even evolved from traditional Remote Procedure Calls. There are various aspects of Web services. Messaging, discovery, portals, roles, and coordination. The format of the messages exchanged between a client and a Web service is specified by a standard called SOAP. The Simple Object Access Protocol (SOAP) is defined by the W3C as a lightweight protocol for exchange of information in a decentralized, distributed environment. SOAP is an XML based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined data types, and a convention for representing remote procedure calls and responses [12],[16].

Web service messages are sent across the network in an XML format defined by the W3C SOAP specification. In most Web services, there are two types of SOAP messages: requests and responses. When a SOAP request is received, the Web service performs an action based on the request and returns a SOAP response. In many implementations, SOAP requests are similar to function calls with SOAP responses returning the results of the function call [12]. With SOAP, a communication between Web services is possible and structured and each participant knows how to send or receive the corresponding SOAP Message. The final step to complete the communication architecture of Web services is to define how to access a service once it is implemented.

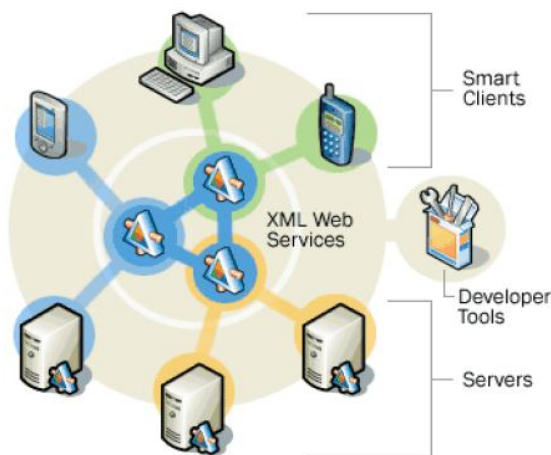


Figure 1.

Web services are the core any type of service, as it holds and connects everything together as shown in Figure 1.

#### IV. META SEARCH ENGINE

Meta Search Engines can be classified into two types. a) General purpose metasearch engine and b) Special purpose Meta Search Engines. The former aims to search the entire Web, while the latter focuses on searching information in a particular domain (e.g., news, jobs).

- **Major Search Engine Approach:** This approach uses a small number of popular major search engines to build a metasearch engine. Thus, to build a general-purpose metasearch engine using this approach, we can use a small number of major search engines such as Google, Yahoo!, Bing (MSN) and Ask. Similarly, to build a special purpose Meta Search Engine for a given domain, we can use a small number of major search engines in that domain.
- **Large scale Metasearch engine approach:** In this approach, a large number of mostly small search engines are used to build a Metasearch engine. For example, to build a general-purpose metasearch engine using this approach, we can perceivably utilize all documents driven search engines on the Web. Such a metasearch engine will have millions of component search engines. Similarly to build a special purpose metasearch engine for a given domain with this approach, we can connect to all the search engines in that domain. For instance, for the news domain, tens of thousands of newspaper and news-site search engines can be used.

Each of the above two approaches has its advantages and disadvantages. An obvious advantage of the major search engine approach is that such a metasearch engine is much easier to build compared to the large-scale metasearch engine approach because the former only requires the metasearch engine to interact with a small number of search engines. Almost all currently popular metasearch engines, such as Dogpile, Mamma and MetaCrawler, are built using the major search engine approach, and most of them use only a handful of major search engine. One example of a large-scale special-purpose metasearch engine is AllInOneNews, which uses about 1,800 news search engines from about 200 countries/regions. In general, more advanced technologies are required to build large-scale metasearch engines. As these technologies become more mature, more large-scale metasearch engines are likely to be built.

#### V. PROPOSED ARCHITECTURE

The proposed architecture takes both approaches into consideration for designing web service metasearch engine system. Significant software components included in this architecture search engine selector, search engine connectors, result extractors and result merger.

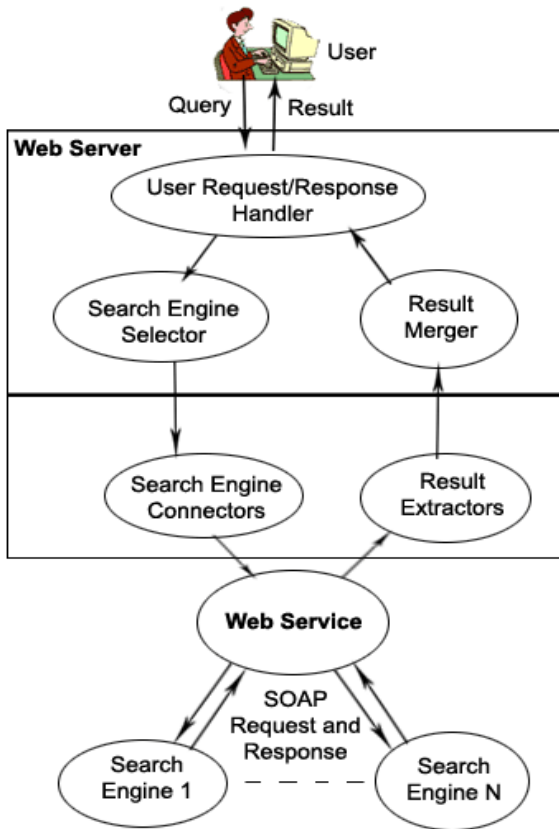


Figure 2 Web Service Meta Search Engine Architecture

#### A. Search Engine Selector:

If the number of component search engines in a metasearch engine is very small, say less than 10, it might be reasonable to send each user query submitted to the metasearch engine to all the component search engines. In this case, the search engine selector is probably not needed. However, if the number of component search engines is large, as in the large scale Meta Search Engine scenario, then sending each query to all component search engines will be an inefficient strategy because most component search engines will be useless with respect to any particular query. For example, suppose a user wants to find 50 best matching results for his/her query from a metasearch engine with 1,000 component search engines. Since the 50 best results will be contained in no more than 50 component search engines, it is clear that at least 950 component search engines are useless for this particular query.

Passing a query to useless search engines may cause serious problems for efficiency. Generally, sending a query to useless search engines will cause waste of resources to the metasearch engine server, each of the involved search engine servers and the Internet. Specifically, dispatching a query, including needed query reformatting, to a useless search engine and handling the returned results, including receiving the returned response pages, extracting the result records from these pages, and determining whether they should be included in the final

merged result list and where they should be ranked in the merged result list if they are to be included, waste the resources of the metasearch engine server; receiving the query from the metasearch engine, evaluating the query, and returning the results back to the metasearch engine waste the resources of each search engine whose results end up useless; and finally transmitting a query from the metasearch engine to useless search engines and transmitting useless retrieved results from these search engines to the metasearch engine waste the network resources of the Internet. Therefore, it is important to send each user query to only potentially useful search engines for processing.

The problem of identifying potentially useful component search engines to invoke for a given query is the search engine selection problem, which is sometimes also referred to as database selection problem, server selection problem, or query routing problem. Obviously, for metasearch engines with more component search engines and/or more diverse component search engines, having an effective search engine selector is more important.

#### B. Search Engine Connectors:

After a component search engine has been selected to participate in the processing of a user query, the search engine connector established a connection with the server of the search engine and passes the query to it. Different search engines usually have different connection parameters. As a result, a separate connector is created for each search engine. In general, the connector for a search engine S needs to know the HTTP (Hyper Text Transfer Protocol) connection parameters supported by S. There are three basic parameters, (a) the name and location of the search engine server, (b) the HTTP request method (usually it is either GET or POST) supported by S, and (c) the name of the string variable that is used to hold the actual query string.

When implementing metasearch engines with a small number of component search engines, experienced developers can manually write the connector for each search engine. However, for large-scale metasearch engines, this can be very time-consuming and expensive. Thus, it is important to develop the capability of generating connectors automatically.

An intelligent metasearch engine may modify a query it receives from a user before passing it to the search engine connector if such a modification can potentially improve the search effectiveness. For example, a query expansion technique may be used by the metasearch engine to add terms that are related to the original user query to improve the chance for retrieving more relevant documents.

#### C. Web Service:

A Meta search engine is a search engine that collects results from other search engine. Web service offers such functionality and then presents a summary of that information as the results of a search. Most search engines available on the Web provide only a browser based interface; however, because Web services

start to be successful, some of those search engines offer also an access to their information through Web services. Two types of search engines are observed, one that acts like a wrapper for the HTML pages returned by the search engine and other one is build upon the Web service offered by the search engine but this difference is visible only when looking at the internal processing of the service. It is difficult to distinguish them from the outside as they implement the same interface.

Web service are built for, any process that can be integrated into external systems through valid XML documents over Internet protocols. This definition outlines the general idea of Web services. Web services can be seen as software components with an interface to communicate with other software components. They have a certain functionality that is available through a special kind of Remote Procedure Call.

SOAP, the Simple Object Access Protocol [16] was developed to enable a communication between Web services. It was designed as a lightweight protocol for exchange of information in a decentralized, distributed environment. SOAP is an extensible, text-based framework for enabling communication between diverse parties that have no prior knowledge of each other. This is the requirement a transport protocol for Web services has to fulfill. SOAP species a mechanism to perform remote procedure calls and therefore removes the requirement that two systems must run on the same platform or be written in the same programming language.

#### D. Result Extractors:

After a component search engine processes a query, the search engine will return one or more response pages. A typical response page contains multiple (usually 10) search result records, each of which corresponds to a retrieved Web page, and it typically contains the URL and the title of the page, a short summary (called snippet) of the page content, and some other pieces of information such as the page size. Fig 2 shows the upper portion of a response page from the Google search engine. Response pages are dynamically generated HTML documents, and they often also contain content unrelated to the user query such as advertisements (sponsored links) and information about the host Web site.

A program (i.e., result extractor) is needed to extract the correct search result records from different component search engines can be merged into a single ranked list. This program is sometimes called an extraction wrapper. Since different search engines often format their results differently, a separate result extractor is usually needed for each component search engine. Although experienced programmers can write the extractors manually, for large-scale metasearch engines, it is desirable to develop techniques that can generate the extractors automatically.

#### E. Result Merger:

After the results from the selected component search engines are returned to the metasearch engine, the result merger combines the results into a single ranked list. The ranked list of

search result records is then presented to the user, possible 10 records on each page at a time, just like more search engines do. Many factors may influence how result merging will be performed and what the outcome will look like. The information that could be utilized includes the local rank of a result record from a component search engine, the title and the snippet of a result record, the full document of each result, the publication time of each retrieved document, the potential relevance of the search engine with respect to the query from where a result is retrieved, and more. A good result merger should rank all returned results in descending order of their desirability.

The existing architecture has many disadvantages in search engine selection, search engine connection and result extractors. We proposed a robust metasearch engine architecture using web services for heterogeneous and dynamic environment.

## VI. IMPLEMENTATION

WSMSE is a web based Meta Search Engine that is developed using Java Web Service, Servlet and JSP. This search engine is developed based on the assumption that an average web user makes searches based on imprecise query keywords or sentences, which in turn leads to unnecessary or inaccurate results. In this work, we demonstrate that with a simple tweak or manipulation of existing search engine functions, which can helps users to get better search results.

To build and deploy Web service Java Web Service Developer Pack (WSDP) will be used. Based on the Java 2 SDK, this toolset adds new API including XML Messaging (JAXM), XML Processing (JAXP), XML Registries (JAXR), XML-based RPC (JAX-RPC) and the SOAP with Attachments (SAAJ). The main advantage Java WSDP is it supported heterogeneous platform and have dynamic behavior.

Every Web service is deployed with a Web Service Description Language (WSDL) file that acts like an instruction manual; in case someone wants to use a particular Web Service, he simply has to look at that WSDL file to learn how to communicate with the corresponding service and use it. Amongst all advantages, the most important one for the Web services is that XML is not platform dependent and it allows easy data processing and exchange between different applications.

HTTP is the most popular transfer protocol and it's supported on almost all platforms. By implementing this standard, combined with XML, Web services remove almost all frontiers between platforms.

The Simple Object Access Protocol (SOAP) exchange of information in a decentralized, distributed environment. SOAP is an XML based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined data types, and a convention for representing remote procedure calls and responses.

```
<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope
xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:soapenc="http://schemas.xmlsoap.org/soap/encoding/"
xmlns:tns="http://adv/WSMSE"
xmlns:types="http://adv/WSMSE/encodedTypes"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <soap:Body
soap:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
    <tns:GetAvailableSEWS />
  </soap:Body>
</soap:Envelope>
```

Figure 3 SOAP request to the WSMSE, invoking the Get Available Search Engine Web Service

Web Service Description Language (WSDL) defines an XML grammar for describing network services as a set of endpoints that accept messages containing either document-oriented or procedure-oriented information. The endpoint is defined by a network protocol and a message format, however, the extensible characteristic of WSDL allow the messages and endpoints being described regardless of what message formats or network protocols are being used to communicate. In other words, a WSDL file is an XML document that describes a set of SOAP messages and how the messages are exchanged.

Universal Discovery Description and Integration (UDDI) is the yellow pages of Web services. A UDDI directory entry is an XML file that describes a business and the services it offers. Both can be categorized and have keys so one can find a provider or a service by different ways. Web services are also defined through a document called a Type Model (tModel) that describes their interface. A tModel is simply a WSDL file without the <service> tag; it contains information to generate the different proxy classes or SOAP messages to communicate with the Web service but it does not specify the access point of the service.

## VII. ADVANTAGES OF WEB SERVICE META SEARCH ENGINE

We attempt to provide a comprehensive analysis of the potential advantages of metasearch engines over search engines. We will also focus on the comparison of metasearch engine and search engine.

### A. Increased Search Coverage:

Metasearch engine can search any document that is indexed by at least one of the search engines. Hence, the search coverage of a metasearch engine is the union of those search engines. Metasearch engine with multiple major search engines as components will have larger coverage than any single component search engine. Different search engines often

employ different document representation and result ranking techniques, and as a result, they often return different sets of top results for the same user query. Thus, by retrieving from multiple major search engines, a metasearch engine is likely to return more unique high quality results for each user query.

### B. Better Content Quality:

The quality of the content of a search engine can be measured by the quality of the documents indexed by the search engine. The quality of a document can in turn be measured in a number of ways such as the richness and the reliability of the contents. General-purpose metasearch engines implemented using the large-scale metasearch engine approach has a better chance to retrieve more up-to-date information than major search engines and metasearch engines that are built with major search engines.

### C. Good Potential for Better Retrieval Effectiveness:

More unique results are likely to be obtained, even among those highly ranked ones, due to the fact that different major search engines have different coverage and different document ranking algorithms. The result-merging component of the metasearch engine can produce better results by taking advantage of the fact that the document collection of major search engines has significant overlaps. This means that many shared documents have the chance to be ranked by different search engines for any given query. If the same document is retrieved by multiple search engines, then the likelihood that the document is relevant to the query increases significantly because there is more evidence to support its relevance. In general, if a document is retrieved by more search engines, the document is more likely to be relevant.

### D. Better Utilization of Resources:

Metasearch engine use component search engines to perform basic search. This allows them to utilize the storage and computing resources of these search engines.

## VIII. CONCLUSIONS AND FUTURE WORK

The goal of this proposal is to shed some light on the reasons that make building metasearch engines, especially large scale metasearch engines, difficult and challenging. Modern search engines providing interfaces that allow external applications to issue Web search queries that are actually processed using their large scale computing infrastructure. This paper proposes a robust Meta search engine, which can communicate to heterogeneous platform using advanced web service techniques. As much progress has been made in advanced metasearch engine technology, several challenges need to be addressed before truly large scale Meta Search Engine can be effectively built and manage.

## REFERENCES

- [1] Lee Underwood, "A Brief History of Search Engines"; [www.webreference.com/authoring/search\\_history/](http://www.webreference.com/authoring/search_history/).
- [2] Ritu Khare, Yuan An and Il-Yeol Song on Understanding Deep Web Search Interfaces: A Survey. SIGMOD Record, March 2010 (Vol. 39, No 1).
- [3] DogPile metasearch engine. <http://www.dogpile.com>.

- [4] Carol L. Barry. The Identification of User Criteria of Relevance and Document Characteristics: Beyond the Topical Approach to Information Retrieval. PhDthesis, Syracuse, 1993.
- [5] Ana B. Benitez, Mandis Beigi, and Shih-Fu Chang. Using relevance feedback in content & based image metasearch. *IEEE Internet Computing*,2(4):58-69, 1998.
- [6] Eric J. Glover, Steva Lawrence, William P. Birmingham and C. Lee Giles on Architecture of the Metasearch Engine that Supports User Information Needs, ACM.
- [7] Tieli Sun Wei Zhao and Zhiyan Zhao An Architecture of Pattern-Oriented Distributed Meta-Search Engine ACM.
- [8] Amir Hossein Keyhanipoor, Maryam Piroozmand, Behzad Moshiri and Caro Lucas A Multi layer/Multi-Agent Architecture for Meta-Search Engine, AIML 05 Conference, 19-21 December 2005, CICC, Cairo, Egypt.
- [9] Jae Hyun Lim, Young-Chan Kim, Hyonwoo Seung, Jun Hwang , Heung-Nam Kim, "Query Expansion for Intelligent Information Retrieval on Internet", International Conference on Parallel and Distributed Systems, 1997.
- [10] Eric J. Glover, Steve Lawrence, William P. Birmingham, C. Lee Giles; "Architecture of a Meta-Search Engine that Supports User Information Needs", Information and knowledge management CIKM '99, ACM Press,Pages: 210 - 216, 1999.
- [11] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", World-Wide Web Conference, 1998.
- [12] Candadai, "A Dynamic Implementation Framework for SOA-based Applications", Web Logic Developers Journal: WLDJ, September/October 2004.
- [13] UDDI specifications – <http://www.uddi.org>.
- [14] SOAP specifications and RFCs - <http://www.w3.org/TR/soap/>.
- [15] WSDL Specifications and RFCs - <http://www.w3.org/TR/wsdl>.
- [16] T. Clements. Overview of SOAP. Java Developers Forum, <http://java.sun.com/developer/technicalArticles/xml/webservices/>.
- [17] Ethan Cerami. Web Services Essentials. O'Reilly, February 2002.
- [18] K.Srinivas, P.V.S.Srinivas and A.Govardhan. "A Survey on the "Performance Evaluation of Various Meta Search Engines" IJCSI Volume 8, Issue 3,No. 2,May 2011 Pages 359-364.
- [19] Nath, R. (2011). An Authorization Mechanism for Access Control of Resources in the Web Services Paradigm. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(6), 36-42.

# A Statistical Approach For Latin Handwritten Digit Recognition

Ihab Zaqout

Dept. of Information Technology  
Faculty of Engineering & Information technology  
Al-Azhar University – Gaza  
Gaza Strip, Palestine

**Abstract**— A simple method based on some statistical measurements for Latin handwritten digit recognition is proposed in this paper. Firstly, a preprocess step is started with thresholding the gray-scale digit image into a binary image, and then noise removal, spurring and thinning are performed. Secondly, by reducing the search space, the region-of-interest (ROI) is cropped from the preprocessed image, then a freeman chain code template is applied and five feature sets are extracted from each digit image. Counting the number of termination points, their coordinates with relation to the center of the ROI, Euclidian distances, orientations in terms of angles, and other statistical properties such as minor-to-major axis length ratio, area and others. Finally, six categories are created based on the relation between number of termination points and possible digits. The present method is applied and tested on training set (60,000 images) and test set (10,000 images) of MNIST handwritten digit database. Our experiments report a correct classification of 92.9041% for the testing set and 95.0953% for the training set.

**Keywords**- Digit recognition; freeman chain coding; feature extraction; classification.

## I. INTRODUCTION

The significant task of handwritten digit recognition has great importance in the recognition of postcodes sort mail, bank check amounts and so on. Since three decades, there is no single classifier performs the best for all pattern classification problems consistently. There are different challenges faced while attempting to solve this problem. The handwritten digits are not always of the same thickness, size, orientation or position relative to margins.

There are several approaches for handwritten digit recognition problem have been reported in the literature in the past. They include SVM (support vector machine) [1, 2, 3], NN (neural network) [4, 5, 6, 7], deformable template matching [8, 9], hybrid method [10, 11, 12, 13] and others. In SVMs for digit classification problems, the training of a large data set is still a bottle-neck and is comparatively slow. NNs have been widely used to solve complex classification problems. However, A single NN often exhibits with the over fitting behavior which results in a weak generalization performance when trained on a limited set of training data. A better deformation algorithms and proper selection of representative prototypes along with its computational requirements are

required for deformable template matching method. Hybrid method has been widely used in pattern recognition applications. It combines two or more of the above mentioned methods or others to overcome their individual weakness and to preserve their individual advantages. But it is still an open problem to obtain a superior hybrid method.

A comprehensive benchmark of handwritten digit recognition with several state-of-the-art approaches, datasets, and feature representations had been reported by [14]. Several classifiers and feature vectors are evaluated on MNIST handwritten digit database [15].

In this paper, our research is focused on an accurate and feasible method applied and tested on training set (60,000 images) and test set (10,000 images) of MNIST handwritten digit database. We agree with others that the preprocess stage is a crucial and it reflects the accuracy of the classification process. Firstly, a preprocess step is started with thresholding the gray-scale digit image into a binary image, and then noise removal, spurring and thinning are performed. Secondly, by reducing the search space, the region-of-interest (ROI) is cropped from the preprocessed image, then a freeman chain code template is applied and five feature sets are extracted from each digit image. Counting the number of termination points, their coordinates with relation to the center of the ROI, Euclidian distances, orientations in terms of angles, and other statistical properties such as minor-to-major axis length ratio, area and others.

Finally, six categories are created based on the relation between number of termination points and possible digits. The advantage of method is that it does not require training, which can save a lot of training time. Experimental results on MNIST database will be reported in the paper to support the feasibility of our method. The remainder of this paper is organized as follows. The system overview is presented in Sec. 2. The feature extraction is proposed in Sec. 3. In Sec. 4, the classification is discussed. Experimental results are shown in Sec. 5 to demonstrate the reliability of our method. Finally our conclusion and future work is given in Sec 6.

## II. SYSTEM OVERVIEW

In this section, the system overview is introduced. The recognition system includes three dependent stages: the



preprocess, the feature extraction and classification as shown in Figure 1.

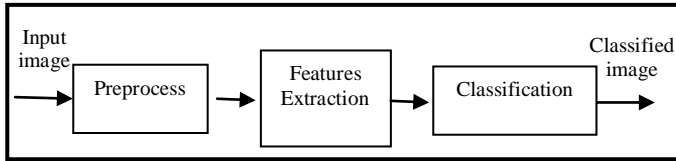


Figure 1. The System Overview

In this paper, a MNIST digit database is used as a dataset to the proposed classification processes. A preprocessed step is started with thresholding the gray-scale digit image into a binary image, and then noise removal, spurring and thinning [16] are performed. During the process of image thinning, we introduce Freeman chain code tracking [17] as shown in Figure 2. To reduce the search space, the region-of-interest (ROI) is cropped from the preprocessed image, then a freeman chain code template is applied and five feature sets are extracted from each digit image. Counting the number of termination points, their coordinates with relation to the center of the ROI, Euclidian distances, orientations in terms of angles, and other statistical properties such as minor-to-major axis length ratio, area and others. Six categories are created based on the relation between number of termination points and possible digits.

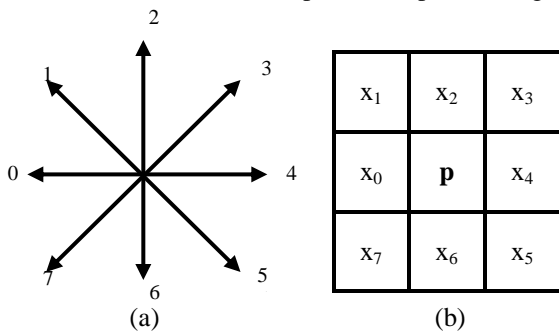


Figure 2. Freeman chain code and 3x3 template

### III. FEATURE EXTRACTION

A fast recognition system has to consider the processing time and its complexity is primarily determined by the feature extraction. In this section, the feature extraction is performed on the thinned image. Five feature sets are extracted from each ROI digit image:

- Number of termination points, xy-coordinates and orientations,
- Minimum-to-maximum distance ratio (Euclidean distance between the xy-coordinate of the center of the ROI and xy-coordinates of termination points), minor-to-major axis length ratio and their spatial distribution compared to the center of the ROI,
- Sum of white pixels after partitioning the thinned digit image into horizontal blocks for 0, 2, 3, 5, 6 and 8,
- Other statistical measurements such as area, minor-to-major axis length ratio, filled area to area ratio,

filled area to convex area ratio and number of holes, and

- Sum Freeman chain code.

As shown in Fig. 2(b), the pixels  $x_1, x_2, \dots, x_8$  are the 8-neighbours of  $p$  in its  $3 \times 3$  template and said to be 8-adjacent to  $p$ . We will use  $x_i$  to denote both the pixel and its value 0 or 1,  $x_i$  is called white or black, accordingly. In the thinned image, the pixel  $p$  is examined for termination point is detected if  $T$  (summation of neighbors of the central pixel  $p$ , where  $p$  is a white pixel and its value equal to 1) is equal to 1 as shown in in the following equation:

$$T = \sum_{i=0}^7 x_i = 1 \quad (1)$$

### IV. CLASSIFICATION

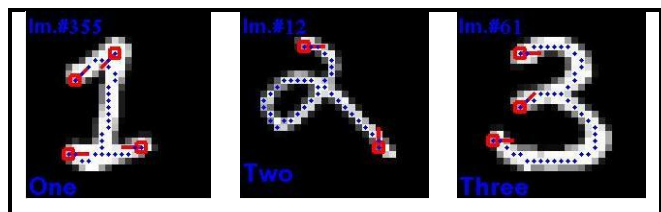
After the features extraction in the previous section, the classification process in our recognition system is divided into two stages. In the first stage, mainly counting the number of termination points, the digits are classified into six categories. The relation between the number of the termination points and possible digits is shown in Table I. In the second stage, other features such as freeman chain code, orientation, positions, distances and others, all the digits are recognized.

TABLE I. THE RELATION BETWEEN NUMBER OF TERMINATION POINTS AND DIGITS

Number of termination points	Possible digits									
	0	1	2	3	4	5	6	7	8	9
0	✓								✓	
1	✓		✓				✓		✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3		✓	✓	✓	✓	✓		✓		✓
4		✓	✓	✓	✓	✓		✓		
5					✓			✓		
>5	Rejected									

The implementation of the above mentioned two-staged classification is depicted in Figure 3. We can obviously notice the following:

1. The result of the thinned operation appears as blue colored dots,
2. Each termination point is bounded with a red square color including its orientation information,
3. The sequence of each digit image as exists in its test set appears on the top-left corner with blue color, and
4. The classification result appears on the bottom-left corner with blue color.



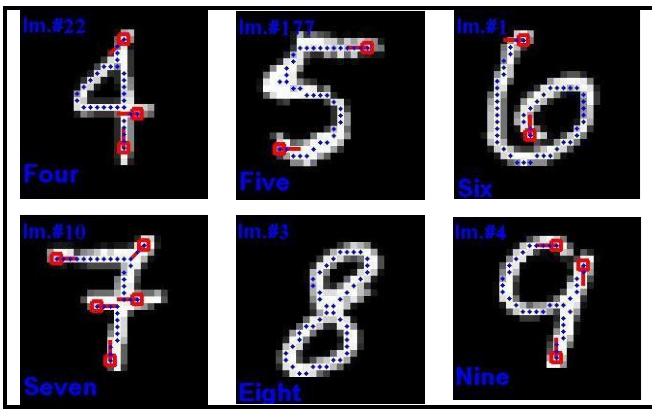


Figure 3. The implementation of the two-staged classification process

### V. EXPERIMENTAL RESULTS

Our experiments are performed on MNIST digit database by using MATLAB 6.5.1 release 13. This database consists of 10,000 gray-scale digit patterns in testing set and 60,000 gray-scale digit patterns in training set. The size of each digit image is 28x28 pixels. All results are obtained by using 2.40 GHz P4 processor under Windows XP. The processing time of each single digit is around 0.488 s.

The results of our recognition system tested on the testing set are summarized in Table II and on the training set are shown in Table III. It's shown that for each class, there exists a true positive (TP) means that the digit is correctly classified, false positive (FP) means that the digit is wrongly classified, rejected (RJ) means that the digit is not classified or unknown and reliability (RL) is calculated as:

$$RL(\%) = \frac{TP(\%)}{100 - RJ(\%)} * 100 \quad (2)$$

TABLE II. THE PERFORMANCE OF THE CLASSIFIERS ON THE TESTING SET

#	No. of Images	True Positive (TP%)	False Positive (FP%)	Rejected (RJ%)	Reliability (RL%)
0	980	95.5102	2.8571	1.6327	97.0955
1	1135	97.0925	2.5551	0.3524	97.4359
2	1032	92.3450	4.6512	3.0039	95.2049
3	1010	94.3564	4.4554	1.1881	95.4909
4	982	94.3992	4.5825	1.0183	95.3704
5	592	90.5405	6.9257	2.5338	92.8943
6	958	91.0230	6.0543	2.9228	93.7635
7	1028	91.1479	6.4202	2.4319	93.4198
8	974	90.5544	8.9322	0.5133	91.0216
9	1009	92.0714	4.9554	2.9732	94.8928
<b>Overall Performance %</b>		<b>92.9041</b>	<b>5.2389</b>	<b>1.8570</b>	<b>94.6589</b>

TABLE III. THE PERFORMANCE OF THE CLASSIFIERS ON THE TRAINING SET

#	True positive		False positive		Rejection		Reliability
	No. of Images	TP%	No. of Images	FP%	No. of Images	RJ%	RL%
0	5698	96.2175	159	2.6849	65	1.6327	97.2853
1	6498	97.7584	113	1.7000	36	0.3524	98.2907
2	5803	94.2658	246	3.9799	108	3.0039	95.9491
3	5918	94.9615	237	3.8030	77	1.1881	96.1495
4	6433	95.2332	259	3.8342	63	1.0183	96.1297
5	5126	93.0309	253	4.5917	131	2.5338	95.2965
6	7810	94.7011	301	3.6498	136	2.9228	96.2890
7	5150	94.8260	196	3.6089	85	2.4319	96.3337
8	3235	95.2872	119	3.5052	41	0.5133	96.4520
9	5401	94.6713	211	3.6985	93	2.9732	96.2402
<b>Σ</b>	<b>57072</b>		<b>2093</b>		<b>835</b>		

Overall Performance %	<b>95.0953</b>	<b>3.5056</b>	<b>1.3991</b>	<b>96.4416</b>
-----------------------	----------------	---------------	---------------	----------------

As shown in the above tables, the overall performance for Latin handwritten digit recognition is 92.9041% and 95.0953, consecutively.

### VI. CONCLUSION AND FUTURE WORK

In this paper, we have described a simple approach for feature extraction and classification for Latin handwritten recognition. Five feature sets are obtained from the thinned digit image using the concept of the freeman chain code template. A two-staged classification is implemented. Firstly, counting the number of termination points, the digits are classified into six categories. The relation between the number of the termination points and possible digits is easily detected. Secondly, other features such as freeman chain code, orientation, positions, distances and others are calculated. The overall performance is 92.9041% for the recognition of 10,000 digit images in the test set and 95.0953% for the recognition of 60,000 digit images in the training set.

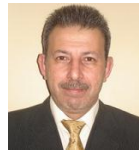
In the future work, a tapped operation for disconnected digits will be tested in the preprocess stage. Furthermore, we will test the above mentioned approach in the classification of the formation of the characters in the Arabic language.

### REFERENCES

- [1] D. Gorgevik, D. Cakmakov and V. Radevski, Handwritten Digit Recognition by Combining Support Vector Machines Using Rule-Based Reasoning, *Proc. of the 23<sup>rd</sup> International Conference on Information Technology Interfaces*. **19**(12) (2001) 139–144.
- [2] D. Gorgevik and D. Cakmakov, Combining SVM Classifiers for Handwritten Digit Recognition, *Proc. of the 16<sup>th</sup> International Conference on Pattern Recognition*. **3** (2002) 102–105.

- [3] S. Maji and J. Malik, Fast and Accurate Digit Classification, *Technical Report No. UCB/EECS-2009-159*. (2009).
- [4] Z. Chi, Z. Lu and F-H. Chan, Multi-Channel Handwritten Digit Recognition Using Neural Networks, *Proc. of IEEE International Symposium on Circuits and Systems*. **1**(1997) 625–628.
- [5] B. Duerr, W. Haettich, H. Tropsch and G. Winkler, A combination of statistical and syntactical pattern recognition applied to classification of unconstrained handwritten numerals, *Pattern Recognition*. **12** (1980) 189–199.
- [6] F. Shah and K. Yousaf, Handwritten Digit Recognition Using Image Processing and Neural Networks, *Proceedings of the World Congress on Engineering*. **1** (2007).
- [7] E. Kussul and T. Baidyk, Improved method of handwritten digit recognition tested on MNIST database, *Image and Vision Computing*. **22**(12) (2004) 971-981.
- [8] A. Jain and D. Zongker, Representation and Recognition of Handwritten Digits Using Deformable Templates, *IEEE Trans. Pattern Analysis and Machine Intelligence*. **19**(19) (1997) 1386–1390.
- [9] V. Mane and L. Ragma, Handwritten digit recognition using elastic matching, *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*. (2011), pp. 1364.
- [10] D. Gorgevik and D. Cakmakov, D, An efficient Three-Stage Classifier for Handwritten Digit Recognition, *Proc. of the 17<sup>th</sup> International Conference on Pattern Recognition*. **4** (2004) 507–510.
- [11] A. Bellili, M. Gilloux and P. Gallinari, An Hybrid MLP-SVM handwritten digit recognizer, *Proc. of the 6<sup>th</sup> International Conference on Document Analysis and Recognition*. (2001), pp. 28–32.
- [12] C. Pereira and G. Cavalcanti, Prototype Selection for Handwritten Connected Digits Classification, *The 10<sup>th</sup> International Conference on Document Analysis and Recognition*. (2009), pp. 1021-1025.
- [13] R. Ebrahimpour and S. Hamed, Hand Written Digit Recognition by Multiple Classifier Fusion based on Decision Templates Approach, *In Proc. of the International Conference on Computer, Electrical, Systems Science, and Engineering (CESSE)*. (2009), pp. 245-250.
- [14] L. Cheng-Lin, N. Kazuki, H. Sako and H. Fujisawa, Handwritten Digit Recognition: Benchmarking of State-of-art Techniques, *Pattern Recognition*. **36** (2003) 2271-2285.
- [15] MNIST Database of Handwritten digits: <http://yann.lecun.com/exdb/mnist/>.
- [16] L. Lam, S-W. Lee and C. Suen, Thinning methodologies-a comprehensive survey, *IEEE Trans. Pattern Analysis and Machine Intelligence*. **14**(9) (1992) 869–885.
- [17] H. Freeman, On the Encoding of Arbitrary Geometric Configurations, *IRE Trans. Pattern Analysis and Machine Intelligence Electronics and Computers*. **10**(1961) 260–268.
- [18] Kekre, H. B. (2010). Texture Based Segmentation using Statistical Properties for Mammographic Images. *IJACSA - International Journal of Advanced Computer Science and Applications*, *1*(5), 102-107.
- [19] Malhotra, R. (2011). Software Effort Prediction using Statistical and Machine Learning Methods. *IJACSA - International Journal of Advanced Computer Science and Applications*, *2*(1), 145-152.

#### AUTHOR PROFILE



**Ihab Zaqout** received the B. S. in Computer Science from the University of Al-Fateh, Libya, in 1987 and the M.S. degree in Computer Science from Jordan University, Jordan in 2000 and the Ph.D. in Computer Science from the University of Malaya, Malaysia in 2006. He is currently at the Dept. of Information Technology, Al-Azhar University - Gaza, Palestine as an assistant professor. His main research interests include image processing, pattern recognition, data mining and machine learning.

# Plant Leaf Recognition using Shape based Features and Neural Network classifiers

Jyotismita Chaki  
School of Education Technology  
Jadavpur University  
Kolkata, India

Ranjan Parekh  
School of Education Technology  
Jadavpur University  
Kolkata, India

**Abstract**—This paper proposes an automated system for recognizing plant species based on leaf images. Plant leaf images corresponding to three plant types, are analyzed using two different shape modeling techniques, the first based on the Moments-Invariant (M-I) model and the second on the Centroid-Radii (C-R) model. For the M-I model the first four normalized central moments have been considered and studied in various combinations viz. individually, in joint 2-D and 3-D feature spaces for producing optimum results. For the C-R model an edge detector has been used to identify the boundary of the leaf shape and 36 radii at 10 degree angular separation have been used to build the feature vector. To further improve the accuracy, a hybrid set of features involving both the M-I and C-R models has been generated and explored to find whether the combination feature vector can lead to better performance. Neural networks are used as classifiers for discrimination. The data set consists of 180 images divided into three classes with 60 images each. Accuracies ranging from 90%-100% are obtained which are comparable to the best figures reported in extant literature.

**Keywords**—plant recognition; moment invariants; centroid-radii model; neural network; computer vision.

## I. INTRODUCTION

Plants play an important role in our environment. Without plants there will be no existence of the earth's ecology. But in recent days, many types of plants are at the risk of extinction. To protect plants and to catalogue various types of flora diversities, a plant database is an important step towards conservation of earth's biosphere. There are a huge number of plant species worldwide. To handle such volumes of information, development of a quick and efficient classification method has become an area of active research. In addition to the conservation aspect, recognition of plants is also necessary to utilize their medicinal properties and using them as sources of alternative energy sources like bio-fuel. There are several ways to recognize a plant, like flower, root, leaf, fruit etc. In recent times computer vision methodologies and pattern recognition techniques have been applied towards automated procedures of plant recognition.

The present paper proposes a scheme for automated recognition of three types of plant species by analyzing shape features from digital images of their leaves. The organization of the paper is as follows: section 2 provides an overview of related work, section 3 outlines the proposed approach with discussions on overview, feature computation and classification

schemes, section 4 provides details of the dataset and experimental results obtained and section 5 provides the overall conclusion and the scope for future research.

## II. PREVIOUS WORK

Many methodologies have been proposed to analyze plant leaves in an automated fashion. A large percentage of such works utilize shape recognition techniques to model and represent the contour shapes of leaves, however additionally, color and texture of leaves have also been taken into consideration to improve recognition accuracies. One of the earliest works [1] employs geometrical parameters like area, perimeter, maximum length, maximum width, elongation to differentiate between four types of rice grains, with accuracies around 95%. Use of statistical discriminant analysis along with color based clustering and neural networks have been used in [2] for classification of a flowered plant and a cactus plant. In [3] the authors use the Curvature Scale Space (CSS) technique and k-NN classifiers to classify chrysanthemum leaves. Both color and geometrical features have been reported in [4] to detect weeds in crop fields employing k-NN classifiers. In [5] the authors propose a hierarchical technique of representing leaf shapes by first their polygonal approximations and then introducing more and more local details in subsequent steps. Fuzzy logic decision making has been utilized in [6] to detect weeds in an agricultural field. In [7] the authors propose a two-step approach of using a shape characterization function called centroid-contour distance curve and the object eccentricity for leaf image retrieval. The centroid-contour distance (CCD) curve and eccentricity along with an angle code histogram (ACH) have been used in [8] for plant recognition. The effectiveness of using fractal dimensions in describing leaf shapes has been explored in [9].

In contrast to contour-based methods, region-based shape recognition techniques have been used in [10] for leaf image classification. Elliptic Fourier harmonic functions have been used to recognize leaf shapes in [11] along with principal component analysis for selecting the best Fourier coefficients. In [12] the authors propose a leaf image retrieval scheme based on leaf venation for leaf categorization. Leaf venations are represented using points selected by the curvature scale scope corner detection method on the venation image and categorized by calculating the density of feature points using non parametric estimation density. In [13] 12 leaf features are extracted and orthogonalized into 5 principal variables which

consist of the input vector of a neural network (NN). The NN is trained by 1800 leaves to classify 32 kinds of plants with accuracy greater than 90%. NNs have also been used in [14] to classify plant based on parameters like size, radius, perimeter, solidity and eccentricity. An accuracy of about 80% is reported. Wavelet and fractal based features have been used in [15] to model the uneven shapes of leaves. Texture features along with shape identifiers have been used in [16] to improve recognition accuracies. Other techniques like Zernike moments and Polar Fourier Transform have also been proposed [17] for modeling leaf structures. An accuracy of 64% has been reported. In [18] authors propose Hybrid Image Segmentation Algorithm for Leaf Recognition and Characterization.

A new approach that combines a thresholding method and H-maxima transformation based method is proposed to extract the leaf veins. Compared with other methods, experimental results show that this combined approach is capable of extracting more accurate venation modality of the leaf for the subsequent vein pattern classification. In [19] authors propose Guiding Active Contours for Tree Leaf Segmentation and Identification. Combining global shape descriptors given by the polygonal model with local curvature-based features, the leaves are classified over nearly 50 tree species. Finally in [20] a combination of all image features viz. color, texture and shape, have been used for leaf image retrieval, with a reported accuracy of 97.9%.

### III. PROPOSED APPROACH

The present paper proposes a scheme for automated detection of 3 classes of plant category by analyzing shapes obtained from a collection of their leaf images, using features based on Moment-Invariants and Centroid-Radii approaches, with various types of neural network classifiers:

#### A. Moment Invariants (M-I)

M-K Hu [21] proposes 7 moment features that can be used to describe shapes and these are invariant to rotation, translation and scaling. For a digital image, the moment of a pixel  $P(x, y)$  at location  $(x, y)$  is defined as the product of the pixel value with its coordinate distances i.e.  $m = x.y.P(x, y)$ . The moment of the entire image is the summation of the moments of all its pixels. More generally the moment of order  $(p, q)$  of an image  $I(x, y)$  is given by

$$m_{pq} = \sum_x \sum_y [x^p y^q I(x, y)] \quad (1)$$

Based on the values of  $p$  and  $q$  the following are defined :

$$\begin{aligned} m_{00} &= \sum_x \sum_y [x^0 y^0 I(x, y)] = \sum_x \sum_y [I(x, y)] \\ m_{10} &= \sum_x \sum_y [x^1 y^0 I(x, y)] = \sum_x \sum_y [x I(x, y)] \\ m_{01} &= \sum_x \sum_y [x^0 y^1 I(x, y)] = \sum_x \sum_y [y I(x, y)] \\ m_{11} &= \sum_x \sum_y [x^1 y^1 I(x, y)] = \sum_x \sum_y [xy I(x, y)] \\ m_{20} &= \sum_x \sum_y [x^2 y^0 I(x, y)] = \sum_x \sum_y [x^2 I(x, y)] \\ m_{02} &= \sum_x \sum_y [x^0 y^2 I(x, y)] = \sum_x \sum_y [y^2 I(x, y)] \\ m_{21} &= \sum_x \sum_y [x^2 y^1 I(x, y)] = \sum_x \sum_y [x^2 y I(x, y)] \\ m_{12} &= \sum_x \sum_y [x^1 y^2 I(x, y)] = \sum_x \sum_y [xy^2 I(x, y)] \\ m_{30} &= \sum_x \sum_y [x^3 y^0 I(x, y)] = \sum_x \sum_y [x^3 I(x, y)] \\ m_{03} &= \sum_x \sum_y [x^0 y^3 I(x, y)] = \sum_x \sum_y [y^3 I(x, y)] \end{aligned} \quad (2)$$

The first four Hu invariant moments which are invariant to rotation are defined as follows

$$\begin{aligned} \phi_1 &= m_{20} + m_{02} \\ \phi_2 &= (m_{20} - m_{02})^2 + (2m_{11})^2 \\ \phi_3 &= (m_{30} - 3m_{12})^2 + (3m_{21} - m_{03})^2 \\ \phi_4 &= (m_{30} + m_{12})^2 + (m_{21} + m_{03})^2 \end{aligned} \quad (3)$$

To make the moments invariant to translation the image is shifted such that its centroid coincides with the origin of the coordinate system. The centroid of the image in terms of the moments is given by:

$$\begin{aligned} x_c &= \frac{m_{10}}{m_{00}} \\ y_c &= \frac{m_{01}}{m_{00}} \end{aligned} \quad (4)$$

Then the central moments are defined as follows:

$$\mu_{pq} = \sum_x \sum_y [(x - x_c)^p (y - y_c)^q I(x, y)] \quad (5)$$

To compute Hu moments using central moments the  $\phi$  terms in equation (2) need to be replaced by  $\mu$  terms. It can be verified that  $\mu_{00} = m_{00}$ ,  $\mu_{10} = 0 = \mu_{01}$ .

To make the moments invariant to scaling the moments are normalized by dividing by a power of  $\mu_{00}$ . The *normalized central moments* are defined as below

$$M_{pq} = \frac{\mu_{pq}}{(\mu_{00})^\omega}, \text{ where } \omega = 1 + \frac{p+q}{2} \quad (6)$$

### B. Centroid-Radii model (C-R)

In [22] K. L. Tan et al. proposes the centroid-radii model for estimating shapes of objects in images. A shape is defined to be an area of black on a background of white. Each pixel is represented by its color (black or white) and its x-y coordinates on the canvas. The centroid is located at the position (Cx, Cy) which are respectively, the average of the x and y coordinates for all black pixels. The boundary of a shape consists of a series of boundary points. A boundary point is a black pixel with a white pixel as its neighbor. A radius is a straight line joining the centroid to a boundary point. In the centroid-radii model, lengths of a shape's radii from its centroid at regular intervals are captured as the shape's descriptor using the Euclidean distance. More formally, let  $\theta$  be the regular interval (measured in degrees) between radii (Figure 1). Then, the number of intervals is given by  $k = 360/\theta$ . All radii lengths are normalized by dividing with the longest radius length from the set of radii lengths extracted. Furthermore, without loss of generality, suppose that the intervals are taken clockwise starting from the x-axis direction. Then, the shape descriptor can be represented as a vector :

$$S = \{r_0, r_\theta, r_{2\theta}, \dots, r_{(k-1)\theta}\} \quad (7)$$

Here  $r_{i\theta}$ ,  $0 \leq i \leq (k-1)$  is the  $(i+1)$ -th radius from the centroid to the boundary of the shape. With sufficient number of radii, dissimilar shapes can be differentiated from each other.

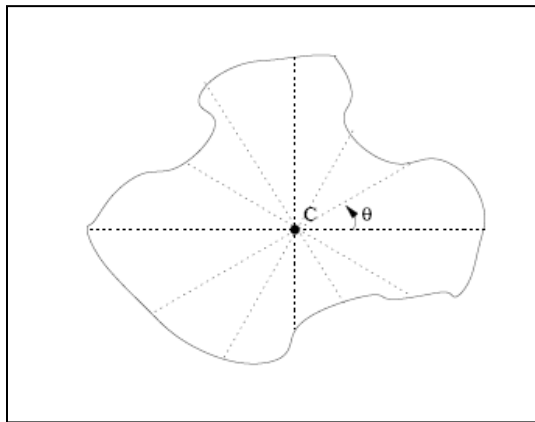


Figure 1. Centroid-radii approach

## IV. EXPERIMENTATIONS AND RESULTS

Experimentations are performed by using 180 leaf images from the Plantscan database available at the URL: [http://imedia-ftp.inria.fr:50012/PI@ntNet/plantscan\\_v2/](http://imedia-ftp.inria.fr:50012/PI@ntNet/plantscan_v2/). The dataset is divided into 3 classes A (Pittosporum Tobira), B (Betula Pendula), C (Cercis Siliquastrum) each consisting of 60 images. Each image is 350 by 350 pixels in dimensions and in JPG format. A total of 90 images are used as the Training set (T) and the remaining 90 images as the Testing set (S). Sample images of each class are shown below.

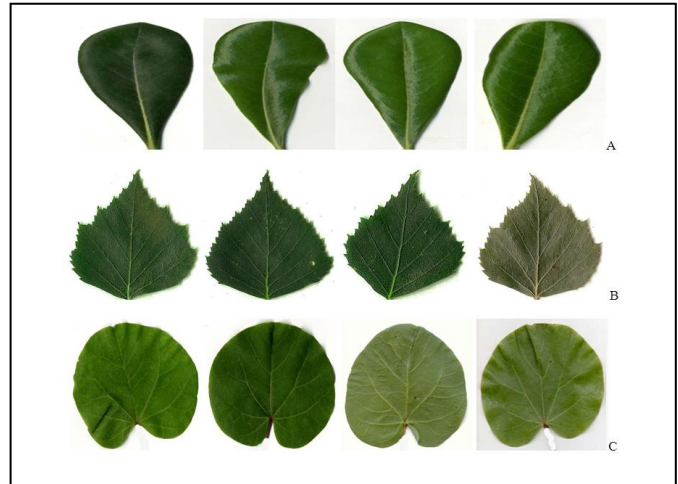


Figure 2. Samples of leaf images belonging to 3 classes

For computing recognition rates, comparisons between training and testing sets are done using neural network (Multi-layer perceptrons : MLP) with feed-forward back-propagation architectures. The legends used in this work are listed in Table I.

TABLE I. LEGENDS

Legend	Meaning	Legend	Meaning
AT	Class-A Training set	AS	Class-A Testing set
BT	Class-B Training set	BS	Class-B Testing set
CT	Class-C Training set	CS	Class-C Testing set

### A. Moment-Invariants (M-I) Representations

The first 4 normalized central moments M1, M2, M3, M4 of each image of the training and testing sets were computed as per equation (6). Various combination of the features in individual and joint configurations were fed to neural network (NN) classifiers to study which combinations produce the best results

#### 1) Individual Features:

Individual features values M1, M2, M3, M4 for the training and testing images for the 3 classes are first used. Results are summarized in Table II below. The first column depicts the feature used, the second column shows the neural network configuration (NNC) viz. 1-3-3 indicates 1 input unit (for the individual feature), 3 units in the hidden layer and 3 units in the output layer (corresponding to the 3 classes to be distinguished).



The third, fourth and fifth columns indicate the percentage recognition accuracies for the three classes, the sixth column provides the overall accuracy for the three classes and the last column indicates the best Mean Square Error (MSE) obtained during the training phase of the NNs.

TABLE II. ACCURACY USING INDIVIDUAL M-I FEATURES

F	NNC	A	B	C	O	MSE
M <sub>1</sub>	1-3-3	70 (21/30)	96.7 (29/30)	100 (30/30)	88.9 (80/90)	0.06
M <sub>2</sub>	1-3-3	73.3	86.7	20	60	0.17
M <sub>3</sub>	1-3-3	10	100	70	60	0.15
M <sub>4</sub>	1-3-3	86.7	46.7	50	61.1	0.16

Table II indicates that of the individual moment values M<sub>1</sub> provides the best results of 88.9 %. The number of images (out of 30) correctly identified, are also indicated in parenthesis. The corresponding NN output for the 3 classes, is shown below in Figure 3. The class files are arranged sequentially i.e. the first 30 files belong to Class-A the next 30 to Class-B and the last 30 to Class-C.

2) Joint Features:

To improve upon the results obtained using individual features, joint features are next considered in 2-D feature spaces i.e. M<sub>1</sub>-M<sub>2</sub>, M<sub>1</sub>-M<sub>3</sub>, M<sub>1</sub>-M<sub>4</sub>. M<sub>1</sub> is kept common since it is seen to produce the best accuracies. Results are summarized in Table III.

TABLE III. ACCURACY USING JOINT 2-D M-I FEATURES

F	NNC	A	B	C	O	MSE
M <sub>1</sub> -M <sub>2</sub>	2-3-3	90	100	93.3	94.4	0.058
M <sub>1</sub> -M <sub>3</sub>	2-3-3	90 (27/30)	96.6 (29/30)	100 (27/30)	95.5 (86/90)	0.053
M <sub>1</sub> -M <sub>4</sub>	2-3-3	86.6	96.6	93.3	92.2	0.068

Table III indicates that of the joint 2-D feature values M<sub>1</sub>-M<sub>3</sub> provides the best results of 95.5 %. The corresponding NN output for the 3 classes, is shown below in Figure 4.

To conclude analysis of joint spaces, features are also considered in 3-D feature spaces : M<sub>1</sub>-M<sub>2</sub>-M<sub>3</sub>, M<sub>1</sub>-M<sub>2</sub>-M<sub>4</sub>, M<sub>1</sub>-M<sub>3</sub>-M<sub>4</sub>. Results are summarized on Table IV below.

TABLE IV. ACCURACY USING JOINT 3-D M-I FEATURES

F	NNC	A	B	C	O	MSE
M <sub>1</sub> -M <sub>2</sub> -M <sub>3</sub>	3-30-3	83.3	100	93.3	92.2	0.028
M <sub>1</sub> -M <sub>2</sub> -M <sub>4</sub>	3-30-3	90	86.7	90	88.9	0.005
M <sub>1</sub> -M <sub>3</sub> -M <sub>4</sub>	3-30-3	90 (27/30)	100 (30/30)	90 (27/30)	93.3 (84/90)	0.005

Table IV indicates that of the joint 3-D feature values M<sub>1</sub>-M<sub>3</sub>-M<sub>4</sub> provides the best results of 93.3 %. The corresponding NN output for the 3 classes, is shown in Figure 5.

Figures 6 and 7 shows the variation of the best features M<sub>1</sub>, (M<sub>1</sub>-M<sub>3</sub>), (M<sub>1</sub>-M<sub>3</sub>-M<sub>4</sub>) over the training and testing files for the 3 classes.

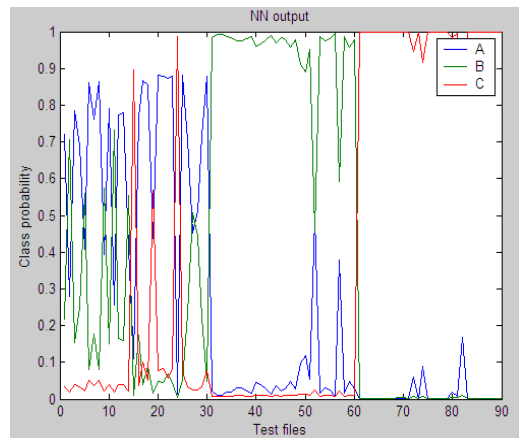


Figure 3. NN output for M<sub>1</sub>

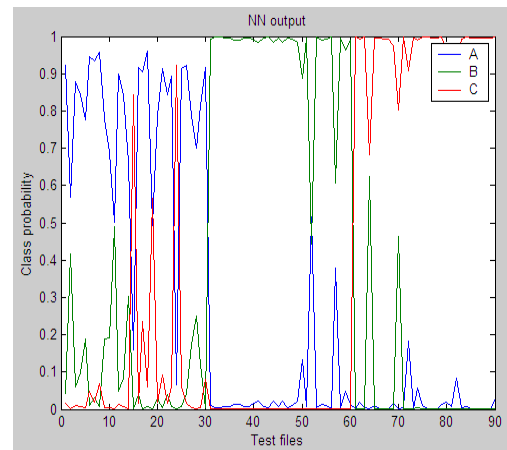


Figure 4. NN output for M<sub>1</sub>-M<sub>3</sub>

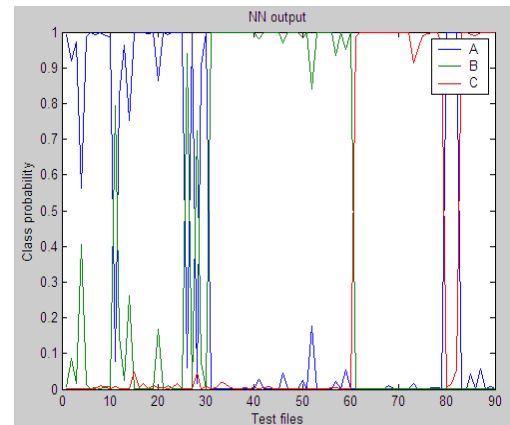


Figure 5. NN output for M<sub>1</sub>-M<sub>3</sub>-M<sub>4</sub>

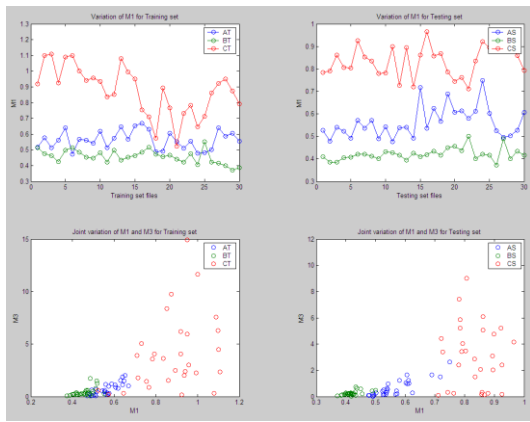


Figure 6. Variation of individual  $M_1$  and  $M_1$ - $M_3$  features in 2-D space

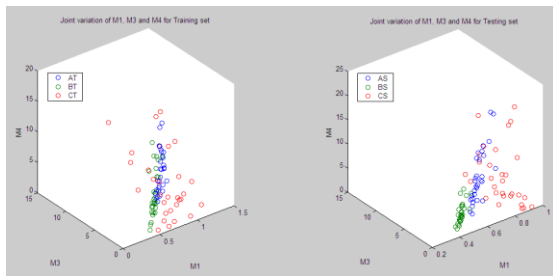


Figure 7. Variation of  $M_1$ - $M_3$ - $M_4$  features in 3-D space

### B. Centroid-Radii (C-R) Representations

Each image is converted to binary form and the Canny edge detector is used to identify its contour. Its centroid is computed from the average of its edge pixels.

Corresponding to each edge pixel the angle it subtends at the centroid is calculated and stored in an array along with its x- and y- coordinate values. From the array 36 coordinate values of edge pixels which join the centroid at 10 degree intervals from 0 to 359 degrees are identified. The radii length of joining these 36 points with the centroid are calculated using the Euclidean distance and the radii lengths are normalized to the range [0,1]. For each leaf image 36 such normalized lengths are stored in an ordered sequence. Figure 8 shows a visual representation of a leaf image, the edge detected version, the location of the centroid and edge pixels, and the normalized radii vector. The average of the 36 radii lengths for each image of each class both for the training and testing sets, is plotted in Figure 9, which depicts the overall feature range and variation for each class.

Classes are discriminated using NN. The results are summarized in Table V. An overall accuracy of 100% is achieved. The NN convergence plot and output are shown in Figure 10. Convergence takes place in 38280 epochs with an MSE of 0.005.

TABLE V. ACCURACY USING C-R FEATURES

F	NNC	A	B	C	O	MSE
36-element C-R vector	36-30-3	100 (30/30)	100 (30/30)	100 (30/30)	100 (90/90)	0.005

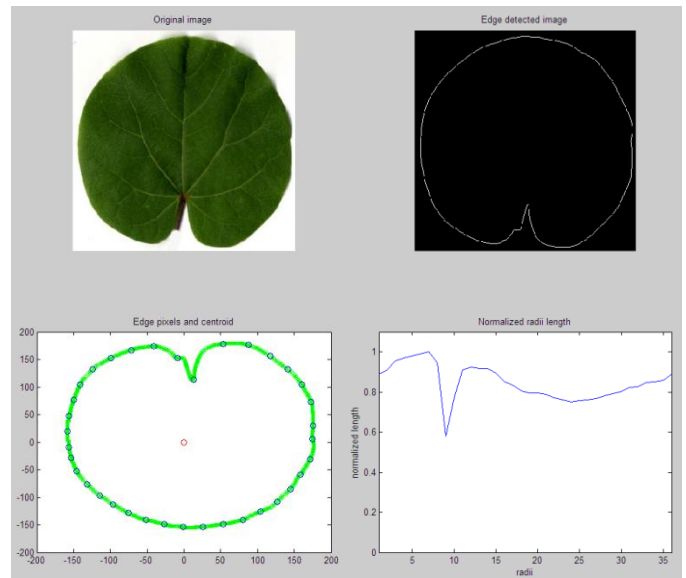


Figure 8. Interface for C-R computations

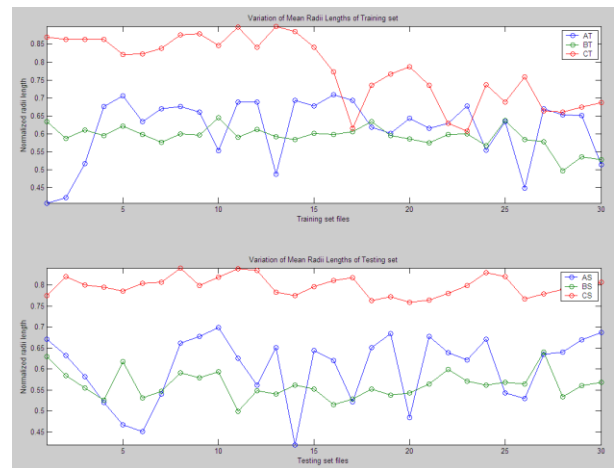


Figure 9. Variation of mean radii length for 3 classes

### C. Hybrid Representation

To study recognition accuracy of a hybrid set of features, two combinations of C-R and M-I features were used : (1) 38-element vector : 36-elements from C-R model and 2 elements ( $M_1$ ,  $M_3$ ) from M-I model (2) 39-element vector : 36-elements from C-R model and 3 elements ( $M_1$ ,  $M_3$ ,  $M_4$ ) from M-I model. The combination vectors are fed to an appropriate NN and results in each case are summarized below in Table VI.

TABLE VI. ACCURACY USING HYBRID FEATURES

F	NNC	A	B	C	O	MSE
38-element hybrid vector	38-30-3	100 (30/30)	100 (30/30)	100 (30/30)	100 (90/90)	0.005
39-element hybrid vector	39-30-3	100	100	93.3	97.7	0.005

The overall accuracy is observed to be 100% using the 38-element hybrid vector, but convergence takes place much quicker requiring only 20706 epochs for an MSE of 0.005, in contrast to 38280 epochs when using only the C-R vector. The convergence plot and NN output is shown in Figure 11.

Regarding the system implementation, computation of 36-radii vector for 30 images takes around 35 seconds, while computation of the first four moments for 30 images takes around 300 seconds, on a 2.66 GHz P4 system with 1 GB RAM. NN convergence for the 38-element hybrid vector takes around 2 minutes for a training set of 30 images. The system has been observed to be stable for all the 3 categories of the leaf images displaying almost the same timings over the entire dataset.

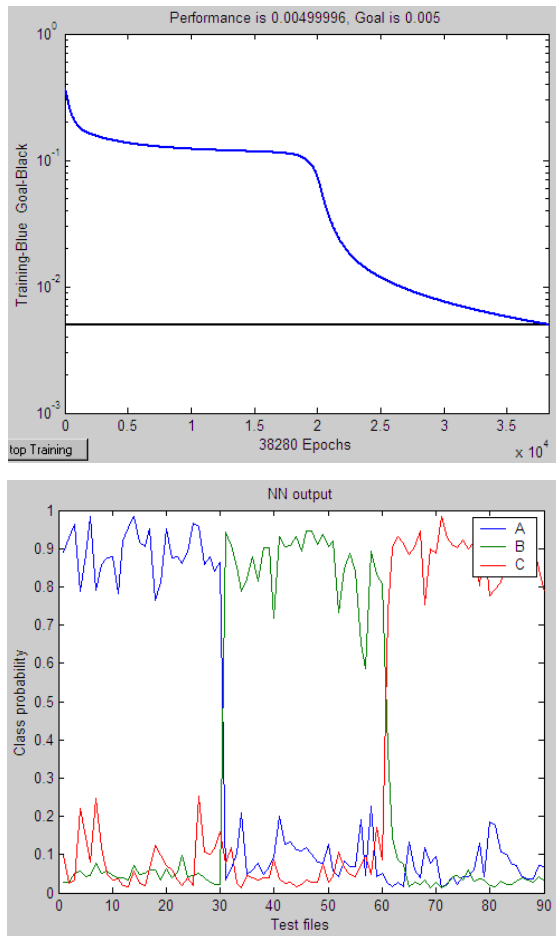


Figure 10. NN convergence and output for C-R vector

### V. ANALYSIS

Automated discrimination between three leaf shapes was done using a variety of approaches to find the optimum results. The study reveals that for M-I approach joint features in general provide better accuracies than individual features. Accuracies based on C-R method are better than individual M-I features and comparable to joint M-I features.

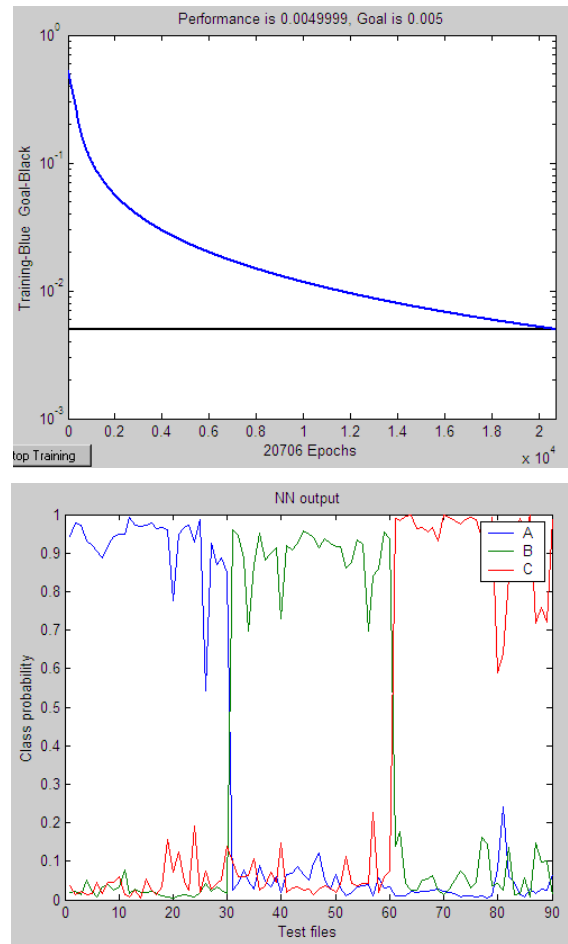


Figure 11. NN convergence and output for 38-e hybrid vector

Performance was found to improve when C-R features were combined with M-I features in a hybrid feature space, the accuracy of 100% being same as that for only the C-R vector, but requiring almost half the number of epochs. Accuracy results obtained using different methods are summarized in Table VII.

TABLE VII. ACCURACY USING VARIOUS FEATURES

F	M-I1	M-I J2	M-I J3	C-R	H
% Accuracy	88.9	95.5	93.3	100	100

To put the above results in perspective with the state of the art, the best results reported in [8] is a recall rate of 60% for discrimination of chrysanthemum leaves from a database of 1400 color images. Accuracy for classification for 10 leaf categories over 600 images is reported to be 82.33% in [10]. Overall classification accuracy reported in [11] for 4 categories of leaf images obtained during three weeks of germination, is around 90%. Accuracy reported in [13] for classification of 32 leaf types from a collection of 1800 images is around 90%. An overall classification of 80% is reported in [14] for identifying two types of leaf shapes from images taken using different frequency bands of the spectrum.

Best accuracies reported in [17] are around 93% using Polar Fourier Transforms. Results reported in [19] are in the region of 80% for classifying 50 species. Accuracies of around 97% have been reported in [20] for a database of 500 images. It therefore can be said that the accuracies reported in the current paper are comparable to the best results reported in extant literature. It may however be noted that in many of the above cases color and geometrical parameters have also been combined with shape based features to improve results, while the current work is based solely on shape characteristics.

## VI. CONCLUSIONS AND FUTURE SCOPES

This paper proposes an automated system for plant identification using shape features of their leaves. Two shape modeling approaches are discussed: one technique based on invariant-moments model and the other on centroid-radii model, and the two are compared with regard to classification accuracies. Such automated classification systems can prove extremely useful for quick and efficient classification of plant species. The accuracy of the current proposed approach is comparable to those reported in contemporary works. A salient feature of the current approach is the low-complexity data modeling scheme used whereby dimensionality of the feature vectors are typically below 40.

Future work would involve research along two directions: (1) combining other shape based techniques like Hough transform and Fourier descriptors, and (2) combining color and texture features along with shape features for improving recognition accuracies.

## REFERENCES

- [1] N. Sakai, S. Yonekawa, and A. Matsuzaki, "Two-dimensional image analysis of the shape of rice and its applications to separating varieties", *Journal of Food Engineering*, vol 27, 1996, pp. 397-407.
- [2] A. J. M. Timmermans, and A. A. Hulzebosch, "Computer vision system for on-line sorting of pot plants using an artificial neural network classifier", *Computers and Electronics in Agriculture*, vol. 15, 1996, pp. 41-55.
- [3] S. Abbasi, F. Mokhtarian, and J. Kittler, "Reliable classification of chrysanthemum leaves through curvature scale space", *Lecture Notes in Computer Science*, vol. 1252, 1997, pp. 284-295.
- [4] A. J. Perez, F. Lopez, J. V. Benlloch, and S. Christensen, "Color and shape analysis techniques for weed detection in cereal fields", *Computers and Electronics in Agriculture*, vol. 25, 2000, pp. 197-212.
- [5] C. Im, H. Nishida, and T. L. Kunii, "A hierarchical method of recognizing plant species by leaf shapes", *IAPR Workshop on Machine Vision Applications*, 1998, pp. 158-161.
- [6] C-C Yang, S. O. Prasher, J-A Landry, J. Perret, and H. S. Ramaswamy, "Recognition of weeds with image processing and their use with fuzzy logic for precision farming", *Canadian Agricultural Engineering*, vol. 42, no. 4, 2000, pp. 195-200.
- [7] Z. Wang, Z. Chi, D. Feng, and Q. Wang, "Leaf image retrieval with shape feature", *International Conference on Advances in Visual Information Systems (ACVIS)*, 2000, pp. 477-487.
- [8] Z. Wang, Z. Chi, and D. Feng, "Shape based leaf image retrieval", *IEEE Proceedings on Vision, Image and Signal Processing (VISIP)*, vol. 150, no.1, 2003, pp. 34-43.
- [9] J. J. Camarero, S. Siso, and E.G-Pelegrin, "Fractal dimension does not adequately describe the complexity of leaf margin in seedlings of *Quercus* species", *Anales del Jardín Botánico de Madrid*, vol. 60, no. 1, 2003, pp. 63-71.
- [10] C-L Lee, and S-Y Chen, "Classification of leaf images", *16<sup>th</sup> IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP)*, 2003, pp. 355-362.
- [11] J. C. Neto, G. E. Meyer, D. D. Jones, and A. K. Samal, "Plant species identification using elliptic Fourier leaf shape analysis", *Computers and Electronics in Agriculture*, vol. 50, 2006, pp. 121-134.
- [12] J-K Park, E-J Hwang, and Y. Nam, "A vention – based leaf image classification scheme", *Alliance of Information and Referral Systems*, 2006, pp. 416-428.
- [13] S. G. Wu, F. S. Bao, E. Y. Xu, Y-X Wang, Y-F Chang, and Q-L Xiang, "A leaf recognition algorithm for plant classification using probabilistic neural network", *The Computing Research Repository (CoRR)*, vol.1, 2007, pp. 11-16.
- [14] J. Pan, and Y. He, "Recognition of plants by leaves digital image and neural network", *International Conference on Computer Science and Software Engineering*, vol 4, 2008, pp. 906 – 910.
- [15] Q-P Wang, J-X Du, and C-M Zhai, "Recognition of leaf image based on ring projection wavelet fractal feature", *International Journal of Innovative Computing, Information and Control*, 2010, pp. 240-246.
- [16] T. Beghin, J. S. Cope, P. Remagnino, and S. Barman, "Shape and texture based plant leaf classification", *International Conference on Advanced Concepts for Intelligent Vision Systems (ACVIS)*, 2010, pp. 345-353.
- [17] A. Kadir, L.E. Nugroho, A. Susanto, and P.I. Santosa, "A comparative experiment of several shape methods in recognizing plants", *International Journal of Computer Science & Information Technology (IJCSIT)*, vol. 3, no. 3, 2011, pp. 256-263
- [18] N. Valliammal, and S. N. Geethalakshmi, "Hybrid image segmentation algorithm for leaf recognition and characterization", *International Conference on Process Automation, Control and Computing (PACC)*, 2011, pp. 1-6.
- [19] G. Cerutti, L. Tougne, J. Mille, A. Vacavant, and D. Coquin, "Guiding active contours for tree leaf segmentation and identification", *Cross-Language Evaluation Forum (CLEF)*, Amsterdam, Netherlands, 2011.
- [20] B. S. Bama, S. M. Valli, S. Raju, and V. A. Kumar, "Content based leaf image retrieval using shape, color and texture features", *Indian Journal of Computer Science and Engineering*, vol. 2, no. 2, 2011, pp. 202-211.
- [21] M-K Hu, "Visual pattern recognition by moment invariants", *IRE Transactions on Information Theory*, 1962, pp. 179-187.
- [22] K-L Tan, B. C. Ooi, and L. F. Thiang, "Retrieving similar shapes effectively and efficiently", *Multimedia Tools and Applications*, vol. 19, 2003, pp. 111-134

## AUTHORS PROFILE

Smt. Jyotismita Chaki is a Masters (M.Tech.) research scholar at the School of Education Technology, Jadavpur University, Kolkata, India. Her research interests include image processing and pattern recognition.

Dr. Ranjan Parekh is a faculty at the School of Education Technology, Jadavpur University, Kolkata, India. He is involved with teaching subjects related to multimedia technologies at the post-graduate level. His research interests include multimedia databases, pattern recognition, medical imaging and computer vision. He is the author of the book "Principles of Multimedia" published by McGraw-Hill, 2006.

# Integrated Information System for reserving rooms in Hotels

Dr. Safarini Osama  
IT Department  
University of Tabuk,  
Tabuk, KSA  
usama.safarini@gmail.com

**Abstract**— It is very important to build new and modern flexible dynamic effective compatible reusable information systems including database to help manipulate different processes and deal with many parts around it. One of these is managing the room reservations for groups and individual, to focus necessary needs in hotels and to be integrated with accounting system. This system provides many tools can be used in taking decision.

**Keywords:** IS - Information System; MIS -Management Information System; DFD - Data Flow Diagram; ER- Entity Relationship; DBMS - Database Management system.

## I. INTRODUCTION

Since tourism is now taking advantage in all over the world, so the role of hotels is coming greater than before, from this point of view our system gains its importance. We had followed the standards of building an information system [1], which is analysis, design, and implementation.

Hotels are big organizations that take place in the market, so using manual system may lead to more and more errors, because people forgot, and their thinking is limited, so we think that using a computerized system is more efficient, especially that we are living in information age.

Reservation is the most important part of any hotel system, this process should be efficient. So it can serve the commercial goals of the hotel. The most important entity in the Reservation process is the Rooms, reserving Rooms in a hotel should be dynamic, and the whole process should be done in an effective way to maximize the hotel profits.

We have two types of customers in the hotel; individual customers and group customers, and the person who is responsible for reservation is the front office manager. Group customers have higher priority over individual customers in high season, because group customers are more stable. So the front office manager should take decisions about the reservation, these decisions can be supported by the computerized system. This system provides many tools that the front office manager could use in taking his decision for allocating rooms and reserve for groups and taking right decisions.

Note that this system can't operate alone in the hotel, because it concentrates mainly in reservation process, the system needs an accounting system and a management

information system, and so it can serve all the needs of the hotel.

We have tried to make it possible to connect our system to other systems (Accounting, MIS) though two screens which are; Accounting screen which deals with entering bills for a customer, and recipient screen which deals with adding a recipient for the hotel. A very important part that we include in our system is the report generator, which enables the manager to have any information about the customers in the hotel. We tried to improve the reusability of our system through setting screen which lets the user change the main settings such as name of the hotel, colors of the background of the screens, and hotel logo.

### A. Similar Projects

#### 1) Fidelio Hotel System

One of the most famous hotels systems is Fidelio, which was developed by Microsoft; it can be considered as a complete hotel system, since it has a hotel accounting system and a management information system in addition to reservation system. It was developed in 1988, but in those days' computers was in its first stages to the commercial business and big organizations, so Fidelio wasn't famous at his first appearance. Over and over when computers take place in the market, big organizations such as hotels start using computerized systems, and then Fidelio starts to be famous.

Fidelio reservation system divides the reservation process into two types, Group and Individual reservation. Both the recipient and the reservation manager can reserve for any type. The reservation information and resident information are entered from the same screen.

When allocating rooms for customers in hotel, you can click on allocate button to show a small screen containing the empty rooms. In other words you can fill all information for customers in the same screen whether this customer was an individual customer or a group customer. As for accounting, you enter other screens to add the customer bills, this screen contains a list of valued facilities, and you can change the cost from this screen. There is no ability in Fidelio to change the color of the screen or the fonts, also the name of hotel in not shown in the main screen.

Fidelio Reservation System contains some static reports about the customers in hotel. These reports contain static data

about customers such as full name and room numbers. You cannot change it by removing or adding new fields.

The user can change his password and nickname from the system that the reservation manager gave it to him. The system doesn't provide the manager with any kind of statistics about the situation of the rooms in the hotel. A reservation manager screen contains two main components. The first is adding recipient information and assign him a shift and sub-manager property. The second component is adding new room to hotel and converts an existing room to maintenance state.

## II. DISCUSSION

The designed Information system begins with a login screen Fig.1

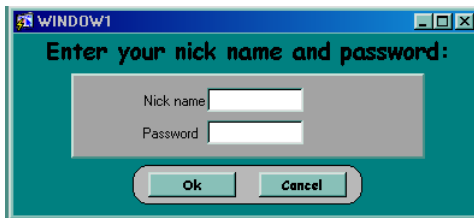


Fig.1 Login Screen

This screen is the first screen in the system which allows the authorized user to access the system. Our project is talking mainly about the reservation process, and components needed to connect the reservation system with other systems needed by the hotel.

First, we divide the reservation process into two types, which are group's reservation and individual's reservation, this division was done so the user will not be confused, and the information needed from each entity differ from the other.

The reservation process is provided with search capabilities, allocating rooms for the customer is one of the most important components of the system which enables the front office manager and recipient to allocate empty rooms for the customers. The Main menu is illustrated in fig.2

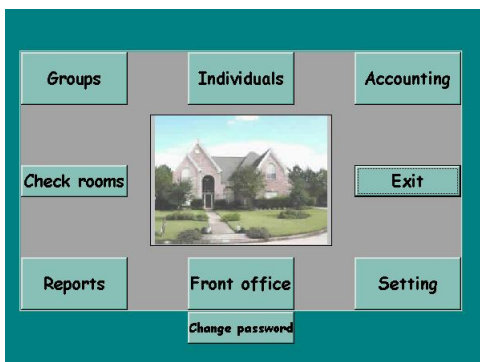


Fig.2 Main Screen

### A. The setting screen:

This screen allows the user to modify some settings such as the color and the hotel name and the hotel's logo (see fig3.)

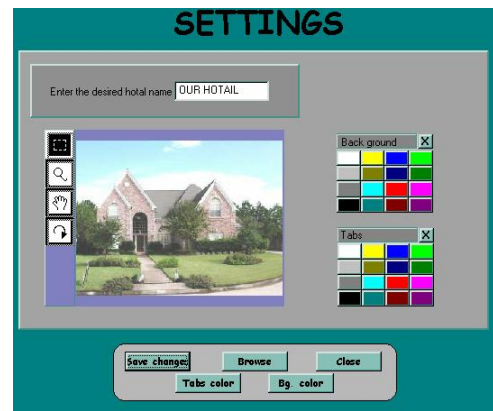


Fig. 3 Settings Screen

### B. Chart Screen (fig. 4)

This screen allow the user to enter only the date from a certain date to a certain date, then by pressing the 'View' button it will give the user a pie chart showing him the every type rooms in the hotel and its partition comparing with other room types.

The user will get also some additional numeric data.

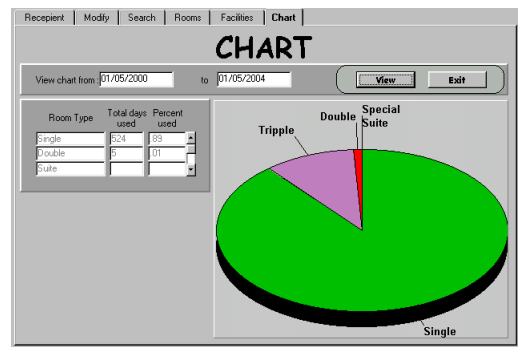


Fig.4 Chart Screen

### C. System Design and Implementation

We have tried to follow the standard phases of building a system, and our main aim is to design a user friendly system interface that helps users to finish their work in an efficient way and the lowest effort, also we tried to add many features and functions to ease the work process. Hotel systems contain a dynamic decision for the work to be finished, so we tried to divide the project into entities according to DFD (data flow diagram). We choose tab page design so the user will feel comfort with working with the system.

### D. The Accounting System

The system does not include the accounting features in a detailed way, because hotels have its own systems that deal with accounting perspective. But in our analysis for the system we found that some models should be involved in the reservation system in order to connect this system with the accounting system of the hotel.

The most important model for doing what mentioned



above is the customer accounts, and enabling the recipient to add bills (fig.5) for the customers depending on their room's numbers.

We made a single screen for adding bills to the customers, the recipient take the bill and choose one facility form the hotel facilities, and enter the amount of the bill, all what he have to do is to give the customer room number, also he can print a bill for the customer at any time while the customer is in the hotel.

Code	Room No.	Date	Customer No.	Payment	Receipt	Bill No. Active

Fig. 5 Accounting Screen

### E. Reports

The reports are considered as a very important component of any system, so we define two categories of reports, which are; static reports and generated reports.

Since the manager takes place in taking decision in the hotel, and the reports is one of the most important tools that enable him to do that, so we concentrate on this component. There is a static report on each screen; we tried to put the most common report on each tab page for every screen. Also we design a report generator that enables the user to have dynamic reports types and control the data to be included in the report.

Our concentration on the reports is for the following reasons;

- The reports is an essential part of any system,
- Helps the managers to take decision about the hotel
- Ease the process of retrieving information about the customers in the group and also the groups.

### F. Implementation Decision

We intended to use Oracle8 [2] and Developer6 which include Forms6i and Report6i; we thought that it's powerful for implementing the proposed system.

Oracle contains many features that make it over choice for implementing the project. The main features are:

- 1) Oracle is a database Management System, which helps in implementation of enterprise systems.
- 2) Oracle supports a stable database with minimum percentage of crash.
- 3) It helps in designing powerful database tables and schemes.

Also, the packages included in Oracle [3], such as Form Builder and Report Builder provides good GUI tools in order to ease the process of prototyping and designing interfaces in a user-friendly manner [4]. We had used ER diagram for database design fig. 6, because it is powerful tool for database design [5] and scheme elicitation.

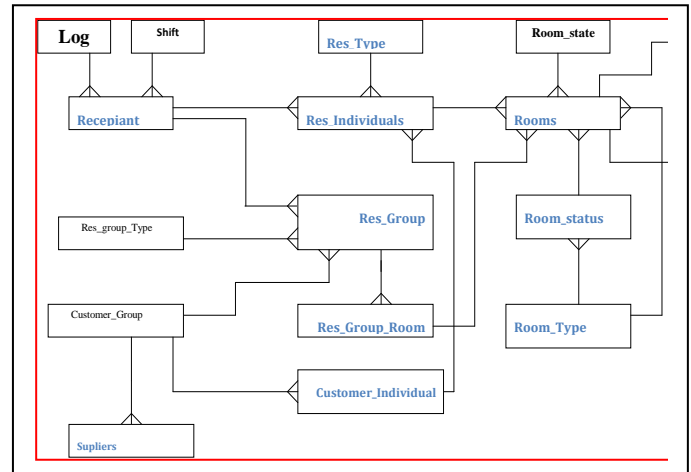


Fig. 6 ER diagram

### G. System Features

#### 1) Fast and Easy to use

From a system point of view nothing is more important to the quality of direct customer service (reservations, check-in/out etc.) than speed and accuracy of response. Our hotel system challenges overcome all other hotel systems in speed and in ease of use.

#### 2) Flexible Reservation and Enquiry Procedure

As much or as little information that any Enquirer offers can be entered in whatever orders it is given. As soon as the name is entered, be it a previous guest or Enquirer, a corporate contact or travel agent, a search will rapidly present their details including e.g. negotiated rates and preferred room numbers. The speed of this name search is of critical importance if it is to be effective - firstly as an aid to prompt customer service and recognition and secondly to ensure the growth of a 'clean' accurate database without record duplication. Our hotel fulfills these criteria.

#### 3) Automatic Report Generator

For any needed Enquirer, simply enter their date and make some choices to have quickly and accurately completed in front of your eyes a report automatically created. The benefits are faster customer service, accurate information and reduced mailing costs. (No special hardware, only a standard PC is required).

#### 4) On-Screen Room Chart

The Room Chart provides availability and occupancy information in a familiar format for those transferring from a manual front office system. Our system is designed to be as friendly and familiar to the first-time computer user as possible, which also makes re-training for experienced users relatively quick and easy.



5) *Accounts*

The majority of data transfer to accounts is usually between the front office (facility) and the accounting, which are totally integrated with our hotel. Finally our hotel System hopes to provide Travelogue Hotels and Resorts with faster and better guest service, including up-to-date history information. The system can provide increased revenues through tighter database control, as well as forecasting data and reports to help make smarter management decisions.

6) *The Front Office Features*

Front office gives the user the ability to enter recipient employees and give them their nick-names and passwords, which can be changed by the recipient later on. Also the search ability for recipients is available.

For new rooms its cost, type, and number can be done through the front office. To add new facility and its cost, can be done through the front office. There is the ability to modify the discount rate and the tax percentage.

Also in the front office there is the ability to the user to view a semi analysis on the usability of the rooms according to a given period of time, it views the data at two statistical ways the first one graphical way by showing a pie chart for rooms for specified type, the second one by showing a table with accurate number for rooms.

III. CONCLUSION

A. *Comparison with Fidelio System*

As we mentioned before, Fidelio is a complete hotel system, which deals with accounting and management information systems, in addition of reservation system. So our comparison, as shown in the following table 1, will be between our system and Fidelio reservation system.

As a main conclusion is the efficiency in generating reports, friendly user interfaces, making decision, enough stable system as we intended to use Oracle8 and Developer6. Previously discussed system features can be considered as conclusions.

B. *Future Scope*

The future plans are to get our developed information system more integrated with Accounting and Management systems.

REFERENCES

[1] Principles of Information Systems, 10th Edition, Ralph M. Stair, George Reynolds; ISBN-10: 0538478292 ISBN-13: 9780538478298, 704 Pages CB ©2012 Published

[2] Mastering Oracle PL/SQL: Practical Solutions Copyright © 2004 by Connor McDonald, with Chaim Katz, Christopher Beck, Joel R. Kallman, and David C. Knox, ISBN (pbk): 1-59059-217-4

[3] Oracle\_Database\_10g\_-\_The\_Complete\_Reference\_-\_McGraw\_Hill\_Osborne.pdf 7935 KB  
[http://rapidshare.com/files/11822933...ll\\_Osborne.pdf](http://rapidshare.com/files/11822933...ll_Osborne.pdf)

[4] Programming with Microsoft Visual Basic 6.0: an object-oriented approach : comprehensive , Edition illustrated; Michael V. Ekedahl, William Arthur Newman; Course Technology, 1999 - 720 pages; ISBN 0760010765, 9780760010761.

[5] Database Design for Mere Mortals®: A Hands-On Guide to Relational Database Design (2nd Edition) [Paperback] 611 pages, Michael J.

Hernandez Publisher: Addison-Wesley Professional; 2 edition (March 13, 2003) ISBN-10: 0201752840 ISBN-13: 978-0201752847.

[6] Dhindsa, K. S. (2011). Modelling & Designing Land Record Information System Using Unified Modelling Language. IJACSA - International Journal of Advanced Computer Science and Applications, 2(2), 26-30.

TABLE 1 COMPARISON BETWEEN OUR SYSTEM WITH FIDELIO SYSTEM

SYSTEM ITEM	Our System	Fidelio System
Group Reservation	Separate screen for reserving hotel for a group	Reserve using reservation screen
Individual Reservation	Separate screen for reserving hotel for an individual	Reserve using reservation screen
Allocating Rooms	Two screens one for the groups and one for the individuals, dynamic search according to check in and check out date.	One allocating screen shows the empty rooms, not dynamic, working parallel with manual system
Search	Every screen has search capability for a customer or group and its customers, able to search by name, room number, date, group or name.	Search for customers by their names or room numbers.
Settings	Able to change screen colors, hotel logo, and hotel name.	"No settings"
Reports	Static reports on each screen and a report generator for individuals and groups.	Four static reports on reservation screen.
Front Office Manager	Provide a screen with ability to add a recipient, room, facility, and charts about the hotel status.	Screen for adding new rooms and its status.

AUTHORS PROFILE



Dr. Osama Ahmad Salim Safarini had finished his PhD. from The Russian State University of Oil and Gaz Named after J. M. Gudkin, Moscow, 2000, at a Computerized-Control Systems Department. He obtained his BSC and MSC in Engineering and Computing Science from Odessa Polytechnic National State University in Ukraine 1996. He worked in different universities and

countries . His research is concentrated on Automation in different branches.

# Automatic Classification and Segmentation of Brain Tumor in CT Images using Optimal Dominant Gray level Run length Texture Features

A.Padma

Research scholar  
Thiyagarajar college of engg  
Tamil nadu-6250015,INDIA

R.Sukanesh

Professor of ECE Dept  
Thiyagarajar college of engg  
Tamil nadu-6250015,INDIA

**Abstract**— Tumor classification and segmentation from brain computed tomography image data is an important but time consuming task performed manually by medical experts. Automating this process is challenging due to the high diversity in appearance of tumor tissue among different patients and in many cases, similarity between tumor and normal tissue. This paper deals with an efficient segmentation algorithm for extracting the brain tumors in computed tomography images using Support Vector Machine classifier. The objective of this work is to compare the dominant grey level run length feature extraction method with wavelet based texture feature extraction method and SGLDM method. A dominant gray level run length texture feature set is derived from the region of interest (ROI) of the image to be selected. The optimal texture features are selected using Genetic Algorithm. The selected optimal run length texture features are fed to the Support Vector Machine classifier (SVM) to classify and segment the tumor from brain CT images. The method is applied on real data of CT images of 120 images with normal and abnormal tumor images. The results are compared with radiologist labeled ground truth. Quantitative analysis between ground truth and segmented tumor is presented in terms of classification accuracy. From the analysis and performance measures like classification accuracy, it is inferred that the brain tumor classification and segmentation is best done using SVM with dominant run length feature extraction method than SVM with wavelet based texture feature extraction method and SVM with SGLDM method. In this work, we have attempted to improve the computing efficiency as it selects the most suitable feature extraction method that can be used for classification and segmentation of brain tumor in CT images efficiently and accurately. An average accuracy rate of above 97% was obtained using this classification and segmentation algorithm.

**Keywords** - Dominant Gray Level Run Length Matrix method (DGLRLM); Support Vector Machine (SVM); Spatial Gray Level Dependence Matrix method (SGLDM); Genetic Algorithm (GA).

## I. INTRODUCTION

In recent years, medical CT Images have been applied in clinical diagnosis widely. That can assist physicians to detect and locate Pathological changes with more accuracy. Computed Tomography images can be distinguished for

different tissues according to their different gray levels. The images, if processed appropriately can offer a wealth of information which is significant to assist doctors in medical diagnosis. A lot of research efforts have been directed towards the field of medical image analysis with the aim to assist in diagnosis and clinical studies [1]. Pathologies are clearly identified using automated CAD system [2]. It also helps the radiologist in analyzing the digital images to bring out the possible outcomes of the diseases. The medical images are obtained from different imaging systems such as MRI scan, CT scan and Ultra sound B scan. The computerized tomography has been found to be the most reliable method for early detection of tumors because this modality is the mostly used in radio therapy planning for two main reasons. The first reason is that scanner images contain anatomical information which offers the possibility to plan the direction and the entry points of radio therapy rays which have to target only the tumor region and to avoid other organs. The second reason is that CT scan images are obtained using rays, which is same principle as radio therapy. This is very important because the intensity of radio therapy rays have been computed from the scanned image.

Advantages of using CT include good detection of calcification, hemorrhage and bony detail plus lower cost, short imaging times and widespread availability. The situations include patient who are too large for MRI scanner, claustrophobic patients, patients with metallic or electrical implant and patients unable to remain motionless for the duration of the examination due to age, pain or medical condition. For these reasons, this study aims to explore methods for classifying and segmenting brain CT images. Image segmentation is the process of partitioning a digital image into set of pixels. Accurate, fast and reproducible image segmentation techniques are required in various applications. The results of the segmentation are significant for classification and analysis purposes. The limitations for CT scanning of head images are due to partial volume effects which affect the edges produce low brain tissue contrast and yield different objects within the same range of intensity. All these limitations have made the segmentation more difficult. Therefore, the challenges for automatic segmentation of the CT brain images have many different approaches. The segmentation techniques

proposed by Nathali Richarda et al and Zhang et al [3][4] include statistical pattern recognition techniques. Kaiping et al [5] introduced the effective Particle Swarm optimization algorithm to segment the brain images into Cerebro spinal fluid (CSF) and suspicious abnormal regions but without the annotation of the abnormal regions. Dubravko et al and Matesin et al [6] [7] proposed the rule based approach to label the abnormal regions such as calcification, hemorrhage and stroke lesion. Ruthmann et al [8] proposed to segment Cerebro spinal fluid from computed tomography images using local thresholding technique based on maximum entropy principle. Luncaric et al proposed [9] to segment CT images into background, skull, brain, ICH, calcifications by using a combination of k means clustering and neural networks. Tong et al proposed [10] to segment CT images into CSF, brain matter and detection of abnormal regions using unsupervised clustering of two stages. Clark et al [11] proposed to segment the brain tumor automatically using knowledge based techniques. Lauric et al [12] proposed to segment the CT brain data into CSF and brain matter using Bayesian classifier. Genesan et al introduced [13] to segment the tumor from CT brain images using genetic algorithm. Hu et al proposed [14] to segment the brain matter from 3D CT images by applying Fuzzy C means and thresholding to 2D slices to create 2D masks and then propagating 2D masks between neighboring slices. From the above literature survey shows that intensity based statistical features are the straightest forward and have been widely used, but due to the complexity of the pathology in human brain and the high quality required by clinical diagnosis, only intensity features cannot achieve acceptable result. In such applications, segmentation based on textural feature methods gives more reliable results. Therefore texture based analysis have been presented for tumor segmentation such as Dominant gray level run length matrix method ,SGLDM method and wavelet based texture features are used and achieve promising results.

Based on the above literature, better classification accuracy can be achieved using dominant run length statistical texture features. In this paper, the authors would like to propose a dominant gray level run length statistical texture feature extraction method [15] The extracted texture features are optimized by Genetic Algorithm(GA)[16] for improving the classification accuracy and reducing the overall complexity. The optimal texture features are fed to the SVM [17] classifier to classify and segment the tumor region in brain CT images.

## II. MATERIALS AND METHODS

Most classification techniques offer grey level (i.e) pixel based statistical features. The proposed system is divided into 4 phases (i) Image preprocessing (ii) Segmentation of Region of Interest (ROI) (iii) Feature extraction (iv) Feature selection (v) Classification and Evaluation. For feature extraction, we discovered three methods which are i) Dominant gray level run length feature extraction method ii) Wavelet based feature extraction method iii) SGLDM method . Once all the features are extracted, then for feature selection, we use Genetic Algorithm (GA) to select the optimal run length statistical texture features. After selecting the optimal run length texture

features, to classify and segment the tumor region from brain CT images using SVM classifier.

### A. Image preprocessing

Brain CT images are noisy, inconsistent and incomplete, thus preprocessing phase is needed to improve the image quality and make the segmentation results more accurate. The cropping operation can be performed to remove the background. The Contrast Limited Adaptive Histogram Equalization (CLAHE) can be used to enhance the contrast within the soft tissues of the brain images and hybrid median filtering technique can also be used to improve the image quality.

### B. Selection of ROI

First the pixel having highest intensity value in the image is selected, then that pixel is compared to the neighboring pixels. The comparison goes till there is change in the intensity level of pixel value. All the pixels having the similar intensity form the Region of Interests.

### C. Feature extraction

As the tissues present in brain are difficult to classify using shape or intensity level of information, the texture feature extraction is founded to be very important for further classification. The purpose of feature extraction is to reduce original data set by measuring certain features that distinguish one region of interest from another .The spatial features from each ROI are extracted by dominant gray level run length method. In this method, by using multilevel dominant eigenvector estimation algorithm and the Bhattacharyya distance measure for texture extraction.

### D. Dominant Gray Level Run Length Matrix method

The DGLRLM is based on computing the number of gray level runs of various lengths. A gray level run is a set of consequence and collinear pixels point having the same gray level value. The length of the run is the number of pixel points in the run. The gray level run length matrix is as follows:

$$\phi(d, \theta) = [p(i, j / d, \theta)], 0 < i \leq Ng, 0 < j \leq Rmax \quad (1)$$

Where Ng is the maximum gray level and Rmax is the maximum Run length. The element p (i, j |  $\theta$ ) specified the estimated number of runs that a given image contains a run length j for a gray level i in the direction of angle  $\theta$ . Four gray level run length matrices corresponding to  $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$  are computed for an each region of interest and the following four textural features such as Short run low gray level emphasis, short run high gray level emphasis, Long run low gray level emphasis, Long run high gray level emphasis are extracted for each gray level run length matrix and take the average all four gray level run length matrices.

#### 1.Short Run Low Gray Level Emphasis(SRLGE)

$$SRLGE = 1/nr \sum_{i=1}^M \sum_{j=1}^N \frac{p(i, j)}{i^2 \cdot j^2} \quad (2)$$

#### 2. Short Run High Gray Level Emphasis(SRHGE)

$$SRLGE = 1/nr \sum_{i=1}^M \sum_{j=1}^N \frac{p(i, j) \cdot i^2}{j^2} \quad (3)$$

3. Long Run Low Gray Level Emphasis(LRLGE)

$$LRLGE = 1/nr \sum_{i=1}^M \sum_{j=1}^N \frac{p(i, j) \cdot j^2}{i^2} \quad (4)$$

4. Long Run High Gray Level Emphasis(LRHGE)

$$LRHGE = 1/nr \sum_{i=1}^M \sum_{j=1}^N p(i, j) \cdot i^2 \cdot j^2 \quad (5)$$

Where p is the run length matrix, p (i, j) is an element of the run length matrix at the position (i, j) and nr is the number of runs in the image. For all the four directions, the dominant gray level run length texture features are extracted.

TABLE 1. FEATURES EXTRACTED USING DGLRLM METHOD

SI-NO	High order DGLRLM features
1.	Short Run Low Gray Level emphasis(SRLGE)
2.	Short Run High Gray Level emphasis(SRHGE)
3.	Long Run Low Gray Level emphasis(LRLGE)
4.	Long Run High Gray level emphasis(LRLGE)

a) Wavelet based feature extraction method

A two level discrete wavelet decomposition [18] of region of interest (ROI) is performed which results in four sub bands such as LL,LH,HH,HL. The sub band LL represents low frequency part of the image and the sub bands LH,HH,HL represents high frequency part of the image. The wavelet decomposition transform is applied this high frequency sub bands only. Daubechies wavelet filter of order two is used.

Algorithm for feature extraction is as follows

- Obtain the sub-image blocks, starting from the top left corner.
- Decompose sub-image blocks using two level 2-D DWT.
- Derive Spatial Gray Level Dependence Matrices (SGLDM) or Gray Level Co-occurrence matrices [19] for each 2 level high frequency sub-bands of decomposed sub image blocks with 1 for distance and 0, 45, 90 and 135 degrees for  $\theta$  and averaged.

From these co-occurrence matrices, the following nine Haralick second order statistical texture features [20] called wavelet Co-occurrence Texture features (WCT) are extracted .

TABLE 2. WCT FEATURES EXTRACTED USING SGLDM METHOD

SI.No	Second order WCT features
1.	Entropy (Measure the disorder of an image)
2.	Energy ( Measure the textural uniformity)
3.	Contrast (Measure the local contrast in an image)
4.	Sum Average ( average of the gray levels)
5.	Variance ( heterogeneity of an image)

6. Correlation (correlation of pixel pairs)
7. Maxprobability(the most prominent pixel pair)
8. Inverse Difference moment( homogeneity )
9. Cluster tendency (Measure the grouping of Pixels with similar values)

b) Spatial Gray Level Dependence Matrix method (SGLDM)

For each ROI, the Spatial Gray Level Dependence Matrix method (SGLDM) can be used to extract the second order statistical texture features for the better diagnosis. This method is based on the estimation of the second order joint conditional probability density functions  $P(i, j | d, \theta)$  where  $\theta = 0, 45, 90, 135$  degrees. Each  $P(i, j | d, \theta)$  is the probability matrix of two pixels which are located with an inter sample distance d and direction  $\theta$  have a gray level i to gray level j. The estimated value for these probability density functions can be represented by

$$\phi(d, \theta) = [P(i, j | d, \theta)], 0 < i, j \leq Ng \quad (6)$$

Where, Ng is the maximum gray level. In this method, four gray level co-occurrence matrixes for four different directions ( $\theta = 0, 45, 90, 135$ ) are obtained for a given distance d (=1, 2) and the following 13 Haralick statistical texture features [20] are calculated for each gray level co-occurrence matrix and take the average of all four gray level co-occurrence matrices. The set of extracted second order statistical texture features from each ROI forms the feature vector or feature set.

TABLE 3. FEATURES EXTRACTED USING SGLDM METHOD

SI-no	Second order SGLDM features
1.	Energy
2.	Entropy
3.	Contrast
4.	Inverse difference moment
5.	Homogeneity
6.	Correlation
7.	Sum Average
8.	Sum Entropy
9.	Difference Average
10.	Difference Entropy
11.	Sum of squares variance
12.	Information measures of Correlation I& II

c) Feature selection

Feature selection is the process of choosing subset of features relevant to particular application and improves classification by searching for the best feature subset, from the fixed set of original features according to a given feature evaluation criterion(i.e., classification accuracy). Optimized feature selection reduces data dimensionalities and computational time and increase the classification accuracy. The feature selection problem involves the selection of a subset of the features from a total number of the features, based on a given optimization criterion. T denotes the subset of selected features and V denotes the set of remaining features. So,  $S = T$

U V at any time. J(T) denotes a function evaluating the performance of T. J depends on the particular application. Here J(T) denotes the classification performance of classifying and segmenting soft tissues from brain CT images using the set of features in T. In this work, genetic algorithm is used.

GENETIC ALGORITHM:

We consider the standard GA to begin by randomly creating its initial population. Solutions are combined via a crossover operator to produce offspring, thus expanding the current population of solutions. The individuals in the population are then evaluated via a fitness function, and the less fit individuals are eliminated to return the population to its original size. The process of crossover, evaluation, and selection is repeated for a predetermined number of generations or until a satisfactory solution has been found. A mutation operator is generally applied to each generation in order to increase variation. In the feature selection formulation of the genetic algorithm, individuals are composed of chromosomes: a 1 in the bit position indicates that feature should be selected; 0 indicates this feature should not be selected. As an example the chromosome 00101000 means the 3rd and 5th features are selected. The features 1, 2, 4, 6, 7, 8th feature are not selected. That is the chromosome represents  $T=\{3,5\}$  and  $V=\{1,2,4,6,7,8\}$ . Fitness function for the given chromosome T is defined as

$$\text{Fitness}(T) = J(T) - \text{penalty}(T) \tag{7}$$

where T is the corresponding feature subset, and  $\text{penalty}(T) = w * (|T| - d)$  with a penalty coefficient w. The size value d is taken as a constraint and a penalty is imposed on chromosomes breaking this constraint. The chromosome selection for the next generation is done on the basis of fitness. The fitness value decides whether the chromosome is good or bad in a population. The selection mechanism should ensure that fitter chromosomes have a higher probability survival. So, the design adopts the rank-based roulette-wheel selection scheme. If the mutated chromosome is superior to both parents, it replaces the similar parent. If it is in between the two parents, it replaces the inferior parent; otherwise, the most inferior chromosome in the population is replaced. The selected optimal feature set based on the test data set is used to train the SVM classifier to classify and segment the tumor region from brain CT images.

a) Classifier

Classification is the process where a given test sample is assigned a class on the basis of knowledge gained by the classifier during training. To make the classification results comparable and for exhaustive data analysis, we have used leave one out classification method for the SVM classifier.

Figure 1 shows the class separation of SVM classifier. Support Vector Machine (SVM) performs the robust non-linear classification with kernel trick. SVM is independent of the dimensionality of the feature space and that the results obtained are very accurate. It outperforms other classifiers even with small numbers of available training samples. SVM is a supervised learning method and is used for one class and n class classification problems.

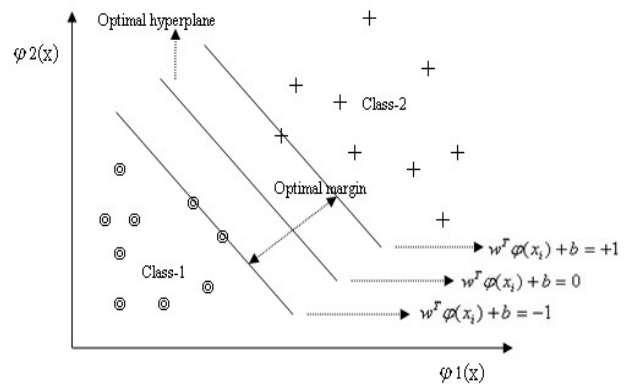


Figure 1.SVM classifier

It combines linear algorithms with linear or non-linear kernel functions that make it a powerful tool in the machine learning community with applications such as data mining and medical imaging applications. To apply SVM into nonlinear data distributions, the data can be implicitly transformed to a high dimensional feature space where a separation might become possible. For a binary classification given a set of separable data set with N samples  $X = \{X_i\}$ ,  $i = 1, 2 \dots N$ , labeled as  $Y_i = \pm 1$ . It may be difficult to separate these 2 classes in the input space directly. Thus they are mapped into a higher dimensional feature space by  $X' = f(x)$ .

The decision function can be expressed as

$$f(x) = W \cdot x + \rho \tag{8}$$

Where  $W \cdot x + \rho = 0$  is a set of hyper planes to separate the two classes in the new feature space. Therefore for all the correctly classified data,

$$Y_i f(x) = Y_i (W \cdot x + \rho) > 0, i = 1, 2 \dots N \tag{9}$$

By scaling W and  $\rho$  properly, we can have  $f(x) = W \cdot x + \rho = 1$  for those data labeled as +1 closes to the optimal hyper plane and  $f(x) = W \cdot x + \rho = -1$  for all the data labeled as -1 closes to the optimal hyper plane. In order to maximize the margin the following problem needs to be solved.

$$\text{Min} (\|W\|/2/2)$$

$$\text{Subject to } Y_i f(x) = Y_i (W \cdot x + \rho) \geq 1, i = 1, 2 \dots N \tag{10}$$

It is a quadratic programming problem to maximize the margins which can be solved by sequential minimization optimization. After optimization, the optimal separating hyper plane can be expressed as

$$f(x) = \sum_{i=1}^N \alpha_i Y_i K(x_i, x) + \rho \tag{11}$$

Where  $K(\cdot)$  is a kernel function,  $\rho$  is a bias,  $\alpha$  is the solutions of the quadratic programming problem to find maximum margin. When  $\alpha$  is non-zero, are called support vectors, which are either on or near separating hyper plane. The decision boundary (i.e.) the separating hyper plane whose decision values  $f(x)$  approach zero, compared with the support vectors, the decision values of positive samples have larger positive values and those of negative samples have larger negative values.

Therefore the magnitude of the decision value can also be regarded as the confidence of the classifier. The larger the magnitude of  $f(x)$ , the more confidence of the classification by choosing a Gaussian kernel function

$$K(x,y) = e^{-\gamma\|x-y\|^2} \quad (12)$$

Where the value of  $\gamma$  was chosen to be 1 and has good performance for the following two reasons. First reason is the Gaussian model has only one parameter and it is easy to construct the Gaussian SVM classifier compared to polynomial model which has multiple parameters. Second reason is there is less limitation in using Gaussian kernel function due to nonlinear mapping in higher dimensional space.

### III. RESULTS AND DISCUSSIONS

This section describes the comparative study of classification performance of the SVM classifier for different texture analysis methods used for classification and segmentation of tumor from brain CT images. The texture features extracted from each ROI of the image to be selected by using pixel based intensity method. The texture features are extracted with same set of images and are obtained from 16 bit gray level images. The SVM is used as a classifier. The results from the SVM classifier for all the texture analysis methods are evaluated by using the statistical analysis.

An experiment has been conducted on a real CT scan brain images based on the proposed system.

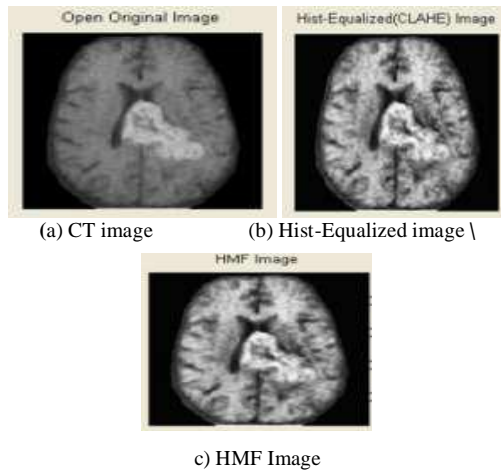


Figure 2(a-c) Preprocessed image

Figure 2(a-c) shows the original input CT image, histogram equalized image, hybrid median filtered image which is used to reduce the different illumination conditions and noises.

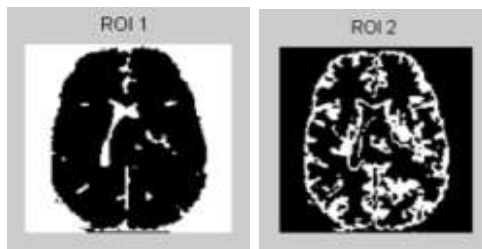


Figure 3(a-d) Selection of ROIS

Figure 3(a-d) represents the selected ROIs of the image to be segmented using pixel based intensity method. For the comparative analysis of texture analysis methods using SVM classifier, 120 images were partitioned arbitrarily into training set, testing set with equal number of images. The accuracy of the classifier for the texture analysis methods are evaluated based on the error rate. This error rate can be described by the terms true and false positive and true and false negative as follows

True Positive (TP): Abnormal cases correctly classified

True Negative (TN): Normal cases correctly classified.

False Positive (FP): Normal cases classified abnormal

False Negative (FN): Abnormal cases classified normal

The above terms are used to describe the clinical efficiency of the classification and segmentation algorithm.

$$\text{Sensitivity} = TP / (TP + FN) * 100$$

$$\text{Specificity} = TN / (TN + TP) * 100$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) * 100.$$

Accuracy is the proportion of correctly diagnosed cases from the total number of cases. Sensitivity measures the ability of the method to identify abnormal cases. Specificity measures the ability of the method to identify normal cases. To make the classification results comparable and for exhaustive data analysis, leave one out cross validation method can be used to estimate the classifier performance in unbiased manner. Here each step, one data set is left out and the classifier is trained using the rest and the classifier is applied to the left out data set. This procedure is repeated such that each data set is left out once. In our application for evaluating classification accuracy, 10 fold cross validation method is done on the data set collected from 120 images (60 normal and 60 abnormal).

The images are divided into 10 sets each consisting of 6 normal images and 6 abnormal images. Then 9 sets are used for training and remaining set is used for testing. In next iteration (2-10), 9 sets are used for training and remaining set is used for testing. This process is repeated for 10 times. Hence 10 iterations are done. The classification accuracy was calculated for by taking the average of all correct classifications. All the textural features were normalized by the sample mean and standard deviation of the image.



TABLE 4. CLASSIFICATION RESULTS OF SVM CLASSIFIER USING CO-OCCURRENCE, WCT, NEW RUN LENGTH FEATURES

Feature name	feature length	features selected	classification accuracy
1.Co-occurrence	52	16	93.3%
2.2 level-WCT features	27	12	95.8%
3.New run length features	16	8	98.3%

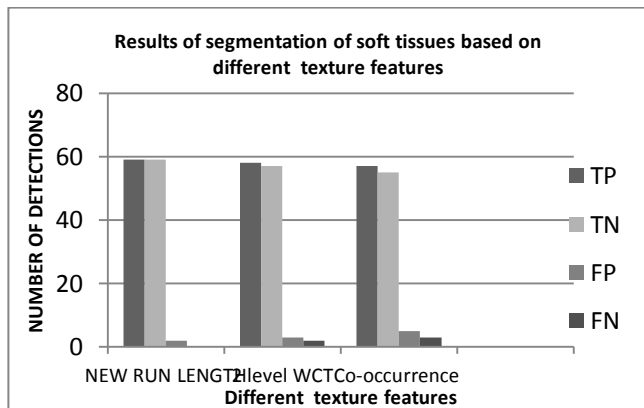


Figure 4. No. of detections based on different texture features

Figure 4 shows the number of abnormal tumor detection based on the three texture analysis methods. The results show that, if the representative samples increased, it gives good classification accuracy for 10 fold cross validation method. The sensitivity and specificity values of texture analysis methods are shown in Figure 5.

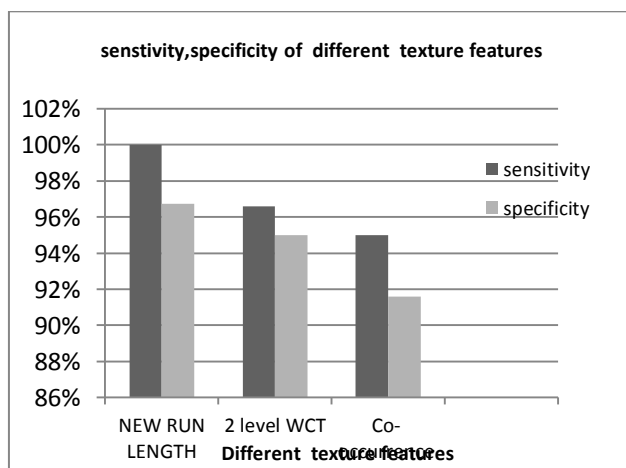


Figure 5. Sensitivity,specificity of texture analysis methods

The sensitivity, specificity values of the new run length method are 96.72%, 98.3% and the wavelet co-occurrence method are 96.6%,95% and co-occurrence method are 95%, 91.6% respectively.

Then the accuracy of the new run length method is 98.3% and wavelet co-occurrence method is 95.8% and the co-occurrence method is 93.3% respectively. We now compare the new run length method with the co-occurrence method and wavelet co-occurrence method on the brain CT images. For the co-occurrence method, 13 co-occurrence features are computed for each of the four directions (i.e) 0,45,90,135 degrees; for the wavelet co-occurrence method, the texture features used for each wavelet decomposition high frequency sub bands are entropy, energy, contrast, sum average, variance, correlation, max probability, inverse difference moment, cluster tendency. Table 5 shows the accuracy of different segmentation methods.

TABLE 5. ACCURACY OF DIFFERENT SEGMENTATION METHODS

SI-n	Segmentation method	Accuracy
1.	Fuzzy C means	85%
2.	K means	87.3%
3.	Bayesian classifier	89%
4.	Genetic algorithm	93%
5.	SVM classifier with feature extraction method	98.3%

The same genetic feature selection method is used for all the three texture analysis methods. In dominant gray level run length feature extraction method, the selected features were long run high gray level emphasis and long run low gray level emphasis. The long run high gray level emphasis captures the inhomogeneous nature of texture features and long run low gray level emphasis captures the homogeneous nature of texture features. About 93% classification accuracy is achieved by most of the feature vectors.

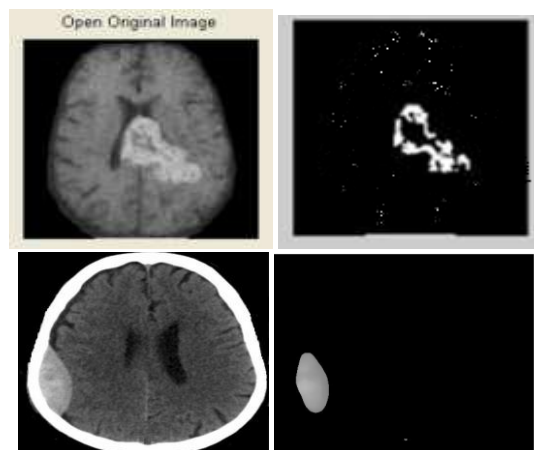


Figure 6(a,b,c,d). CT image and the segmented tumor image

Figure 6 (a,b,c,d) shows the input CT image and the corresponding segmented tumor image. From the classification results of the SVM classifier for all the three texture analysis methods, the new run length method perform comparably well with the wavelet co-occurrence features and better than the co-occurrence features. This demonstrates that there is more texture information contained in run length matrices and that a good method of extracting such information is important to classification and segmentation algorithm.

#### IV. CONCLUSION

In this work a new dominant run length feature extraction methodology for the classification and segmentation of tumor in brain CT images using support vector machine with genetic algorithm feature selection is proposed. The algorithm has been designed based on the concept of different types of brain soft tissues (CSF, WM, GM, Abnormal tumor region) have different textural features. The selection method of ROI is simple and accurate. The results show that the segmentation and classification of tumor for the new run length feature extraction method yields better results compared to the other texture analysis methods based on SVM classifier. It is found that this method gives favorable result with accuracy percentage of above 98% for the images that are being considered. This would be highly useful as a diagnostic tool for radiologists in the automated segmentation of tumor in CT images.

The goal of this work is to compare the classification performance of the SVM classifier based on different texture analysis methods. Hence it is concluded that the neural network supported by conventional texture analysis methods can be effectively used for classification and segmentation of brain tumor from CT images. Use of large data bases is expected to improve the system robustness and ensure the repeatability of the resulted performance. The automation procedure proposed in this work using a SVM enables proper abnormal tumor region detection and segmentation thereby saving time and reducing the complexity involved.

In this research work, the four segmentation methods, Bayesian classification, fuzzy C means, K means and expectation maximization algorithm were produced good results. In this work, the dominant feature extraction method with SVM classifier is proposed for the classification and segmentation of brain CT images and the results are compared with Bayesian classification and fuzzy c means, k means clustering method and expectation maximization algorithm through statistical analysis. The proposed method has better performance compared to the other existing methods. Plans for future work include the specific annotation of the abnormal regions such as hemorrhage, calcification and lesion.

#### REFERENCES

- [1]. Duncan J.S., Ayache N., "Medical Image Analysis- Progress Over two decade and challenges ahead", IEEE Trans on PAMI, Vol 22, pp. 85 – 106, 2000.
- [2]. G.P. Tourassis, "Journey towards computer aided Diagnosis – Role of Image Texture Analysis", Radiology, Vol 2, pp. 317 – 320, 1999.
- [3]. Nathalie Richard, Michel Dojata, Catherine Garbayvol, "Distributed Markovian Segmentation Application to MR brain Scans", Journal of Pattern Recognition, Vol 40, pp. 3467 – 3480, 2007.
- [4]. Y. Zhang, M. Brady, S. Smith, "Segmentation of Brain MR Images through hidden Markov random field model and the expectation-maximization algorithm", IEEE Transactions on Medical Imaging, Vol 20, pp. 45 – 57, 2001.
- [5]. Kaiping Wei, Bin He, Tao Zhang, Xianjun Shen, "A Novel Method for segmentation of CT Head Images", International conference on Bio informatics and Biomedical Engineering, Vol 4, pp. 717 – 720, 2007.
- [6]. Dubravko Cosic, Sven Loncaric, "Rule based labeling of CT head images", 6th conference on Artificial Intelligence in Medicine, pp. 453 – 456, 1997.
- [7]. Matesn Milan, Loncaric Sven, Petravic Damir, "A rule based approach to stroke lesion analysis from CT brain Images", 2nd International symposium on Image and Signal Processing and Analysis, June (2001), pp. 219 – 223, 2001.
- [8]. Ruthmann V.E., Jayce E.M., Reo D.E, Eckaidt M.J., (1993), "Fully automated segmentation of cerebro spinal fluid in computed tomography", Psychiatry research: Neuro imaging, Vol 50, pp. 101 – 119, 1993.
- [9]. Loncaric S and D. Kova Cevic, "A method for segmentation of CT head images", Lecture Notes on Computer Science, Vol 1311, pp. 1388 – 305, 1993.
- [10]. Tong Hau Lee, Mohammad Faizal, Ahmad Fauzi and Ryoichi Komiya, "Segmentation of CT Brain Images using unsupervised clusterings", Journal of Visualization, Vol 12, pp. 31-138, 2009.
- [11]. Clark M C., Hall L O., Goldgof D B, Velthuzien R., Murtagh F R., and Silbiger M S., "Automatic tumor segmentation using knowledge based techniques", IEEE Transactions on Medical Imaging, Vol 17, pp. 187-192, 1998.
- [12]. Lauric A., Frisken S., "Soft segmentation of CT brain data", Technical Report TR-2007-3 Tufts University, M A.
- [13]. Ganesan R, Radhakrishnan R., "Segmentation of Computed Tomography Brain Images using genetic algorithm", International Journal of Soft computing, Vol 4, pp. 157-161, 2009.
- [14]. Hu Q., Qian G, Aziz A and Nowinski W L., "Segmentation of brain from computed tomography head images", IEEE Eng. Medicine and Biology Society, Vol 6, pp. 3375-3378, 2005.
- [15]. Tang Xiaou, "Texture Information in Run Length Matrices", IEEE Transaction on Image Processing, Vol 7, pp. 234-243, 1998.
- [16]. Frank Z. Brill., Donald E. Brown., and Worthy N. Martin, "Fast Genetic Selection of Features for Neural Network Classifiers", IEEE Trans., on Neural Networks, Vol 3, pp. 324-328, 1992.
- [17]. El-Naqa I, Yang Y, Wernick M N, Galatsanos N P and Nishikawa R M, "A support vector machine approach for detection of micro calcifications", IEEE Trans. on Medical Imaging, Vol 21, pp. 1552-1563, 2002.
- [18]. Van G., Wouwer P., Scheunders and D. Van Dyck, "Statistical texture characterization from discrete wavelet representation", IEEE Trans. Image processing, Vol 8, pp. 592-598, 1999.
- [19]. Haddon J F, Boyce J F, Co-occurrence Matrices for Image analysis. IEE Electronic and Communications Engineering Journal, Vol 5, pp. 71 – 83, 1993.
- [20]. Haralick R M, Shanmugam K and Dinstein I, "Texture features for Image classification", IEEE Transaction on System, Man, Cybernetics, Vol 3, pp. 610 – 621, 1973.
- [21]. Chaabane, L. (2011). Improvement of Brain Tissue Segmentation Using Information Fusion Approach. IJACSA - International Journal of Advanced Computer Science and Applications, 2(6), 84-90.
- [22]. Chaabane, L. (2011). Evaluation of the Segmentation by Multispectral Fusion Approach with Adaptive Operators: Application to Medical Images. IJACSA - International Journal of Advanced Computer Science and Applications, 2(9), 1-7.

# An Ontology- and Constraint-based Approach for Dynamic Personalized Planning in Renal Disease Management

Normadiah Mahiddin<sup>1</sup>, Yu-N Cheah<sup>2</sup>, Fazilah Haron<sup>3</sup>

<sup>1</sup>Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

<sup>2,3</sup>School of Computer Sciences, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

<sup>3</sup>College of Computer Science and Engineering, Taibah University, P.O. Box 30002, Madinah, Saudi Arabia

**Abstract**—Healthcare service providers, including those involved in renal disease management, are concerned about the planning of their patients' treatments. With efforts to automate the planning process, shortcomings are apparent due to the following reasons: (1) current plan representations or ontologies are too fine grained, and (2) current planning systems are often static. To address these issues, we introduce a planning system called Dynamic Personalized Planner (DP Planner) which consists of: (1) a suitably light-weight and generic plan representation, and (2) a constraint-based dynamic planning engine. The plan representation is based on existing plan ontologies, and developed in XML. With the available plans, the planning engine focuses on personalizing pre-existing (or generic) plans that can be dynamically changed as the condition of the patient changes over time. To illustrate our dynamic personalized planning approach, we present an example in renal disease management. In a comparative study, we observed that the resulting DP Planner possesses features that rival that of other planning systems, in particular that of Asgaard and O-Plan.

**Keywords**-patient care planning; treatment protocols; dynamic treatment planning; personal health services.

## I. INTRODUCTION

Healthcare service providers are undoubtedly concerned about updating their patients' health records or profiles, and the planning of their patients' treatments to support the efficient and effective delivery of healthcare services. However, not all healthcare service providers are carrying out planning activities effectively, especially when it comes to automated or computer-based planning, due to shortcomings in current planning systems.

The first problem is that most of the current plan representations or ontologies are too fine grained (detailed). This means that the plan representations or ontologies are not suited for all situations and for all levels. We need to have a portable and intuitively easy representation that facilitates the storage and manipulation of generic plans. The second problem is that current planning systems are often static. This means planning is carried out once without taking into account changes that may take place as time goes on. These plans also do not consider past events. Dynamic planning is therefore required to allow plans to be updated as new situations arise.

To address the concerns above, we present a methodology for generic and dynamic healthcare planning, resulting in a system called the Dynamic Personalized Planner (DP Planner) [1]. For this purpose, we define (1) a suitably light-weight and generic plan representation based on existing plan ontologies, and (2) a constraint-based dynamic planning engine.

## II. STATE OF THE ART

A popular approach for plan representation is via ontologies, i.e. plans are designed based on project specific ontologies and domain description languages [2]. Fig. 1 shows the structure of the Plan Ontology proposed by Tate [3].

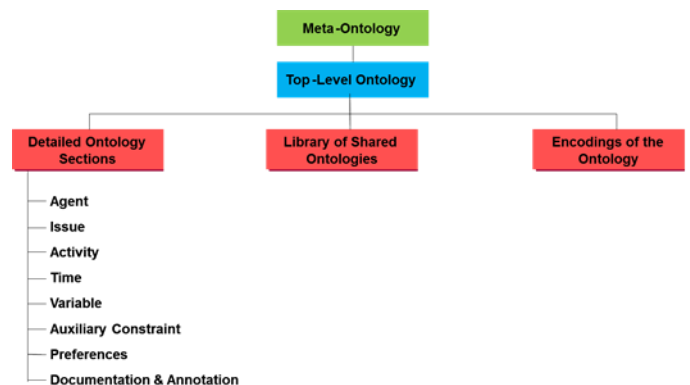


Figure 1. Plan Ontology structure [3].

Another useful plan ontology for generic planning is the task-specific ontology and approach called Asbru which uses skeletal plans [4]. Asbru represents skeletal plans using a task-specific, intention-based, and time-oriented language. It can be used to design specific plans [5]. In Asbru, a plan contains a name and a set of arguments (time annotation and knowledge roles). All plans and actions have a temporal dimension and the plan's execution is controlled by a number of conditions such as filter, setup, suspend, reactivate, abort and complete [6]. Fig. 2 shows the Asbru plan ontology structure.

Besides plan ontologies (which contribute towards plan representation), there are also a number of plan generation initiatives. An example of such an initiative is the RAX Planner/Scheduler (RAX-PS) [7]. The RAX-PS generates plans

that are temporally flexible, allowing the execution system to adjust to actual plan execution conditions without breaking the plan. The result is a system capable of building concurrent plans.

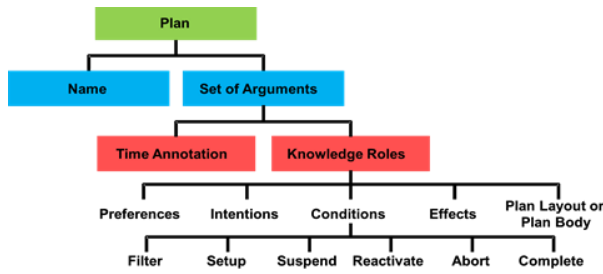


Figure 2. The Asbru plan ontology structure [4].

Fig. 3 shows the architecture of the planning system in RAX-PS. For the RAX-PS, the domain model describes the dynamics of the system [8] to which the planner is being applied, i.e. the Deep Space One Spacecraft. The plan database is initialized by a plan request which consists of an initial state and a set of goals. The initialized database is then modified by a search engine to generate a complete and valid plan. This complete plan is then put into operation by an execution agent. The heuristics and planning experts are also integral parts of the Deep Space One planning system. The heuristics guide the search engine while the planning expert interfaces with external systems which provide critical inputs such as altitude and speed.

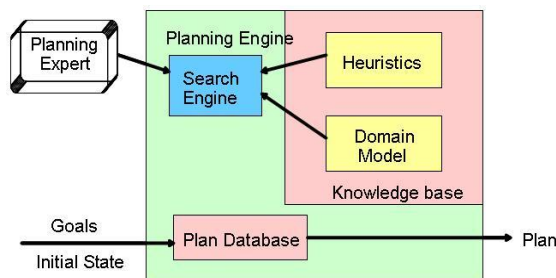


Figure 3. The RAX Planner/Scheduler (RAX-PS) [7].

The Capture the Flag (CTF) [9] game project uses the notion of critical points (time during the execution of an action or plan where a decision might be made) to define states in the continuous domain. These states are then used to efficiently evaluate plans. An action or a plan posts a goal, G. This invokes the CTF planning algorithm [10].

In the effort to generate outputs that are dynamically assembled from smaller fragments, the Personalized Healthcare Information (PHI) system [11] composes personalized documents that conform to an individual’s health profile. The composition of PHI is carried out in a three-step procedure which are (1) selection of a set of Topic Specific Documents (TD), where each selected TD addresses some of the individual’s healthcare concerns, (2) combination of the selected TDs to produce an aggregated PHI document, and (3) verification of the accuracy of the aggregated PHI document. Each individual illness/concern/issue noted in the health profile is addressed by at least a single TD. Constraint satisfaction

techniques are used to ensure that the aggregated PHI document is medically consistent. Fig. 4 shows the processes for PHI composition.

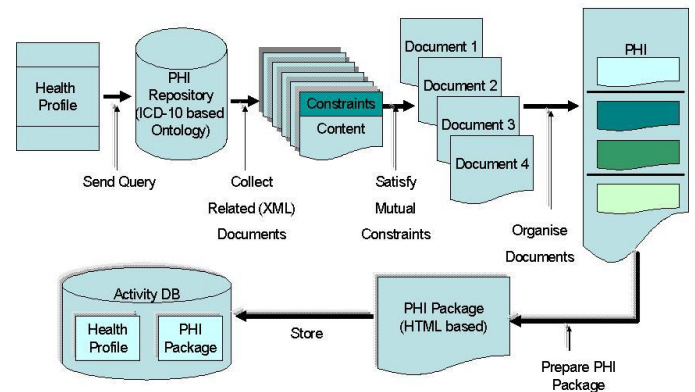


Figure 4. The process flow for PHI composition [11].

We have found the PHI system’s approach in using constraints attractive as this approach could be adapted in our DP Planner to ensure the coherency of plan fragments that are assembled to form a complete plan.

### III. THE DYNAMIC PERSONALISED PLANNING METHODOLOGY

The dynamic personalized planning approach consists of two phases:

1. Plan ontology definition and representation.
2. Planning algorithm definition.

Fig. 5 briefly shows the two phases together with the techniques and approaches used.

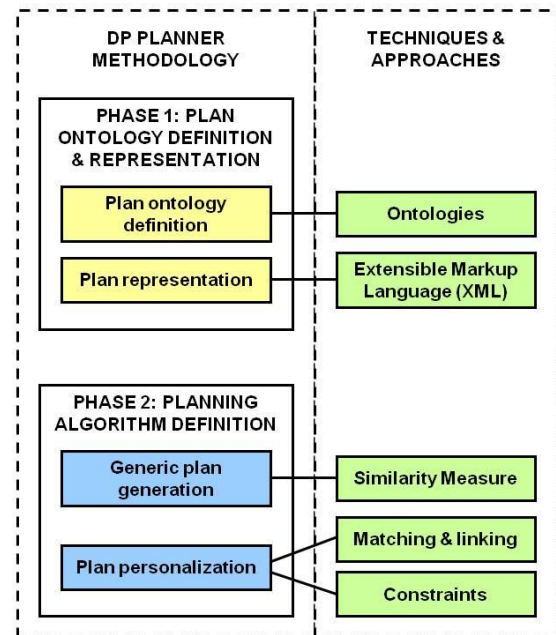


Figure 5. DP Planner methodology with related techniques and approaches [1].



### A. Phase 1: Plan Ontology Definition and Representation

In defining the plan ontology, the Asbru plan ontology was adopted as the basis for the DP Planner's plan ontology in view that it is reasonably concise compared to other ontologies that were surveyed. Besides this, some elements of Goals, Operators, Methods and Selection (GOMS) analysis [12] were also incorporated. GOMS is a method for analyzing and modeling the knowledge and skills that a user must develop in order to perform tasks on a device or system.

#### 1) Plan ontology definition

Fig. 6 shows the DP Planner's plan ontology [13].

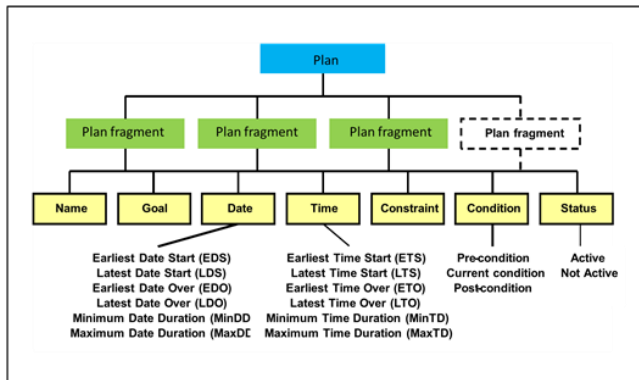


Figure 6. DP Planner's plan ontology structure [13].

Referring to Fig. 6, the plan is positioned at the highest position in the plan hierarchy, and is basically the task that needs to be performed. Examples of plans are those for kidney patient treatment monitoring, gestational diabetes mellitus monitoring, student performance monitoring, etc. The plan consists of a sequence of plan fragments. These plan fragments are the necessary steps to achieve the task and can be viewed as the most crucial component of the plan ontology.

Each plan fragment consists of seven attributes:

- Name: Identifies a plan.
- Goal: States the target to be achieved.
- Date: States the date of plan execution (if required).
- Time: States the duration of plan execution.
- Constraints: Information of the plan execution limit.
- Condition: Situations in which the task takes place.
- Status: Keeps track of the situation of plan execution.

Some of the plan component's attributes have detailed sub-attributes. Examples are as follows:

- Date has six sub-attributes: Earliest Date Start (EDS), Latest Date Start (LDS), Earliest Date Over (EDO), Latest Date Over (LDO), Minimum Date Duration (MinDD) and Maximum Date Duration (MaxDD),
- Time also has six sub-attributes: Earliest Time Start (ETS), Latest Time Start (LTS), Earliest Time Over (ETO), Latest Time Over (LTO), Minimum Time

Duration (MinTD) and Maximum Time Duration (MaxTD).

- Condition consists of three sub-attributes:
  - Pre-condition: This is to ensure that certain conditions are met before the execution of a particular plan fragment.
  - Current condition: This ensures that certain conditions are currently met in a particular plan fragment.
  - Post-condition: This is to ensure that certain conditions are met after the execution of a particular plan fragment and before the next plan fragment.

#### 2) Plan representation

Fig. 7 illustrates an example of a plan represented in XML for the treatment of a patient with renal disease. Note that the plan ontology can be naturally implemented in XML as the hierarchical nature of ontologies maps very well into the nested nature of XML. With the defined DP Planner plan ontology and representation, the DP Planner planning engine implementation is discussed next.

```
<?xml version="1.0" encoding="UTF-8"?>
<PlanRepository>
  <Patient IC="781002046086">Ani</Patient>
  <Plan ID="1">
    <PlanFragment ID="1.1" Name="Glomerular Filtration (GFR) Test">
      <Goal>Detect the stage of Kidney Disease</Goal>
      <Date>
        <EarliestDateStart>20/4/2006</EarliestDateStart>
        <LatestDateStart>21/4/2006</LatestDateStart>
        <EarliestDateOver>00/00/0000</EarliestDateOver>
        <LatestDateOver>00/00/0000</LatestDateOver>
        <MinimumDateDuration>0 years 0 months 1 days</MinimumDateDuration>
        <MaximumDateDuration>0 years 0 months 1 days</MaximumDateDuration>
      </Date>
      <Time>
        <EarliestTimeStart>08:00 AM</EarliestTimeStart>
        <LatestTime_Start>04:00 PM</LatestTimeStart>
        <EarliestTimeOver>09:00 AM</EarliestTimeOver>
        <LatestTimeOver>05:00 PM</LatestTimeOver>
        <MinimumTimeDuration>0 hours 30 minutes</MinimumTimeDuration>
        <MaximumTimeDuration>9 hours 0 minutes</MaximumTimeDuration>
      </Time>
      <Condition>
        <PreCondition>Diabetes</PreCondition>
        <PreCondition>High Blood Pressure</PreCondition>
        <CurrentCondition>Tiredness</CurrentCondition>
        <CurrentCondition>Nausea</CurrentCondition>
        <PostCondition>Percentage of Renal Damage = 60%</PostCondition>
      </Condition>
      <Constraints>Not taken any medications</Constraints>
      <Constraints>Not pregnant</Constraints>
      <Status>Completed</Status>
    </PlanFragment>
  </Plan>
</PlanRepository>
```

Figure 7. Plan representation in XML format.

### B. Phase 2: Planning Algorithm Definition

The DP Planner's planning strategy involves reusing plans that are stored in the plan repository and subsequently personalizing these plans according to the constraints defined by the user.

The planning algorithm is divided into two parts (see Fig. 8):

1. Generic plan generation.
2. Plan personalization.

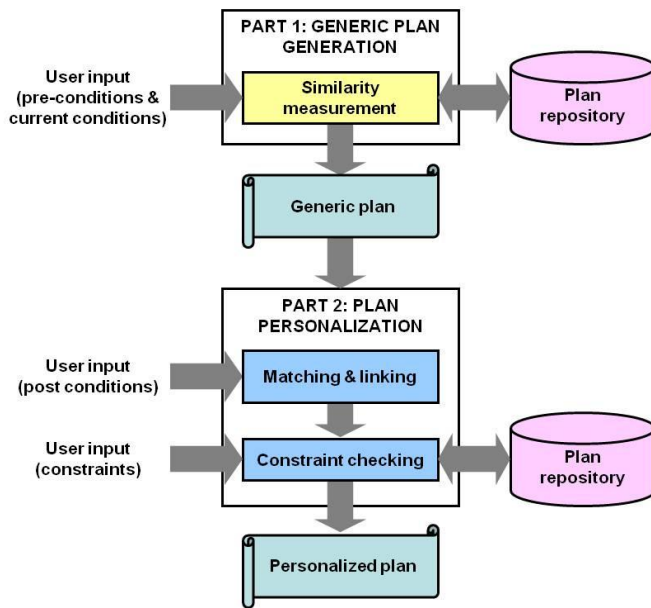


Figure 8. The planning algorithm.

1) Generic plan generation

Firstly, user inputs are matched against each of the existing plans in the plan repository. A concentration unit is calculated during the matching process to compare the closeness of the match. This unit is based on the number of matches between the user’s inputs (pre-conditions and current conditions) with those of the plan fragments which constitute each plan. The similarity between the user’s pre-condition, and current condition inputs, and those of a particular plan,  $P$ , is expressed as  $C_P$ , (see Equation 1).

$$C_P = \frac{n}{T_f} \quad (1)$$

where  $n$  = number of matches of pre-conditions + matches of current conditions, and  $T_f$  = total number of plan fragments in a plan. The values of  $C$  for each plan will be compared. The plan with the highest value will be chosen as the generic plan. After a generic plan has been identified, the process of plan personalization will follow

2) Plan personalization

The personalization method employs a combination of (1) a simple matching and linking technique, and (2) a constraint-based approach to form certain restrictions [9] so that only plan fragments that meet the predetermined criteria and user’s inputs can form the finalized and personalized plan for a particular user or situation. Personalization is only carried out when the user’s status is active, i.e. the plan is still relevant to the user’s condition. For a start, the user will be asked for the outcomes of following the activities defined in a particular plan fragment. These outcomes are called post-conditions. Here, a simple matching technique will be applied to match the post-conditions of a particular (current) plan fragment with the pre-conditions in the next plan fragment. If these match each other, that next plan fragment in the generic plan will be recommended as the subsequent plan for the user. However, if they do not match, the plan (or sequence of plan fragments) will terminate at that current plan fragment.

The matching technique is applied to ensure that each plan fragment in the finalized plan links to each other (see Fig. 9). This is to ensure that the final plan generated by the system corresponds to the needs of the user. After ensuring that the plan fragments in the generic plan fulfils the initial matching of post-conditions with the pre-conditions, the actual personalization can then take place using constraints.

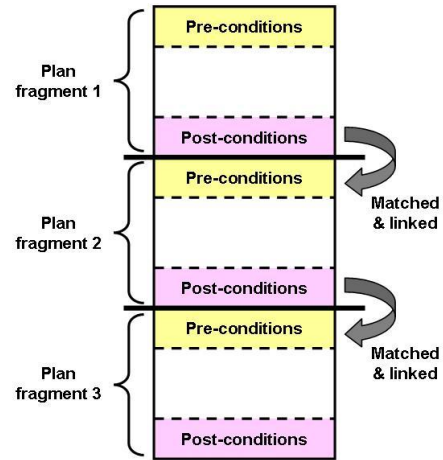


Figure 9. The linking between plan fragments in a plan.

Constraints are utilized to ensure the consistency of the multiple fragments of a plan in order to form a complete plan. The individual plan fragments must not contradict each other or lead to improper recommendations [11]. Therefore, for the purpose of the DP Planner, a constraint is simply a variable which restricts the execution of a particular plan fragment. It also describes the relationship between plan fragments in a plan. Ultimately, the constraints are used to find suitable replacements for plan fragments which are not suited for the plan fragments preceding or following it. Fig. 10 illustrates how constraints are used to personalize a generic plan.

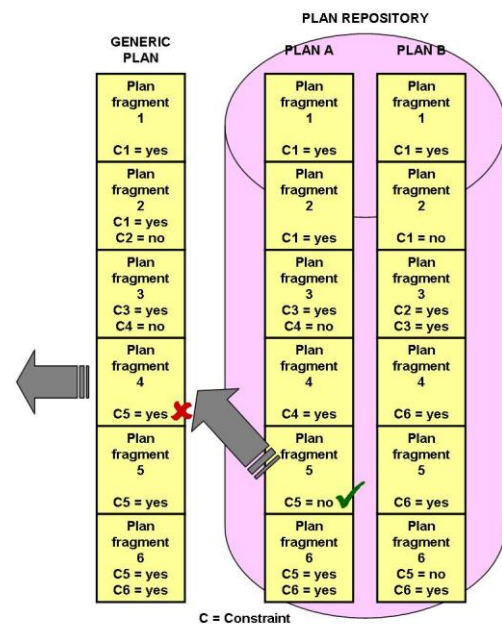


Figure 10. The use of constraints in plan personalization.



In Fig. 10, the generic plan’s Plan Fragment 4 shows constraint 5 (C5) = yes. However, let us assume this does not fulfill a constraint specified by the user, i.e. the user has C5 = no. Therefore, plan personalization is carried out to find a plan fragment in the plan repository which fulfils the user’s constraint. The example in Fig. 10 shows Plan Fragment 5 in Plan A has C5 = no. Therefore, our system will use this plan fragment to replace Plan Fragment 4 in the generic plan provided other details are also met (e.g. pre-conditions, post-conditions, etc.).

#### IV. RESULTS: EXAMPLE CASES

Fig. 11 shows the command-line system interface which obtains inputs from the user to generate a generic and personalized plan in the domain of renal disease.

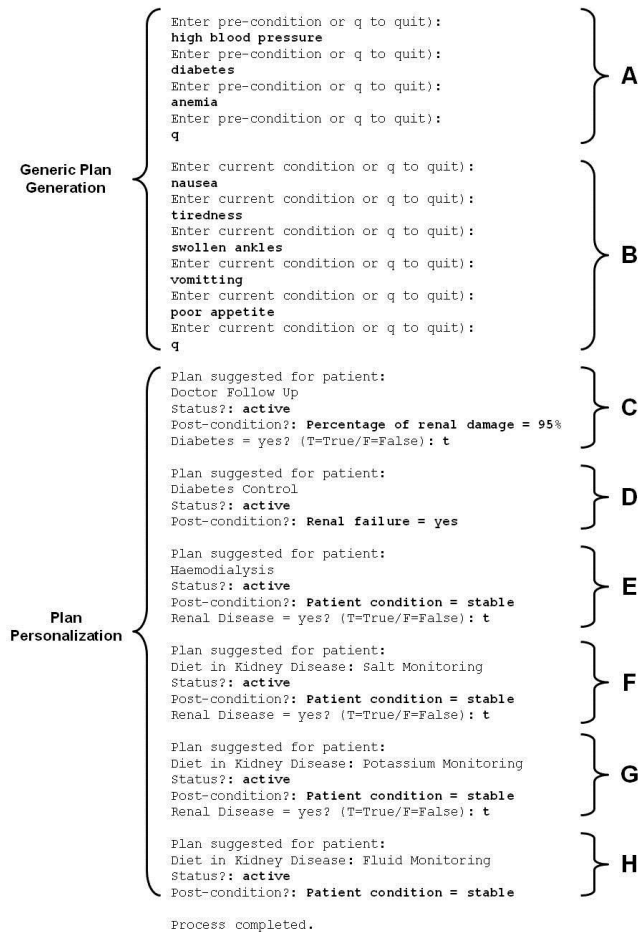


Figure 11. The process of finding the best-matched plan in the plan repository.

##### A. Generic Plan Generation

From Fig. 11, the interaction labeled A gets the inputs of pre-conditions while the interaction labeled B gets the inputs of current conditions from the user (bold text indicates text entered by the user). Both of these conditions will be used to match the pre-conditions and current conditions in the plans found in the plan repository. As described in Section III.B.1, this matching is carried out to find the plan that best matches the user’s pre-condition and current condition inputs, i.e. the

plan with the highest concentration of matches value (see Equation 1) is identified. Then, a copy of that plan is created as an XML file and assigned with a new ID. The values for the pre-conditions and current conditions are assigned with those inputted by the user while dates, times, and post-conditions are left empty.

##### B. Plan Personalization

After the generic plan has been generated, personalization then takes place. Firstly, the user will be presented with the first plan fragment in the plan as shown in Fig. 11 (labeled C). Then, the planning system will prompt the user about their status, i.e. whether active (user is implementing the plan) or not active (not implementing the plan). Either response will result in a different output for the user. Hence, we show the results of different cases in the following sub-sections. This process is repeated with other plan fragments (labeled D to H in Fig. 11).

###### 1) Case 1: Status is active

This case is for situations when the user is implementing a particular plan fragment. When this happens, the system will prompt for details about the user’s post-condition as shown in Fig. 12. Assuming that the user’s input for post-condition was *Percentage of renal damage = 95%* therefore, this input will be matched with the next plan fragment’s pre-condition. If they match, the constraints of the next plan fragment will be highlighted to the user: *Diabetes = yes?: (T=True/F=False)* as shown in Fig. 12. If the user fulfils the constraint, the next plan fragment will be presented. The output for this user will be generated as shown in Fig. 13.

```

Plan suggested for patient:
Doctor Follow Up
Status?: active
Post-condition?: Percentage of renal damage = 95%
Diabetes = yes? (T=True/F=False): t

Plan suggested for patient:
Diabetes Control
Status?: active
Post-condition?: Renal failure = yes

Plan suggested for patient:
Diabetes Control
Status: active
Post-condition: Renal failure = no

Process completed.
    
```

Figure 12. Part of the system interface when status of the user is “active”.

```

<Plan ID="7.0">
  <PlanFragment ID="7.1" Name="Doctor Follow Up">
    <Goal>Detect the stage of kidney disease</Goal>
    ...
  <Condition>
    <PreCondition>high blood pressure</PreCondition>
    <PreCondition>diabetes</PreCondition>
    <PreCondition>anemia</PreCondition>
    <CurrentCondition>nausea</CurrentCondition>
    ...
  </Constraints />
  <Status>Active</Status>
</PlanFragment>
</Plan>
    
```

Figure 13. Part of the system output when status of the user is “active”.

###### 2) Case 2: Status is not active

This case shows that the user is not implementing a particular plan fragment. When this is indicated by the user, the system will not perform any personalization in the subsequent plan fragments. Therefore, the planning process is deemed to

be complete. The output for the user will be generated as shown in Fig. 14. Here, the result is personalized by removing the subsequent plan fragments that are not necessary.

```
<Plan ID="8.0">
  <PlanFragment ID="8.1" Name="Doctor Follow Up">
    <Goal>Detect the stage of didney disease</Goal>
    <Date>
      ...
    </Date>
    <Time>
      ...
    </Time>
    <Constraints />
    <Condition>
      <PreCondition>diabetes</PreCondition>
      <CurrentCondition>nausea</CurrentCondition>
      <PostCondition />
    </Condition>
    <Constraints />
    <Status>Not active</Status>
  </PlanFragment>
</Plan>
```

Figure 14. Part of the system output when status of the user is "not active".

### 3) Case 3: Inputs do not fulfill conditions or constraints in the generic plan

This situation can be further divided into three sub-cases as follows:

1. Inputs which do not fulfill conditions in the generic plan.
2. Inputs which do not fulfill some conditions but fulfill constraints in the generic plan.
3. Inputs which do not fulfill constraints in the generic plan.

In the first case, consider the example system interface in Fig. 15. Assume that the post-condition in the generic plan is *Percentage of renal damage = 95%*, and this does not match with the user's input which is *Percentage of renal damage = 60%*. Therefore, the personalization process is terminated and the entire planning process is deemed to have completed. As a result, the planning system generates the output as shown in Fig. 16 as the personalized plan for this case.

```
Enter pre-condition or q to quit):
high blood pressure
Enter pre-condition or q to quit):
diabetes
Enter pre-condition or q to quit):
q

Enter current condition or q to quit):
nausea
Enter current condition or q to quit):
vomitting
Enter current condition or q to quit):
swollen ankles
Enter current condition or q to quit):
tiredness
Enter current condition or q to quit):
q

Plan suggested for patient:
Doctor Follow Up
Status?: active
Post-condition?: Percentage of renal damage = 60%

Process completed.
```

Figure 15. System interface for inputs which do not fulfill conditions in the generic plan.

```
<Plan ID="9.0">
  <PlanFragment ID="9.1" Name="Doctor Follow Up">
    <Goal>Detect the stage of kidney disease</Goal>
    <Date>
      ...
    </Date>
    <Time>
      ...
    </Time>
    <Constraints />
    <Condition>
      <PreCondition>high blood pressure</PreCondition>
      <PreCondition>diabetes</PreCondition>
      <CurrentCondition>nausea</CurrentCondition>
      <CurrentCondition>vomitting</CurrentCondition>
      <CurrentCondition>swollen ankles</CurrentCondition>
      <CurrentCondition>tiredness</CurrentCondition>
      <PostCondition>Percentage of renal damage = 60%</PostCondition>
    </Condition>
    <Constraints />
    <Status>Not active</Status>
  </PlanFragment>
</Plan>
```

Figure 16. Personalized plan for inputs which do not fulfill conditions in the generic plan.

In the second case, consider the example system interface in Fig. 17. Let us assume that the post-condition in the generic plan, i.e. *Patient condition = stable* (see Fig. 18), does not match the user's input which is *Patient condition = not stable* (Fig. 17). Therefore, the process terminates. As a result, the planning system generates the personalized plan as showed in Fig. 19.

```
Enter pre-condition or q to quit):
high blood pressure
Enter pre-condition or q to quit):
diabetes
Enter pre-condition or q to quit):
q

Enter current condition or q to quit):
nausea
Enter current condition or q to quit):
vomitting
Enter current condition or q to quit):
swollen ankles
Enter current condition or q to quit):
tiredness
Enter current condition or q to quit):
q

Plan suggested for patient:
Doctor Follow Up
Status?: active
Post-condition?: Percentage of renal damage = 95%
High blood pressure = yes? (T=True/F=False): t

Plan suggested for patient:
High Blood Pressure Control
Status?: active
Post-condition?: Renal failure = yes
Fistula operation = yes? (T=True/F=False): t

Plan suggested for patient:
Haemodialysis
Status?: active
Post-condition?: Patient condition = not stable

Process completed.
```

Figure 17. System interface for inputs which do not fulfill some conditions but fulfill constraints in the generic plan.

In the third case, consider the example system interface in Fig. 20. Let us assume that the user does not fulfill the constraint of *High blood pressure = yes*, the planning system would search the plan repository for a plan fragment which fulfils the user's input. This plan fragment is then retrieved and

used to replace the one in the generic plan which did not fulfill the user's input. Fig. 21 shows the personalized plan for this case.

```
<Plan ID="10.0">
  <PlanFragment ID="10.1" Name="Doctor Follow Up">
    ...
  <PlanFragment ID="10.2" Name="High Blood Pressure Control">
    ...
  <PlanFragment ID="10.3" Name="Haemodialysis">
    ...
    <Condition>
      <PreCondition>Renal failure = yes</PreCondition>
      <CurrentCondition>Patient condition = not stable</CurrentCondition>
      <PostCondition>Patient condition = stable</PostCondition>
    </Condition>
    ...
  <PlanFragment ID="10.4" Name="Diet in Kidney Disease: Salt Monitoring">
    ...
</Plan>
```

Figure 18. Generic plan for inputs which do not fulfill some conditions but fulfill constraints in the generic plan.

```
<Plan ID="10.0">
  <PlanFragment ID="10.1" Name="Doctor Follow Up">
    <Goal>Detect the stage of kidney disease</Goal>
    <Date>
      ...
    <Time>
      ...
    <Condition>
      <PreCondition>high blood pressure</PreCondition>
      <PreCondition>diabetes</PreCondition>
      <CurrentCondition>nausea</CurrentCondition>
      <CurrentCondition>swollen ankles</CurrentCondition>
      <CurrentCondition>tiredness</CurrentCondition>
      <CurrentCondition>vomitting</CurrentCondition>
      <PostCondition>Percentage of renal damage = 95%</PostCondition>
    </Condition>
    <Constraints />
    <Status>Active</Status>
  </PlanFragment>
  <PlanFragment ID="10.2" Name="High Blood Pressure Control">
    <Goal>To reduce and delay the kidney damage</Goal>
    <Date>
      ...
    <Time>
      ...
    <Condition>
      <PreCondition>Percentage of renal damage = 95%</PreCondition>
      <CurrentCondition>tiredness</CurrentCondition>
      <CurrentCondition>nausea</CurrentCondition>
      <CurrentCondition>vomitting</CurrentCondition>
      <CurrentCondition>swollen Ankles</CurrentCondition>
      <PostCondition>Renal failure = yes</PostCondition>
    </Condition>
    <Constraints>High blood pressure = yes</Constraints>
    <Status>Active</Status>
  </PlanFragment>
</Plan>
```

Figure 19. Personalized plan for inputs which do not fulfill some conditions but fulfill constraints in the generic plan.

```
Enter pre_condition or q to quit):
diabetes
Enter pre_condition or q to quit):
q

Enter current condition or q to quit):
nausea
Enter current condition or q to quit):
vomitting
Enter current condition or q to quit):
swollen ankles
Enter current condition or q to quit):
q

Plan suggested for patient:
Doctor Follow Up
Status?: active
Post_condition?: Percentage of renal damage = 95%
High blood pressure = yes? (T=True/F=False): f

Suggested treatment plan:
Glomerular Filtration (GFR) Test

Process completed.
```

Figure 20. System interface for inputs which do not fulfill constraints in the generic plan.

```
<Plan ID="11.0">
  <PlanFragment ID="11.1" Name="Doctor Follow Up">
    <Goal>Detect the stage of kidney disease</Goal>
    <Date>
      ...
    <Time>
      ...
    <Condition>
      <PreCondition>diabetes</PreCondition>
      <CurrentCondition>nausea</CurrentCondition>
      <CurrentCondition>vomitting</CurrentCondition>
      <CurrentCondition>swollen ankles</CurrentCondition>
      <PostCondition>Percentage of renal damage = 95%</PostCondition>
    </Condition>
    <Constraints />
    <Status>Active</Status>
  </Plan_Fragment>
  <PlanFragment ID="11.2" Name="Glomerular Filtration Test">
    <Goal>To reduce and delay the kidney damage</Goal>
    <Date>
      ...
    <Time>
      ...
    <Condition>
      <PreCondition>Percentage of Renal Damage = 95%</PreCondition>
      <CurrentCondition>nausea</CurrentCondition>
      <CurrentCondition>vomitting</CurrentCondition>
      <CurrentCondition>swollen ankles</CurrentCondition>
      <PostCondition>Renal failure = yes</PostCondition>
    </Condition>
    <Constraints />
    <Status>Not active</Status>
  </PlanFragment>
</Plan>
```

Figure 21. Personalized plan for inputs which do not fulfill constraints in the generic plan.

## V. DISCUSSION AND COMPARISON

In general, the generic and personalized plan generation processes performs up to expectation. However, as a limitation

of the DP Planner system, these processes would not function fully when a replacement cannot be found in the plan repository. When the system encounters this situation, it will advise the user to refer the case to a medical practitioner.

TABLE I. COMPARISON OF DP PLANNER WITH OTHER PLANNING SYSTEMS.

Features/Planning system	DP Planner	Asbru/Asgaard	O-Plan	I-X	Prodigy	STRIPS	PLANET	RAX-PS	
Plan ontology & representation	Adoption of a conceptual model for planning	Yes	Yes	No	No	No	No	No	
	Simplicity of the plan ontology	Yes	Yes	Yes	Yes	No	No	No	
	Simplicity of the plan representation	Yes	Yes	No	Yes	Yes	No	No	
Planning method	Planning approach	Non-linear	Non-linear	Non-linear	Non-linear	Non-linear	Linear	-	Non-linear
	Genericity	Domain-Configurable	Domain-Configurable	Domain-Configurable	Domain-Configurable	Domain-Configurable	Domain-Independent	-	Domain-Specific
	Application of case-based plan adaptation techniques	Yes	Yes	No	No	No	No	-	No
	Tolerance towards planning system execution failures	Yes	Yes	Yes	Yes	Yes	No	-	No

Table 1 shows the comparison between the DP Planner with other planning systems, i.e. Asgaard, O-Plan, Prodigy, STRIPS, PLANET and RAX-PS. From our observation, Asgaard and O-Plan are established planning systems that the DP Planner can be compared to in view that they have the relevant plan ontology, plan representation, as well as the planning engine for their planning system. Further comparisons with Asgaard and O-Plan are discussed in the following sub-sections.

### A. Plan Ontology Definition and Representation

Asgaard was inspired by Belief-Desire-Intention (BDI) model [14] while DP Planner was developed based on Goals-Operators-Method-Selection (GOMS) model. Using the BDI framework, Asgaard has been used to build large-scale, highly capable agent system [15]. Therefore, Asgaard is more suited for domains with large and complex but partly vague and incomplete knowledge. In contrast, DP Planner is based on the GOMS framework that has not been used to develop large-scale systems. GOMS has been mainly used to represent human knowledge necessary for performing certain tasks and complex human activities. As a result, DP Planner which is based on GOMS is more suited in domains with obvious knowledge, i.e. knowledge that is confirmed and complete, for performing certain tasks and knowledge.

Due to its simplicity, the DP Planner plan ontology was developed without the need for any ontology tool such as Protégé. The DP Planner plan representation in XML is also intuitively easy to understand. Asbru (which is Asgaard's plan representation language) uses a machine-readable language (Backus-Naur Form or BNF syntax) to annotate guidelines based on the task-specific ontology.

Asbru also requires the use of an ontology editor such as Protégé for the acquisition of clinical guidelines based on the same ontology and GMT (Guideline Markup Tool) to translate the guideline into a formal representation written in XML [16].

DP Planner's ontology is also easier and simpler compared to O-Plan which has its own detailed ontology structure. The O-Plan plan representation is in Task Formalism form and will change in different O-Plan agents in which it is situated. This is quite complex compared to DP Planner in which the plan representation remains in XML form in any situation.

### B. Planning Algorithm Definition

DP Planner generates a generic plan by retrieving an existing plan with similar characteristics to the current planning requirements, and adapting the generic plan by reusing existing plans to produce a personalized plan. This is akin to a case-based approach with adaptation. Asgaard employs a similar approach to DP Planner whereby it also applies plan adaptation in its planning process.

However, the difference is that Asgaard adopts the transformational type of adaptation whereas DP Planner adopts a derivational analogy type followed by the transformational type. Derivational analogy potentially reduces the search space by ignoring the unnecessary choices. This is achieved using the DP Planner's similarity measurement technique. This is only suitable in situations when most of the previous plans require extensive adaptation and when the cost of saving rationale is low [15]. The cost of saving rationale here means the ability to fulfill the requirements of a particular plan fragment that was defined as a constraint. However, it presupposes that the derivational traces exist. In contrast, when this is not possible, transformational analogy is the better choice because the plans themselves can be used for adaptation. Thus, in DP Planner, derivational analogy is applied in generic plan generation while transformational analogy is applied in the plan personalization in view that the cost of saving rationale is higher.

In general, Asgaard appears more robust in view that it is a fully deployed system, and that it has a monitoring component which monitors changes to the user's situation, while DP Planner is not a deployed system and therefore relies on users to report any changes.

O-Plan on the other hand is based on software agents and provides a hierarchical planning architecture to support planning and control with temporal and resource constraint handling [17]. O-Plan is also designed as a fully deployed system. O-Plan's architecture shows that it has five major components which are Domain Information, Knowledge Sources, Support Tools, Plan State, and Controller. O-Plan also has the agent architecture since it has a Task Assignment, Planner and Execution agent.

In O-Plan, its plan repair algorithm involves two tables (see Fig. 22): TOME (Table of Multiple Effect), and GOST (Goal Structure Table) [18]. An execution failure occurs when one or more of the expected effects at a node-end fail to be asserted. Each effect is recorded in the TOME and when an action depends on an effect asserted earlier, it is recorded in the GOST (Step 1).

When an execution failure occurs, the TOME will be updated and its relation with GOST entries will be found. If it is related with any of the GOST entries, then the Knowledge Sources is used to fix the problem (Step 2). The Knowledge Sources are responsible for determining the consequences of unexpected events or of actions that do not execute as intended, for deciding what action to take when a problem is detected, and for making repairs to the effected plan (Step 3 and 4) [17].

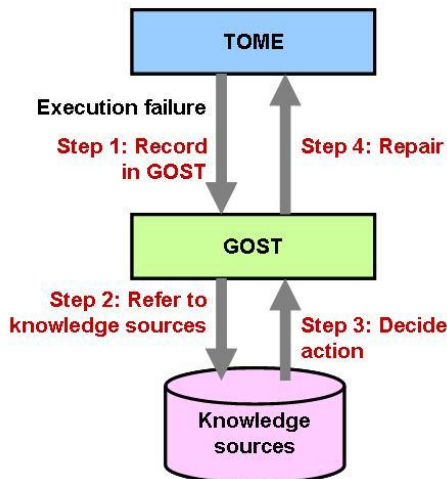


Figure 22. Solving execution failure in O-Plan [12].

When comparing the DP Planner's approach with that of O-Plan, it seems that the DP Planner approach is simpler as only two stages are needed to solve the failure whereas O-Plan requires four stages to solve the failure (see Fig. 23).

### VI. FUTURE WORK

Presently, the DP Planner is implemented in a local environment. Its capabilities can be extended further by deploying it in a grid environment with distributed plan repositories and planning engines.

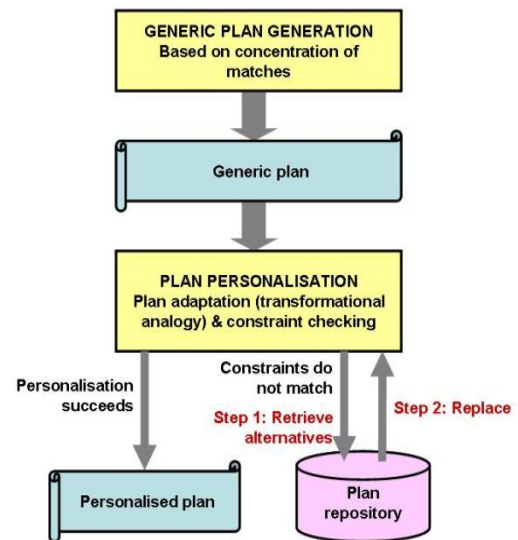


Figure 23. Solving execution failure in DP Planner.

For this purpose, the open source Globus Tool Kit can be utilized to allow sharing of computing power, databases and other tools online across corporate, institutional, and geographic boundaries autonomously and safely. The Globus Tool Kit is used in tandem with the Open Grid Services Architecture with Data Access and Integration (OGSADAI) which provides a service group registry that can be used to identify database services that offer specific data tables [19].

In order to communicate with OGSADAI, the DP Planner would potentially require middleware software to communicate with the DBMS which will store the planning data. Fig. 24 shows the OGSADAI with Globus in a possible DP Planner planning scenario. Here, the planning engines and plan repositories are distributed across different locations and each of these planning engines accesses the XML database containing the plan repository (or medical treatment plans) via Xindice (a suggested XML DBMS) while the planning results are delivered through the Client Tool Kit middleware whereby it provides the communication channel between the requesting node and the processing planning engine [20].

### VII. CONCLUSION

There are many types of planning systems currently available in the literature though most are static in nature. In this paper, we presented (1) a plan ontology and representation, and (2) a dynamic planning engine which makes use of generic plans and plan fragments, based on our plan ontology, as a planning template. The processing of the planning engine is based on similarity measure, matching, linking, and constraint techniques. In the comparative study carried out, it was observed that the DP Planner possesses features that rival that of other planning systems, in particular that of Asgaard and O-Plan. It is hoped that the DP Planner will make planning initiative more efficient and effective in delivering applicable plans to users especially to healthcare providers and patients.



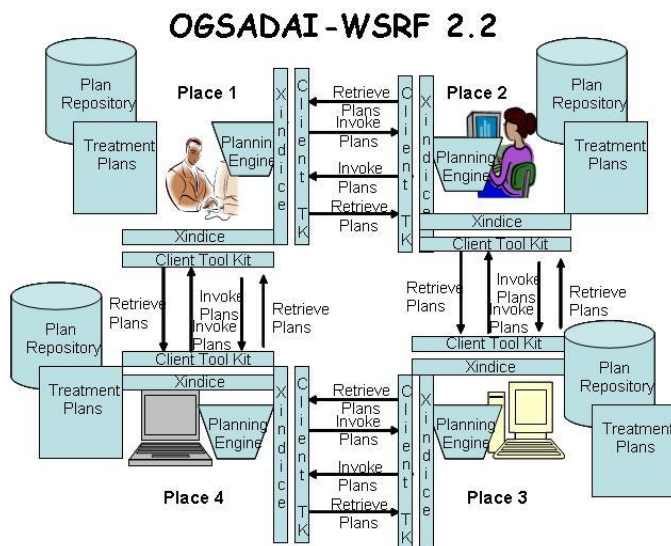


Figure 24. The DP Planner in a grid computing environment.

#### ACKNOWLEDGMENT

The authors wish to thank the Ministry of Higher Education, Malaysia and Universiti Sains Malaysia for funding this work through the Fundamental Research Grant Scheme (FRGS) under the project entitled "A Dynamic Planning Algorithm Using Ontologies and Constraints on the Grid".

#### REFERENCES

- [1] N. Mahiddin, "An Ontology and Constraint-based Approach for Dynamic Personalised Planning," Master Thesis, Universiti Sains Malaysia, 2009.
- [2] A. Tate, "Towards a plan ontology," Journal of the Italian Association for AI, AI\*IA Notizie, Special Issues on Aspects of Planning Research, vol. 9, no. 1, pp. 19-26, March 1996.
- [3] A. Tate, "A plan ontology - a working document," in Proceedings of the Workshop on Ontology Development and Use, La Jolle, CA, U.S.A., 1994.
- [4] Y. Shahar, S. Miksch, and P. Johnson, "A task-specific ontology for design and execution of time-oriented skeletal plans," in Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, 1996.
- [5] S. Miksch, Y. Shahar, and P. Johnson, "Asbru: a task specific, intention-based and time-oriented language for representing skeletal plans," in Proceedings of the Seventh Workshop on Knowledge Engineering Methods and Languages (KEML-97), Milton Keynes, U.K., 1997.
- [6] S. Miksch, A. Seyfang, and R. Kosara, "Plan management: supporting all steps of protocol development and deployment," in Proceedings of the EUNITE-Workshop on Intelligent Systems in Patient Care, Vienna, Austria, pp. 35-42, 2001.
- [7] A. K. Jónsson, P. H. Morris, N. Muscettola, and K. Rajan, "Planning in interplanetary space: Theory and practice," in Proceedings of the Fifth International Conference on Artificial Intelligence Planning and Scheduling (AIPS-2000), Breckenridge, CO, U.S.A., pp. 177-186, 2000.
- [8] J. Frank, A. K. Jónsson, and P. H. Morris, "On reformulating planning as dynamic constraint satisfaction," in Proceedings of the 4th International Symposium on Abstraction, Reformulation and Approximation, London, U.K.: Springer-Verlag, 2000.

- [9] M. S. Atkin, and P. R. Cohen, "Physical planning and dynamics," in Working Notes of the AAAI Fall Symposium on Distributed Continual Planning, Orlando, FL, U.S.A., pp. 4-9, 1998.
- [10] M. S. Atkin, and P. R. Cohen, "Using simulation and critical points to defines states continous search spaces," in Proceedings of Simulation Conference, Orlando, FL, U.S.A., vol. 1, pp. 464-470, 2000.
- [11] S. S. R. Abidi, Y. H. Chong, S. R. Abidi, "An intelligent info-structure for composing and pushing personalised healthcare information over the internet," in Proceedings of the 14th IEEE Symposium on Computer Based Medical Systems (CBMS 2001), Bethesda, MD, U.S.A., pp. 225-230, 2001.
- [12] D. Jonassen, M. Tessmer, and W. Hannum, Task Analysis Methods for Instructional Design, Mahwah, NJ: Lawrence Erlbaum Associates, 1999.
- [13] N. Mahiddin, Y.-N. Cheah, and F. Haron, "A generic plan ontology for dynamic health plans," in Proceedings of the International Conference of Knowledge Engineering 2005 (IKE '05), Las Vegas, NV, U.S.A., 2005.
- [14] S. Miksch, and A. Seyfang, "Continual planning with time-oriented, skeletal plans," in Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000), Amsterdam, The Netherlands: IOS Press, pp. 512-515, 2000.
- [15] H. M. Avila, and M. T. Cox, "Case-based plan adaptation: An analysis and review," IEEE Intelligent Systems, vol. 23, no. 4, pp.75-81, 2008.
- [16] Y. Shahar, S. Miksch, and P. Johnson, "The Asgaard project: A task specific framework for the application and critiquing of time-oriented clinical guidelines," Artificial Intelligence in Medicine, vol. 14, pp. 29-51, 1998.
- [17] O-Plan Team, O-plan: Architecture guide (version 2.3), 1995. URL: <http://www.aiai.ed.ac.uk/oplan/documents/ANY/oplan-architecture-guide.pdf>. Retrieved: 18 October 2011.
- [18] B. Drabble, A. Tate, and J. Dalton, "Repairing plans on-the-fly," in Proceedings of the NASA Workshop on Planning and Scheduling for Space, 1997.
- [19] Globus Project Team, Globus tool kit homepage, 2007. URL: <http://www-unix.globus.org/toolkit/about.html>. Retrieved: 18 October 2011.
- [20] Globus Project Team, Software components for grid systems and applications, 2007. URL: [http://www.globus.org/grid\\_software](http://www.globus.org/grid_software). Retrieved: 18 October 2011.
- [21] Felix, A. A., & Taofiki, A. A. (2011). On Algebraic Spectrum of Ontology Evaluation. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(7), 159-168.
- [22] Vanitha, K., Yasudha, K., & Soujanya, K. N. (2011). The Development Process of the Semantic Web and Web Ontology. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(7).

#### AUTHORS PROFILE

Normadiah Mahiddin received her B.Comp.Sc. (Hons) degree from Universiti Sains Malaysia in 2003, and M.Sc. (Computer Science) from the same university in 2009. She is currently a Ph.D. candidate at Universiti Kebangsaan Malaysia. Her research interests include knowledge management, intelligent systems, health informatics, and automated planning.

Yu-N Cheah received his B.Comp.Sc. (Hons) and Ph.D. degrees from Universiti Sains Malaysia in 1998 and 2002 respectively. He is currently lecturing at the School of Computer Sciences, Universiti Sains Malaysia. His research interests include knowledge management, intelligent systems, health informatics, and semantic technologies.

Fazilah Haron received her B.Sc. (in Computer Science) from the University of Wisconsin-Madison, U.S.A. and her Ph.D. from the University of Leeds, U.K. She is an Associate Professor at the School of Computer Sciences, Universiti Sains Malaysia and currently on secondment at Taibah University, Madinah, Saudi Arabia. Her research interests include modeling and simulation of crowd, parallel and distributed processing, and grid computing.



# Asynchronous Checkpointing and Optimistic Message Logging for Mobile Ad Hoc Networks

Ruchi Tuli

Affiliation 1 : Research Scholar, Department of Computer Science

Singhania University, Pachari Bari (Rajasthan) INDIA

Affiliation 2: Department of Computer Sc. & Engg., P.O  
Box 31387, Yanbu University College,  
Kingdom of Saudi Arabia

Parveen Kumar

Professor,

Department of Computer Science,  
Meerut Institute of Engineering & Technology, Meerut  
INDIA

**Abstract** - In recent years the advancements in wireless communication technology and mobile computing fueled a steady increase in both number and types of applications for wireless networks. Wireless networks can roughly be classified into cellular networks which use dedicated infrastructure (like base stations) and ad hoc networks without infrastructure. A Mobile Ad Hoc Network (MANET) is a collection of mobile nodes that can communicate with each other using Multihop wireless links without using any fixed infrastructure and centralized controller. Since this type of networks exhibits a dynamic topology, that is, the nodes move very frequently, it is hard to establish some intermittent connectivity in this scenario. Fault tolerance is one of the key issues for MANETs. In a cluster federation, clusters are gathered to provide huge computing power. Clustering methods allow fast connection and also better routing and topology management of mobile ad hoc networks (MANET). To work efficiently on such systems, networks characteristics have to be taken into account, for e.g. the latency between two nodes of different clusters is much higher than the latency between two nodes of the same cluster. In this paper, we present a message logging protocol well-suited to provide fault tolerance for cluster federations in mobile ad hoc networks. The proposed scheme is based on optimistic message logging.

**Keywords** – MANETs; clusterhead; checkpointing; pessimistic logging; fault tolerance; Mobile Host.

## I. INTRODUCTION

Wireless networks include infrastructure-based networks and ad hoc networks. Most wireless infrastructure-based networks are established by a one hop radio connection to a wired network. On the other hand, mobile ad hoc networks are decentralized networks that develop through self-organization [1]. The original idea of MANET started out in the early 1970s. At this time they were known as packet radio networks. Lately, substantial progress has been made in technologies like microelectronics, wireless signal processing, distributed computing and VLSI (Very Large Scale Integration) circuit design and manufacturing [2]. This has given the possibility to put together node and network devices in order to create wireless communications with ad hoc capability.

MANETs are formed by a group of nodes that can transmit and receive data and also relay data among themselves. Communication between nodes is made over wireless links. A

pair of nodes can establish a wireless link among themselves only if they are within transmission range of each other. An important feature of ad hoc networks is that routes between two hosts may consist of hops through other hosts in the network [3]. When a sender node wants to communicate with a receiver node, it may happen that they are not within communication range of each other. However, they might have the chance to communicate if other hosts that lie in-between are willing to forward packets for them. This characteristic of MANET is known as multihopping. An example is shown in figure 1. Node A can communicate directly (single-hop) with node B, node C and node D. If A wants to communicate with node E, node C must serve as an intermediate node for communication between them. Therefore, the communication between nodes A and E is multi-hop.

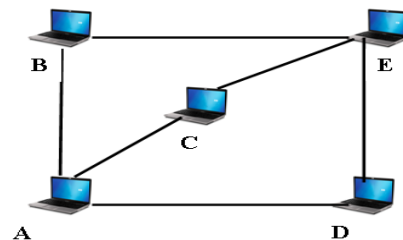


Figure 1 – Multi-hop communication in a mobile ad hoc network

Today wireless bluetooth, personal area networks (PAN), IEEE 802.11 a/b/g, wireless local area networks (WLAN) and HIPERLAN/2, are communication standards that include ad hoc features [4]. Figure 2 shows an example of a mobile ad hoc network composed of commonly used wireless devices.

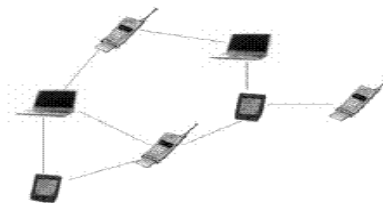


Figure 2 – Wireless Mobile Ad hoc Network

Checkpoint and message logging protocols are designed for saving the execution state of mobile application, so that when a MH recovers from a failure, the mobile application can roll back to the last saved consistent state, and restart execution with recovery guarantees. The existing protocols assume that the MH's disk storage is not stable and thus checkpoint and log information are stored at the base stations [5], [6].

Log-based rollback recovery exploits the fact that a process execution can be modeled as a sequence of deterministic state intervals, each starting with the execution of a non-deterministic event. A non-deterministic event can be the receipt of a message from another process or an event internal to the process. A message send event is not a non-deterministic event. Log-based rollback recovery assumes that all non-deterministic events can be identified and their corresponding determinants can be logged into the stable storage.

During failure-free operation, each process logs the determinants of all non-deterministic events that it observes onto the stable storage. Additionally, each process also takes checkpoints to reduce the extent of rollback during recovery. After a failure occurs, the failed processes recover by using the checkpoints and logged determinants to replay the corresponding non-deterministic events precisely as they occurred during the pre-failure execution. Because execution within each deterministic interval depends only on the sequence of non-deterministic events that preceded the interval's beginning, the pre-failure execution of a failed process can be reconstructed during recovery up to the first non-deterministic event whose determinant is not logged. The deterministic intervals composing the process execution are called state intervals. The state intervals are partially ordered by the Lamport's happen-before relation [7].

Message logging techniques are classified into pessimistic [8], optimistic [9], [10], [11] and causal [12], [13], [14], [15]. Pessimistic logging protocols assume that a failure can occur after any non-deterministic event in the computation. This assumption is "pessimistic" since in reality failures are rare. In their most straightforward form, pessimistic protocols log to the stable storage the determinant of each non-deterministic event before the event affects the computation. A pessimistic protocol is one in which each process  $p$  never sends a message until it knows that all messages delivered before sending the previously sent messages are logged. Pessimistic protocols will never create any inconsistent process (orphans), and so the reconstruction of the state of a crashed process is very straightforward. The pessimistic protocols potentially block a process for each message it receives.

In optimistic logging protocols, processes log determinants asynchronously to the stable storage. These protocols optimistically assume that logging will be complete before a failure occurs. Determinants are kept in a volatile log, and are periodically flushed to the stable storage. Thus, optimistic logging does not require the application to block waiting for the determinants to be written to the stable storage, and therefore incurs much less overhead during failure-free execution. However, the price paid is more complicated recovery, garbage collection, and slower output commit. If a process fails, the determinants in its volatile log are lost, and the state intervals

that were started by the non-deterministic events corresponding to these determinants cannot be recovered. Furthermore, if the failed process sent a message during any of the state intervals that cannot be recovered, the receiver of the message becomes an orphan process and must roll back to undo the effects of receiving the message. Optimistic logging protocols do not implement the always-no-orphans condition.

Causal logging combines the advantages of both pessimistic and optimistic logging at the expense of a more complex recovery protocol. Like optimistic logging, it does not require synchronous access to the stable storage except during output commit. Like pessimistic logging, it allows each process to commit output independently and never creates orphans, thus isolating processes from the effects of failures at other processes. Moreover, causal logging limits the rollback of any failed process to the most recent checkpoint on the stable storage, thus minimizing the storage overhead and the amount of lost work. Causal logging protocols make sure that the always-no-orphans property holds by ensuring that the determinant of each non-deterministic event that causally precedes the state of a process is either stable or it is available locally to that process.

In this paper we focus on optimistic based message logging for communications in a clustered ad hoc network, since the checkpointing-only schemes are not suitable for the mobile environment and also in ad hoc environments in which unreliable mobile hosts and fragile network connection may hinder any kind of coordination for checkpointing and recovery. In order to cope with the storage problem, the task of logging is assigned to the CH instead of MHs, since each message heading to a MH is routed through the CH. Also, in order to reduce the overhead imposed on mobile hosts, cluster heads take charge of logging and dependency tracking, and mobile hosts maintain only a small amount of information for mobility tracking.

The rest of this paper is organized as follows: Section 2 discusses related work and problem formulation. In section 3, system model is described. Section 4 explains the basic algorithm and comparison with existing schemes is described in section 5. Finally section 6 concludes the paper.

## II. RELATED WORK AND PROBLEM FORMULATION

### A. Related Work

Application failure recovery in the mobile computing environment has received considerable attention in the recent years. The schemes that have been proposed employ checkpointing, logging or a combination of both, recognizing the inherent limitations of the mobile computing environments. Since the requirements of an ad hoc network are different from the mobile computing environment, these issues need to be addressed in this area as well.

Prakash and Singhal describe in [16] a checkpointing algorithm for Mobile Computing System. Checkpoint collection is synchronous and non-blocking. A minimum number of nodes are forced to take checkpoints. Each MH maintains a dependence vector. MHs maintain causal relationships through message. This scheme reduces energy consumption by

powering down individual components during periods of low activity.

In [17] T.Park et.al has presented an efficient movement based recovery scheme. This scheme is a combination of message logging and independent checkpointing. Main feature of this algorithm is that a host carrying its information to the nearby MSS can recover instantly in case of a failure. To enhance failure-free execution, concept of a 'certain range' is introduced. An MH moving inside a range, recovery information remains in host MSS otherwise it moves recovery information to nearby MSS. Though recovery is ensured, failure-free execution cost increases. Due to this out of range concept overheads due to transfer of checkpoint from one MSS to another MSS increases many fold.

Sapna E. George [18] et.al describes a checkpointing and logging scheme based on mobility of MHs. A checkpoint is saved when hand-off count exceeds a predefined optimum threshold. Optimum threshold is decided as a function of MH's mobility rate, failure rate and log arrival rate. Recovery probability is calculated and recovery cost is minimized in this scheme.

Acharya et al. [6] describes uncoordinated checkpointing, where multiple MHs can arrive at a global consistent checkpoint without coordination messages. However, neither it takes into account how failure recovery is achieved nor does it address the issue of recovery information management in the face of MH movement.

In [19] authors proposed an independent checkpointing scheme which saves the state of processes in the computer to which a mobile host is currently attached.

The authors in [20] presents a low overhead recovery scheme based on a communication induced checkpointing, which allows the processes to take checkpoints asynchronously and uses communication-induced checkpoint coordination for the progression of the recovery line. The scheme also uses selective pessimistic message logging at the receiver to recover the lost messages. However, the recovery scheme can handle only a single failure at a time.

P. Kumar and A. Khunteta [22] proposed a minimum-process coordinated checkpointing algorithm for deterministic mobile distributed systems, where no useless checkpoints are taken, no blocking of processes takes place, and anti-messages of very few messages are logged during checkpointing. In their algorithm they have tried to reduce the loss of checkpointing effort when any process fails to take its checkpoint in coordination with others.

### B. Problem Formulation

Cluster federations are hierarchical systems. The latency between two clusters is much higher than the latency between two nodes of the same cluster. For efficient execution on such systems, applications must take into account the topology of the cluster federation. Communications between nodes of the same cluster should be favored over communications between nodes of different clusters.

The objective of the present work is to design an optimistic based message logging for communications in a clustered ad

hoc network, since the checkpointing-only schemes are not suitable for the mobile environment and also in ad hoc environments. In order to cope with the storage problem, the task of logging is assigned to the CH instead of MHs, since each message heading to a MH is routed through the CH. In order to reduce the size of dependency information carried in each message for asynchronous recovery, only the messages between the CHs carry the information, and the dependency between the MHs residing in the same Cluster can be traced through the message order within the CH. Using the restricted dependency tracking, no extra overhead is imposed on MHs. Of course, there is a possibility of unnecessary rollbacks due to the imprecise dependency information, however, comparing with the checkpointing-only schemes, the chance of rollback propagation in the message logging schemes is very low.

### III. SYSTEM MODEL

A successful approach for dealing with the maintenance of mobile ad hoc networks is by partitioning the network into clusters. In this way the network becomes more manageable. Clustering is a method which aggregates nodes into groups. These groups are contained by the network and they are known as clusters. Clusters are analogous to cells in a cellular network. However, the cluster organization of an ad hoc network cannot be achieved offline as in fixed networks [21]. In most clustering techniques nodes are selected to play different roles according to a certain criteria. In general, three types of nodes are defined:

*Ordinary nodes* :- Ordinary nodes are members of a cluster which do not have neighbors belonging to a different cluster.

*Gateway nodes*:- Gateway nodes are nodes in a non-clusterhead state located at the periphery of a cluster. These types of nodes are called gateways because they are able to listen to transmissions from another node which is in a different cluster. To accomplish this, a gateway node must have at least one neighbor that is a member of another cluster.

*Clusterheads*:- Most clustering approaches for mobile ad hoc networks select a subset of nodes in order to form a network backbone that supports control functions. A set of the selected nodes are called clusterheads and each node in the network is associated with one. Clusterheads are connected with one another directly or through gateway nodes. The union of gateway nodes and clusterheads form a connected backbone. This connected backbone helps simplify functions such as channel access, bandwidth allocation, routing power control and virtual-circuit support.

Clusterheads are analogous to the base station concept in current cellular systems. They act as local coordinators in resolving channel scheduling and performing power control. However, the difference of a clusterhead from a conventional base station resides in the fact that a clusterhead does not have special hardware, it is selected among the set of stations and it presents a dynamic and mobile behavior. Since clusterheads must perform extra work with respect to ordinary nodes they can easily become a single point of failure within a cluster. For this reason, the clusterhead election process should consider for the clusterhead role, those nodes with a higher degree of

relative stability. The main task of a clusterhead is to calculate the routes for long-distance messages and to forward inter-cluster packets. Figure 3 shows the system model and different roles of nodes in a mobile ad hoc network organized by clusters.

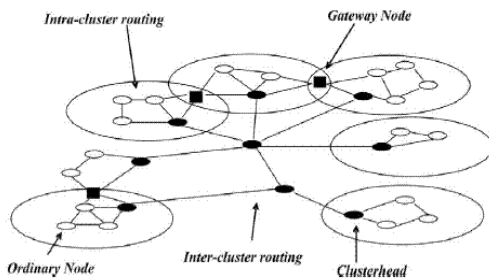


Figure: - 3 System Model

The clustering system considered in this paper follows the model presented in the figure above. The system is organized into various clusters, each having a clusterhead and ordinary nodes, which will be termed as Mobile Hosts; a set of dynamic links can be established between a MH and a Cluster head. The area covered by a cluster head is called a *cell*. A MH residing in a cluster can be connected to the clusterhead servicing the cluster and the MH can communicate to another MH only through the cluster head. The links in the dynamic network support FIFO communication in both directions.

For a MH to leave a cluster and enter into another cluster, it first has to end its current connection by sending a *leave(x)* message to the cluster head, where *x* is the sequence number of the last message received from the cluster head, and then establish a new connection by sending *join* (MH-id, previous cluster-id, previous clusterhead-id) message to the new cluster head in the new cluster. Each cluster head maintains a list of identifiers, called a *Current\_Nodelist*, with which nodes it connected at current time.

A MH can also disconnect itself from the cluster head voluntarily without leaving the cluster by sending a *disconnect(x)* message to conserve power. When the cluster head receives a *disconnect* message from a Node, it marks that node to be “disconnected” by setting a flag that maintains a list of voluntarily disconnected MHs, called *disconnected\_Nodelist*. Later on, the MH can *reconnect* to any cluster by sending a *reconnect* (MH-id, previous cluster-id, previous clusterhead\_id) message to the current cluster head. If the MH is reconnected to a new cluster, the new cluster head informs the previous cluster head of the reconnection so that the previous cluster head can perform the proper hand-off procedures.

#### IV. PROPOSED ALGORITHM

The ordinary Nodes in the clustered ad hoc networks are considered highly vulnerable to failures, while the cluster heads are relatively reliable as they are chosen among the nodes with a higher degree of relative stability. By reliable CH, we mean that recovery information for MHs can never be lost due to its own failure. With this assumption, the volatile memory space

of a CH can be utilized as a stable storage to save checkpoints and message logs of MHs.

##### A. Data Structures and Notations

$Chk_i^x$  = checkpoint sequence number

$i=0 \dots n$

$j=0 \dots n$

$MH_i$  = No. of Mobile hosts

$i=0 \dots n$

CH = Cluster head

$m_i^{rcv}$  = No. of messages a mobile host has received

$i=0 \dots n$

$rcv=0 \dots n$

$Locate_i$  = variable to retrieve information after failure of MH

$chk\_seq$  = sequence number of latest checkpoint

$cp\_loc$  = current ID of a cluster

$cp\_ch$  = current cluster head ID

$msg\_seq$  = sequence number of the first message logged after checkpoint

$log\_set$  = IDs of cluster heads that store MH logs

$timeToCkp_i$  = Timer to take checkpoint on MH

##### B. Checkpointing and Message Logging

Each mobile host MH independently takes a checkpoint and a unique sequence number is assigned to each checkpoint. For the checkpointing, MH first saves its current state as a checkpoint and then transfers the checkpoint to CH to which it is currently connected.  $Chk_i^x$  denotes the *x*th checkpoint of  $MH_i$ . Each checkpoint is identified by a pair of (*i*, *x*).  $MH_i$  then sends the checkpoint to its current CH, say  $CH_p$ . Each Mobile host also maintains a variable  $m_i^{rcv}$ , to count the number of messages a mobile host has received, and the value of  $m_i^{rcv}$  is sent with the checkpoint to the cluster head. On the receipt of a checkpoint and other related information CH saves it into stable storage. The value of  $m_i^{rcv}$  sent with the checkpoint to CH is used to decide the correct position of the checkpoint with respect to logged messages.

Each  $CH_p$  also maintains the message log for the nodes residing in that cluster. Since each message that is delivered to MH in the cluster is routed through CH, logging of messages incurs little overhead. Let  $Msg_i^x$  denotes *x*th message delivered to  $MH_i$ . In addition,  $CH_p$  also logs the messages related to the mobility of the nodes, such as join, leave, disconnect and reconnect. Any of these messages sent from MH must also carry the value of  $m_i^{rcv}$  and sequence number is logged with message.

A node after a failure should be able to locate its latest checkpoint and the message log for recovery. Each mobile host also contains a variable  $Locate_i$ , to retrieve the information after failure.  $Locate_i$  contains  $chk\_seq$ ,  $cp\_loc$ ,  $cp\_ch$ ,  $msg\_seq$  and  $log\_set$ .  $chk\_seq$  stores the sequence number of latest checkpoint and  $cp\_loc$  stores the ID of the cluster and  $cp\_ch$  stores the ID of the CH that has recorded the latest checkpoint. Let this cluster be called  $cluster_{in}$  and CH be called  $CH_p$ .  $msg\_seq$  denotes the sequence number for the first message logged after the checkpoint and  $log\_set$  contains the IDs of the cluster heads that stores its logs. At every checkpoint,  $cp\_loc$  is updated with the current CH,  $cp\_seq$  is updated with the

sequence number of the latest checkpoint and  $log\_set$  is cleared. At every logging activity, the IDs of the current CHs are added to  $log\_set$  if it is not present already.  $Locate_i$  is logged by the CHs which a MH has visited. When a mobile host joins or reconnects to a new cluster, say  $CH_p$ , it sends  $Locate_i$  with the connection message. Also when mobile host disconnects itself from or leaves  $CH_p$ , it sends  $Locate_i$  with disconnection message if information in the  $Locate_i$  has been changed since the connection was established.  $CH_p$  on receipt of  $Locate_i$  logs it with the message. Each mobile host also maintains a variable  $timeToCkp_i$  which defines Time interval until next checkpoint

### C. Algorithm

We describe pseudocode for the checkpointing and message logging protocol here

### D. Checkpointing at MH

```
If (  $timeToCkp_i = \text{Expire}$  ) then
     $chk\_seq_i = chk\_seq_i + 1$ ; //increment checkpoint
    sequence number
    Perform checkpointing ,  $Chk_i^{chk\_seq_i}$ 
    Save (  $i, chk\_seq_i, m_i^{rcv}$  ) with  $Chk_i^{chk\_seq_i}$ 
    // updating the  $Locate$  field

     $Locate_i.chk\_seq = chk\_seq_i$ ;
     $Locate_i.cp\_loc = cluster_{in}$ ;
     $Locate_i.cp\_ch = p$ ;
     $Locate_i.msg\_seq = m_i^{rcv} + 1$ ;
     $Locate_i.log\_set = \text{NULL}$ ;
    Send (  $Chk_i^{chk\_seq_i}[i, chk\_seq_i, m_i^{rcv}]$  ) to  $CH_p$ 
Else
    Continue computation;
If (  $CH_p = \text{rcv } Chk_i^{chk\_seq_i}$  ) from MH
    Save (  $Chk_i^{chk\_seq_i}[i, chk\_seq_i, m_i^{rcv}]$  );
Else
    Continue computation
```

### E. Message logging at CH

```
a) When cluster head delivers a message M, to Mobile
host
     $msg\_seq_i = msg\_seq_i + 1$ ;
    Insert (  $M_i^{msg\_seq_i}[i, msg\_seq_i]$  ) into Log;

b) When Mobile host receives a message from cluster
head (  $CH_p$  )
    If (  $p \notin Locate_i.log\_set$  )
         $Locate_i.log\_set = Locate_i.log\_set \cup p$ ;

c) When Mobile host sends a message to Cluster head
    If (  $M \in \{join, leave, disconnect, reconnect\}$  )
        Send (  $M [Locate_i, m_i^{rcv}]$  )

d) When Cluster head receives a message M, from
mobile host
    If (  $M \in \{join, leave, disconnect, reconnect\}$  )
        Insert (  $M (Locate_i, m_i^{rcv})$  ) into log;
```

### F. Proof of Correctness

**Theorem I :-** If a MH fails, its state can be reconstructed independently

**Proof :-** Let  $MH_i$  state be  $[s_i^0, s_i^1, s_i^2 \dots s_i^l]$  before failure, which indicates messages  $e_i^0, e_i^{x-1}, e_i^x, \dots, e_i^y$ , where  $1 \leq y$ ,  $e_i^x$  is the first message from the last checkpoint and  $e_i^y$  is the last message before failure. Since all the messages delivered to  $MH_i$  are logged in CH and  $Locate_i.log\_set$  indicates the order in which  $MH_i$  has contacted CH since its last checkpoint. After a failure  $MH_i$  should rollback to the latest checkpoint and the logged messages in the same order and it can reconstruct the same state intervals as the ones before failure. Because all the messages sent and received events are recorded, the  $MH_i$ 's state can be reconstructed.

### V. HANDLING FAILURES AND DISCONNECTIONS

We distinguish here failures and disconnections. Failures can be categorized as – mobile host falls and is damaged, lost or stolen, battery is discharged. Disconnections are termed as hand-off. Since the mobility rate of mobile hosts in ad hoc networks is very high, so while connected a mobile host can change its position and can join another cluster. This movement is termed as disconnection. We will discuss these two issues separately.

After a MH recovers from a failure, either a mobile host is in the same cluster or the cluster can be changed. When a  $MH_i$  recovers from a failure, it first sends a  $join(MH\text{-id, previous cluster-id, previous clusterhead-id})$  to its current CH, say  $CH_p$ .  $CH_p$  checks its  $Active\_nodelist$  list and  $Disconnected\_nodelist$ . If  $MH_i$  is found in any of these lists that mean  $MH$  after a failure has recovered in current cell and thus the  $Locate_i$  must have been logged in  $CH_p$ .

Sometimes it may happen that CH is not able to find  $MH_i$  either in  $Active\_nodelist$  list or  $Disconnected\_nodelist$ , which means that  $MH$  after a failure has moved to another cluster. In this case, firstly  $MH$  sends a  $join(MH\text{-id, previous cluster-id, previous clusterhead-id})$  to its current CH, say  $CH_q$ . Now,  $CH_q$  will broadcast the recovery message, so that previous CH which has been contacted by  $MH_i$  can deliver the most recent  $Locate_i$  to  $CH_q$ . After  $CH_q$  receives the most recent  $Locate_i$ , it starts with the recovery procedure. During the recovery of  $MH_i$ , new messages heading to  $MH_i$  can be logged at  $CH_q$ . However, these messages are delivered to  $MH_i$  after consuming all the messages in the log. Only the  $MH$ s which have failed rollback to the latest checkpoint and replay the logged messages to ensure the global recovery line and no other  $MH$ s need to rollback.

A node after a disconnection should be able to locate its latest checkpoint and the message log for recovery. Each mobile host contains a variable  $Locate_i$ , to retrieve the information after failure which we have discussed in section 4.1. For each hand-off or disconnection, a  $MH$  within a cluster transfers the checkpoint and message logs to the current cluster head, so that the recovery information can be retrieved later from the cluster head. For a  $MH, MH_i$ , connected to the cluster head  $CH_p$ , in cluster  $CL_i$  first saves its checkpoints and message logs and updates the information in  $Locate_i$ . Let

$MH\_data_{(i,p)}$  denotes the checkpoints and message logs of  $MH_i$  saved by  $CH_p$  of  $CL_i$ . When  $MH_i$  leaves  $CL_i$  and joins another cluster head say  $CH_{new}$  of cluster  $CL_k$ , a hand-off procedure is initiated by  $CH_{new}$  sending a handoff-request for  $MH_i$  to  $CH_p$ . While the hand-off procedure is performed, the recovery information is transferred from  $CH_p$  to  $CH_{new}$ . Figure 4 depicts the sequence of events that take place for the recovery information transfer.

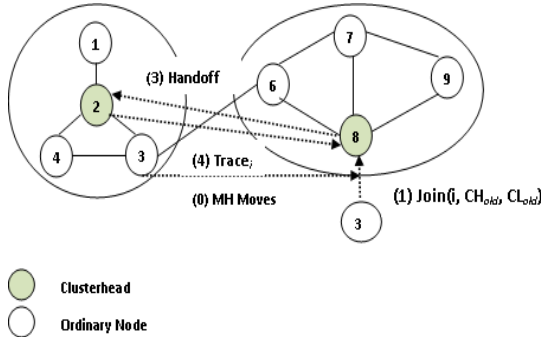


Figure 4 : Handling failures and disconnections

## VI. PERFORMANCE COMPARISON

In [1] authors proposed a communication pattern based checkpointing scheme to save consistent global states, in which a checkpoint is taken whenever a message reception is preceded by a message transmission. An independent checkpointing scheme which saves the state of processes in the computer to which a mobile host is currently attached was proposed by authors in [12]. Neither of the above approaches needs the checkpointing coordination, however, they may enforce a large number of checkpoints. [13] proposed a low-cost synchronous checkpointing scheme, in which a process can advance its checkpoint asynchronously, however, it may result in considerable message overhead and an inconsistency. [10] presents a low overhead recovery scheme based on a communication induced checkpointing, which allows the processes to take checkpoints asynchronously and uses communication-induced checkpoint coordination for the progression of the recovery line. The scheme also uses selective pessimistic message logging at the receiver to recover the lost messages. However, the recovery scheme can handle only a single failure at a time. In [18] authors describe a checkpointing and logging scheme based on mobility of MHs. A checkpoint is saved when hand-off count exceeds a predefined optimum threshold. Optimum threshold is decided as a function of MH's mobility rate, failure rate and log arrival rate. Recovery probability is calculated and recovery cost is minimized in this scheme.

We have described an optimistic based message logging scheme since checkpointing-only schemes are not suitable for ad hoc environments and most of the schemes described above are based on checkpointing-only approach. Also, we have followed the asynchronous checkpointing approach, as asynchronous recovery is desirable in ad hoc environments in which MH can be disconnected any time from the network and co-ordination may not be possible.

## VII. CONCLUSION

In this paper, we have proposed an optimistic based message logging approach for cluster based ad hoc networks in which each MH in the cluster takes checkpoint independently. Also, each message that is delivered to MH in the cluster is routed through CH which avoids the overhead of message logging at MH. MH only carries minimum information and all the dependency tracking and mobility of MH can be properly traced by CH. The asynchronous checkpointing scheme relieves the MH from any kind of coordination and they can take their checkpoints whenever they want.

## REFERENCES

- [1] C. Prehofer, C. Bettstetter. "Self organization in communication networks: Principles and design paradigms". *IEEE Communications Magazine*. Vol. 43. Issue 7. 2005. pp. 78-85.
- [2] Y. P. Chen, A. L. Liestman, J. Liu. *Ad Hoc and Sensor Networks, Wireless Networks and Mobile Computing. Clustering Algorithms for Ad hoc Wireless Networks*. Vol. 2. Chapter 7: Nova Science Publishers, Hauppauge NY, 2004. pp. 145-164. 7. J. Wu, J. Cao. "Connected k-hop clustering in ad hoc networks". *ICPP*. 2005. pp 373-380.
- [3] I. Chatzigiannakis, S. Nikolettseas. "Design and analysis of an efficient communication strategy for hierarchical and highly changing ad-hoc mobile networks". *Mobile Networks and Applications*. Vol. 9. 2004. pp. 248-263.
- [4] M. Frodigh, P. Johansson, P. Larsson. "Wireless Ad Hoc Networking--The Art of Networking without a Network". *Ericsson Review*. Vol. 77. 2000. pp. 248-263.
- [5] S. Gadiraju, Vijay Kumar, "Recovery in the mobile wireless environments using mobile agents", *IEEE Transactions on Mobile Computing*, June 2004, Vol. 3.
- [6] A. Acharya, B. R. Badrinath, "Checkpointing distributed applications on mobile computers", in *Proc. 3rd Int. Conf. Parallel and Distributed Information Systems*, Austin, Texas, 1994, pp. 73-80.
- [7] L. Lamport. *Time, Clocks, and the Ordering of Events in a Distributed System*. *Communications of the ACM*, 21(7):558-565, 1978.
- [8] D. B. Johnson and W. Zwaenpoel, "Sender-Based Message Logging," In *Digest of Papers: 17th International Symposium on Fault-Tolerant Computing*, pp. 14-19, 1987.
- [9] O. P. Damani and V. K. Garg, "How to Recover Efficiently and Asynchronously when Optimism Fails," In *Proc. the 16th International Conference on Distributed Computing Systems*, pp. 108-115, 1996.
- [10] D. B. Johnson and W. Zwaenpoel, "Recovery in distributed systems using optimistic message logging and checkpointing," In *Proc. the 7th Annual ACM Symposium on Principles of Distributed Computing*, pp. 171-181, 1988.
- [11] R. B. Strom and S. Yemeni, "Optimistic recovery in distributed systems," *ACM Transactions on Computer Systems*, Vol.3, No.3, pp. 204-226, 1985.
- [12] L. Alvisi, B. Hoppe, and K. Marzullo, "Nonblocking and Orphan-Free Message Logging Protocols," In *Proc. the 23th Symposium on Fault-Tolerant Computing*, pp. 145-154, 1993.
- [13] L. Alvisi and K. Marzullo, "Message Logging: Pessimistic, Optimistic, Causal and Optimal," *IEEE Transactions on Software Engineering*, Vol.24, No.2, pp. 149-159, 1998.
- [14] E. Elnozahy, "On the relevance of Communication Costs of Rollback Recovery Protocols," In *Proc. the 15th ACM Symposium on Principles of Distributed Computing*, pp. 74-79, 1995.
- [15] E. N. Elnozahy and W. Zwaenpoel, "Manet: Transparent rollback-recovery with low overhead, limited rollback and fast output commit," *IEEE Transactions on Computers*, Vol.41, No.5, pp. 526-531, 1992.
- [16] R. Prakash, M. Singhal, (1996) "Low Cost Checkpointing and Failure Recovery in Mobile Computing Systems", *IEEE Transactions on Parallel and Distributed Systems*, VOL. 7, NO. 10, OCTOBER 1996



- [17] Taesoon Park, Namyoon Woo, Heon Y. Yeom, (2003) "An Efficient recovery scheme for fault tolerant mobile computing systems", *Future Generation Computer System*, 19(1): 37-53
- [18] Sapna E. George, Ing-Ray Chen, Ying Jin, (2006) "Movement-Based Checkpointing and Logging for Recovery in Mobile Computing Systems", *MobiDE*, 51-58
- [19] D.K. Pradhan, P. Krishna, and N.H. Vaiday. Recoverable mobile environment : Design and trade-off analysis. In *Proc. of the 26th Int'l Symp. on Fault Tolerant Computing*, 1996.
- [20] D. Manivannan and M. Singhal. Failure recovery based on quasi-synchronous checkpointing in mobile computing systems. OSU-CISRC-796-TR36, Dept. of Computer and Information Science, The Ohio State University, 1996.
- [21] A. B. McDonald, T. F. Znati. "A mobility-based framework for adaptive clustering in wireless Ad Hoc networks". *IEEE Journal on Selected Areas in Communications*. Vol. 17. 1999. pp. 1466-1487.
- [22] Parveen Kumar, Ajay Khunteta, "Anti-message Logging based coordinated checkpointing protocol for Deterministic Mobile Computing Systems", *International Journal of Computer Applications* (0975-887), Vol. 3-No. 1, June, 2010.
- [23] Parveen Kumar, "A Low-Cost Hybrid Coordinated Checkpointing Protocol for Mobile Distributed Systems", *Mobile Information Systems* [An International Journal from IOS Press, Netherlands] pp 13-32, Vol. 4, No. 1, 2007. [Listed in ACM Portal & Science Citation Index Expanded]
- [24] Lalit Kumar, Parveen Kumar "A Synchronous Checkpointing Protocol for Mobile Distributed Systems: A Probabilistic Approach", *International Journal of Information and Computer Security* [], pp 298-314, Vol. 3 No. 1, 2007. [An International Journal from Inderscience Publishers, USA, Listed in ACM Portal]
- [25] Sunil Kumar, R K Chauhan, Parveen Kumar, "A Minimum-process Coordinated Checkpointing Protocol for Mobile Computing Systems", *International Journal of Foundations of Computer science*, Vol 19, No. 4, pp 1015-1038 (2008). [ Listed in Science Citation Index Expanded ]
- [26] Parveen Kumar, Lalit Kumar, R K Chauhan, "A Hybrid Coordinated Checkpointing Protocol for Mobile Computing Systems", *IETE journal of research*, Vol. 52, Nos 2&3, pp 247-254, 2006. [Listed in Science Citation Index Expanded ]
- [27] Lalit Kumar, Parveen Kumar, R.K. Chauhan, "Logging based Coordinated Checkpointing in Mobile Distributed Computing Systems", *IETE Journal of Research*, vol. 51, no. 6, pp. 485-490, 2005. [Listed in Science Citation Index Expanded ]
- [28] Obaida, M. A., Faisal, S. A., & Roy, T. K. (2011). AODV Robust ( AODV R ): An Analytic Approach to Shield Ad-hoc Networks from Black Holes. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(8), 97-102.
- [29] Journal, I., Science, A. C., & Hod, M. (2011). A Survey on Attacks and Defense Metrics of Routing Mechanism in Mobile Ad hoc Networks. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(3), 7-12.
- [30] Indukuri, R. K. R. (2011). Dominating Sets and Spanning Tree based Clustering Algorithms for Mobile Ad hoc Networks. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(2).

# An Experimental Improvement Analysis of Loss Tolerant TCP (LT-TCP) For Wireless Network

Md. Abdullah Al Mamun  
Dept. of CSE, DUET  
Gazipur – 1700, Bangladesh

Momotaz Begum  
Lecturer, Dept. of CSE, DUET, Gazipur – 1700 Bangladesh

Sumaya kazary  
Asst. Prof., Dept. of CSE, DUET, Gazipur – 1700 Bangladesh

Md. Rubel  
Dept. of CSE, DUET  
Gazipur – 1700, Bangladesh

**Abstract**—Now-a-days TCP is a famous protocol used in Internet but the main problem is packet losses due to congestion. In this thesis we proposed a new Loss Tolerant TCP (LT-TCP), an enhancement of TCP which makes it robust and applicable for extreme wireless environment. In the proposed LT-TCP two additional term, data and data header compression are added in existing LT-TCP. We reduce the total volume of data and packet size in our adaptive method which is able to minimize the congestion and increase the reliability of wireless communication. The ECN respond about random data packet loss and disruption process. The overhead of Forward Error Control FEC is imposed just-in-time process and target to maximize the performance even if the path characteristics are uncertain. This proposal show that it will perform better over regular TCP and it is possible to reduce packet losses up to 40-50%.

**Keywords**- *LT-TCP; Mix reliability; Timeouts; Congestion avoidance; end-to-end Algorithms; Compression technique; Packet erasure rate; TCP SACK; RFEC; PFEC; RAR; ECN; throughput; RTT; Goodput.*

## I. INTRODUCTION

TCP is the most popular protocol to deliver data reliably regardless of the form and construction of the network [1]. When a data packet traverse a wireless link, a major fraction of packet losses due to transmission error [2]. TCP is robust in that it can adapt to disparate network conditions [4, 7]. LT-TCP uses an adaptive, end-to-end hybrid Automatic Repeat request/Forward Error Correction ARQ/ FEC [3, 4] reliability strategy and exploits Explicit Congestion Notification ECN for incipient congestion detection [5]. Congestion control is the major problem of managing network traffic where the total demand of resources such as bandwidth among the computing users exceeds the available capacity [8]. In terms of packet loss is occurred more often due to high Bit Error Rates (BERs) than due to congestion [9]. When using TCP over wireless network, it considers each packet loss as a sign of congestion and invokes congestion control measures at the source [6]. This results in severe performance degradation. To improve the TCP performance such as WLAN using congestion response, mix of reliability mechanism, time out avoidance and handle a large volume of data [10]. It studies as extensive literature on the performance of TCP and emphasis the ways distinguish the effects due to congestion. It address the TCP congestion avoidance and control issue over the wireless links from the

end-to-end communication, thereby leading to efficient network resource utilization and improving the application response time. It also analyze widely used TCP end-to-end algorithm and presents a new technique that enables the TCP to better adapt to the wireless environment [11].

We provision a data compression technique which reduce the volume of data [12], hence reduce the total number of packets and proactive FEC in the original window as a function of the estimate of the actual packet erasure rate. Subsequently reactive FEC is used to mitigate the effects of erasures, during the transmission phase. An adaptive maximum segment size (MSS) provides a minimum number of packets in the TCP window, again seeking to risk of timeouts [10]. We seek to adaptively balance the FEC and packet overhead. While reducing the risk timeouts and also rapidly erased packets [3].

## II. PROPOSED LT-TCP

### A. Overview of Flow Graph

In accordance to Fig: 1 application data are providing compression technique which yield compressed data. This compressed data broken into TCP segments where the MSS is chosen to accommodate proactive FEC (PFEC) packets in the window. Reactive FEC (RFEC) packets are computed at the same time and held in reserve. Feedback from the receiver provides not only the loss estimate but also information (e.g. SACK blocks) that can be used to compute the number of RFEC packets to send for each block [10]. When the sender receives ACK, it determines the type of packets to send (Data/PFEC/RFEC) and transmits them. This provides self-clocking and follows the semantics of TCP behavior. LT-TCP comprises the following building blocks that complement each other and extend SACK to provide resilience. PFEC and RFEC help to data recovery. PFEC operates in conjunction with adaptive MSS and determined by the current estimates of loss rate. RFEC is computed based on feedback from the receiver and the loss rate estimate. The receiver can reconstruct the data packets as soon as any k out of n packets arrives at the receiver.

### B. Data and Header Compression

Compression is a technique which reduces the volume of original information [7, 12]. This is done either to reduce the volume of size of text and images to be transmitted or to reduce the bandwidth that is required for its transmission audio and

video. It required less bandwidth during transmission it is also secured because of its encrypted form.

**C. RAR Compression Technique**

RAR is a powerful allowing you to manage and control archive files.

Data Type	Original Data Size	Compressed Data Size	Number Of Segment (MSS 1500 Bytes)	
			Before	After
Text	10 KB	4.3 KB	7	4
	10 MB	3.5 MB	6990	2567
	100 MB	47.9 MB	69905	33745
Image	100 KB	69.5 KB	65	46
	1MB	789 KB	687	535
	10 MB	8 MB	6989	5800
Audio	10 MB	8.9 MB	6991	5786
	100 MB	89.6 MB	69905	62633
	1 GB	985 MB	715800	688565
Video	10 MB	8.9 MB	6990	6222
	100 MB	91.9 MB	69905	64733
	1 GB	999 MB	715698	697444

Compressed output for different data type [12] Console RAR [12] supports archives only in RAR format, the names of which usually have a “.rar” extension. ZIP and other format are not supported.

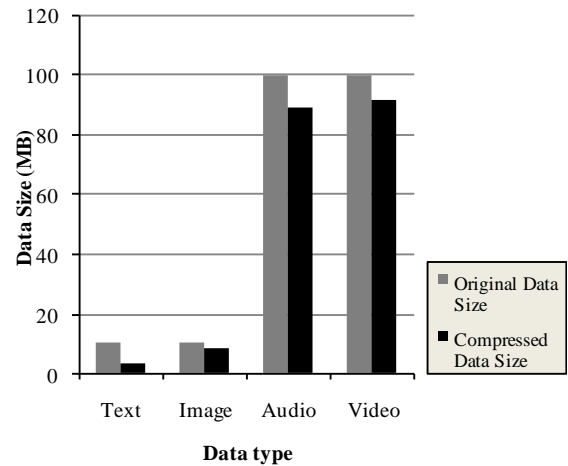


Figure 1. Compression comparison in terms of data size

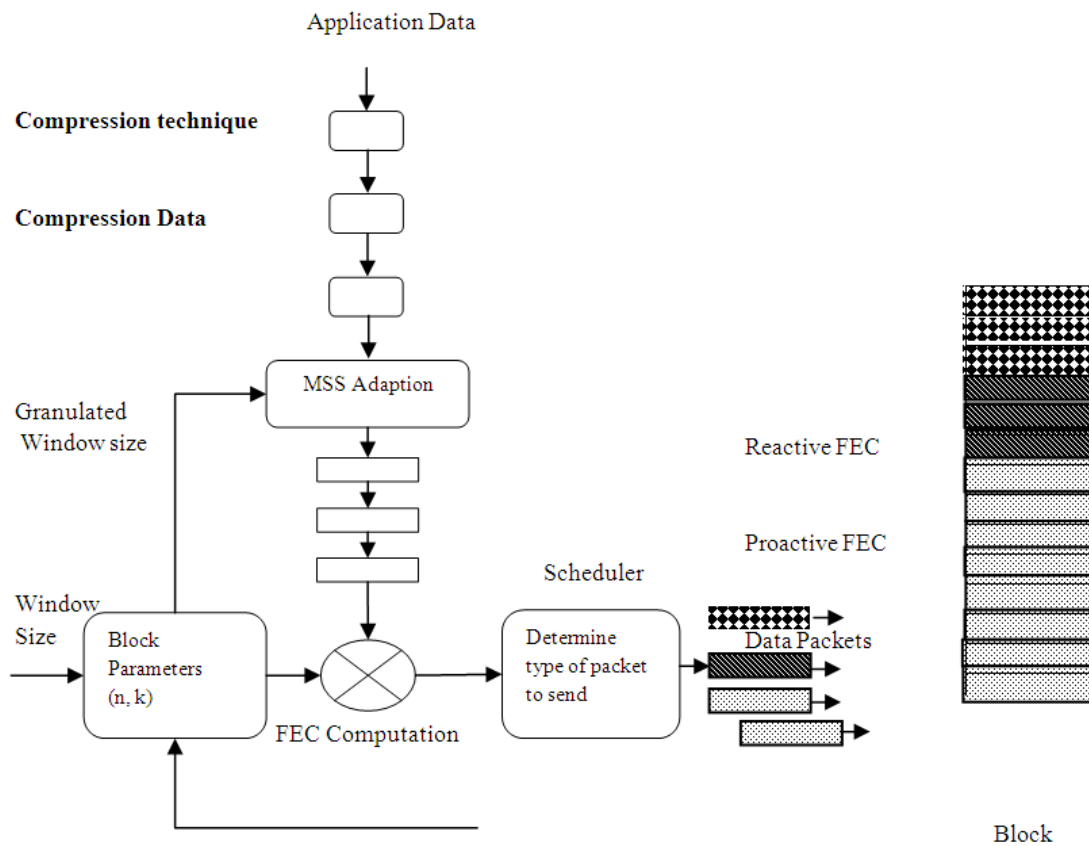


Figure 2. Proposed LT-TCP Flow graph

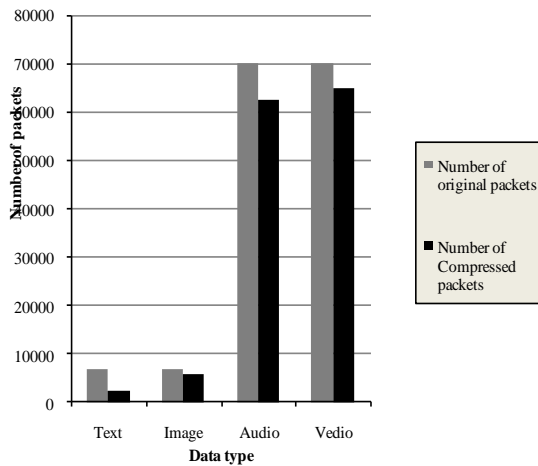


Figure 3. Compression comparison in terms of number of packets.

#### D. Van Jacobson TCP/IP Header Compression

The IPv4 header is 20 bytes and when carrying UDP (8 bytes) and RTP (12 bytes) at least the packet header become 40 bytes. A compression scheme usually compresses such header to 2-4 bytes. On and average, considering a few uncompressed packets and few relatively large packets, more than 80% savings [7] can be observed. When the compared with the payload being carried, in such cases as voice type where payload size is usually static in range of 20-60 bytes, the header size presents a huge overhead. In header compression in such cases results in major bandwidth savings. The IPv6 with a header size of 40 bytes is gaining wide acceptance and has been included in Release 5 and onward version of 3G wireless network. In this case, header compressions yield in even more savings.



Figure 4. Abstract view of header compression [13].

TABLE I. THE HEADER COMPRESSION GAINS [13].

Protocol Headers	Total header size (bytes)	Min. compressed header size (bytes)	Compression gain (%)
IPV4/TCP	40	4	90
IPV4/UDP	28	1	96.4
IPv4/UDP/RTP	40	1	97.5
IPv6/TCP	60	4	93.3
IPv6/UDP	48	3	93.75
IPv6/UDP/RTP	60	3	95

These benefits lead to improved *QoS* in the network and the possibility for operators to improve their average revenue per user. The operators will be able to retain and attract customers with better *QoS* on the network and more services and content the links.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

### III. PERFORMANCE ANALYSIS

#### A. Data & Header Compression Comparison

We use single bottleneck test case fig 5 with 10 flows and erasure rates varying from 0-50%. Host is ECN enabled; bottleneck implements RED/ECN on a 250 KB buffer (i.e. up to 500 packets of size of 500 bytes). *minthres* and *maxthres* values are as shown. According to table the 1, 2 and 3 we observed the performance for uncompressed data and compressed data & header both are given below

TABLE II. PERFORMANCE OF UNCOMPRESSED DATA

Parameter	LT-TCP		
	10 ms	80 ms	200 ms
Goodputs (MB/s)	5.4	4.5	3.9
Number of timeout	17	10	5
Throughput (MB/s)	7.4	6.3	5.1

TABLE III. PERFORMANCE OF COMPRESSED DATA

Parameter	LT-TCP		
	10 ms	80 ms	200 ms
Goodputs (MB/s)	7.3	6.5	5.1
Number of timeout	12	7	3
Throughput (MB/s)	9.5	7.3	6.4

From the above tables we observed that we compressed the data and header it mitigate the congestion and packet losses. The flow congestion is lasting for 100s for each operation. To assess the contribution of LT-TCP components, we use a 30% PER test case. Metrics include aggregate throughput, goodputs, number of timeouts and congestion window dynamics. We account for all packet header overheads.

#### B. Proposed LT-TCP vs. Traditional LT-TCP

In terms of packet erasure both perform well. However, the performance of traditional LT-TCP drops to 10% sometimes even more. In addition an error rate (40%) is sufficient to break a single LT-TCP connection due to repeated timeouts but proposed LT-TCP can reduce it significantly. Fig 6 shows different error rates for a number of RTT that means the degradation of performance is linear. Proposed LT-TCP manages packet error rates by avoiding timeouts and able to recover lost packets using proactive and reactive FEC even if RTT is too high.

#### C. Performance Evaluation

Finally we observe the Goodputs of both LT-TCP are shown in table IV. As we can see the proposed LT-TCP's Goodputs is higher than the traditional LT-TCP and performance is increasing with respect in time. It perform maximum at 80ms.

TABLE IV. PERFORMANCE OF PROPOSED LT-TCP

Parameter	LT-TCP		
	10 ms	80 ms	200 ms
LT-TCP	5.4	4.5	3.9
Proposed LT-TCP	7.3	6.5	5.1
Increased performance (%)	19	20	12

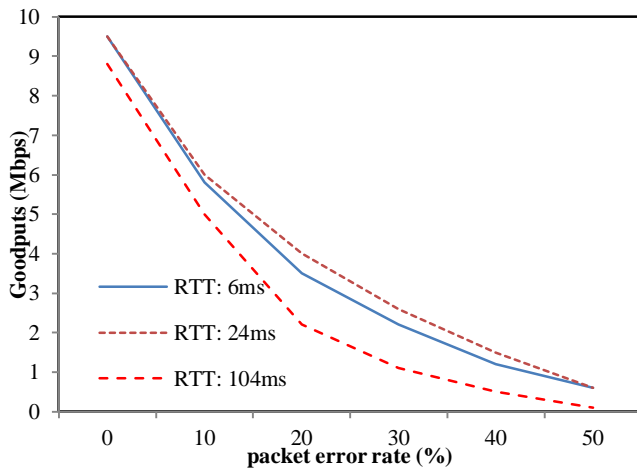


Figure 5. LT-TCP performance with increase erasure rate and RTT

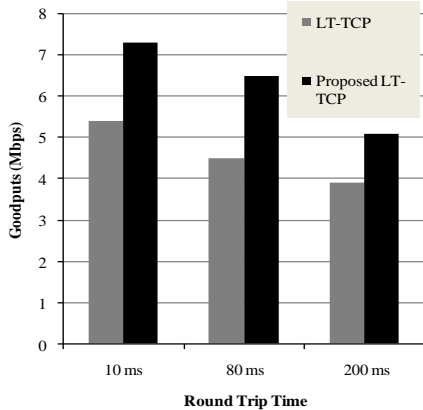


Figure 6. Performance graph of proposed LT-TCP

#### IV. CONCLUSION

Our TCP is the dominant reliable protocol used in internet; we have proposed a Loss-Tolerant TCP (LT-TCP) which introduces additional mechanisms as data & header compression in an adaptive manner. Our enhancement allow good performance even under demanding conditions through reduce the total volume of data that made less number of packet and packet size which mitigate congestion and perform better than traditional LT-TCP.

#### REFERENCES

- [1] Computer Networks by Andrew S. Tanenbum.
- [2] "TCP PERFORMANCE ENHANCEMENT OVER WIRELESS NETWORKS" By Aiyathurai ir.canterbury.ac.nz/bitstream/10092/1229/1/thesis\_fulltext.pdf.
- [3] S. Biswas, G.Judd, D. Aguayo, J. Bicket and Morris: Link-level measurements from an 802.11b mesh network. In SIGCOMM, Aug'02
- [4] C. Ladas, R. M. Edwards AMIEE, M. Mahdavi, and G. A. Manson (2002): Class based selective ARQ Scheme for high performance TCP and UDP over wireless links. Paper presented at the mobile and wireless network, 2002. 4th international workshop on it.
- [5] K. K. Ramkrishnan S. Floyd and D. Black: The addition of Explicite Congestion Notification (ECN) to IP, Sept 2001.
- [6] C. Barakat and E. Altman: Bandwidth tradeoff between TCP and link level FEC. Computer Networks 39(2): 133-150, June 2002.
- [7] S. Casner and V. Jacobson: Compressing IP/UDP/RTP Headers for low speed serial links. RFC 2508, Feb 1999. IETF Network working group.
- [8] RFC 793 (1981). Transmission control protocol. From <http://www.ietf.org/rfc/rfc793.txt>.
- [9] Data Communication and Networking by behrouz A. Forouzan.
- [10] Omesh Tickoo, Vijaynarayanan Subramanian, Shivkumar Kalyanaraman and K. K. Ramakrishnan: LT-TCP: End-to-End Framework to Improve TCP Performance over Networks with Lossy Channels. In thirteen international workshop on Quality of Service (IWQoS 2005), Passau, Germany 2005.
- [11] CHIU D. M and JAIN R.: analysis of increase and decrease algorithms for congestion avoidance in computer networks. Computer Networks and ISDN system, Vol 17 June, pp. 1-14
- [12] Compression principles techniques and tools RAR.
- [13] "White paper: An introduction to IP header compression" from EFFNET AB.

#### AUTHORS PROFILE



Md. Abdullah al mamun obtained his Bachelor of Science in Engineering degree from Department of Computer Science and Engineering (CSE) of Dhaka University of Engineering & Technology (DUET), Gazipur-1700, Bangladesh. At present he is performing extensive research on Network and Web Security, Wireless networking, Software Architecture, Machine Vision, Artificial Intelligence, Protocol cryptography. His key research interest includes Cryptography, analysis and Algorithm designs, E-mail: aamcse@gmail.com.



Momotaz Begum, Lecturer, Department of CSE, DUET and she achieved her B. Sc. in Engineering degree from Department of Computer Science and Engineering (CSE), DUET, Bangladesh. E-mail: momotaz03\_duet@yahoo.com



Sumaya kazary, Asst. Prof., Department of CSE, DUET and she achieved her B. Sc. in Engineering degree from Department of Computer Science and Engineering (CSE), DUET, Bangladesh. E-mail: kazal\_duet@yahoo.com

# Comparison of Workflow Scheduling Algorithms in Cloud Computing

Navjot Kaur  
CSE Department, PTU Jalandhar  
LLRIET  
Moga, India

Taranjit Singh Aulakh  
CSE Department, PTU Jalandhar  
BGIET  
Sangrur, India

Rajbir Singh Cheema  
IT Department, PTU Jalandhar  
LLRIET  
Moga, India

**Abstract**— Cloud computing has gained popularity in recent times. Cloud computing is internet based computing, whereby shared resources, software and information are provided to computers and other devices on demand, like a public utility. Cloud computing is technology that uses the internet and central remote servers to maintain data and applications. This technology allows consumers and businesses to use application without installation and access their personal files at any computer with internet access. The main aim of my work is to study various problems, issues and types of scheduling algorithms for cloud workflows as well as on designing new workflow algorithms for cloud Workflow management system. The proposed algorithms are implemented on real time cloud which is developed using Microsoft .Net Technologies. The algorithms are compared with each other on the basis of parameters like Total execution time, Execution time for algorithm, Estimated execution time. Experimental results generated via simulation shown that Algorithm 2 is much better than Algorithm 1, as it reduced makespan time.

**Keywords**- Cloud Computing; Workflows; Scheduling; Makespan; Task ordering; Resource Allocation.

## I. INTRODUCTION

Cloud computing is Internet-based computing, whereby shared resources, software and information are provided to computers and other devices on-demand, like a public utility. Cloud computing is a technology that uses the internet and central remote servers to maintain data and applications. Cloud computing allows consumers and businesses to use applications without installation and access their personal files at any computer with internet access. This technology allows for much more efficient computing by centralizing storage, memory, processing and bandwidth.

### A. Workflows

The WfMC (Workflow Management Coalition) defined a workflow as “the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules.”

WfMC published its reference model in [1], identifying the interfaces within this structure which enable products to interoperate at a variety of levels. This model defines a workflow management system and the most important system interfaces (see Fig 1).

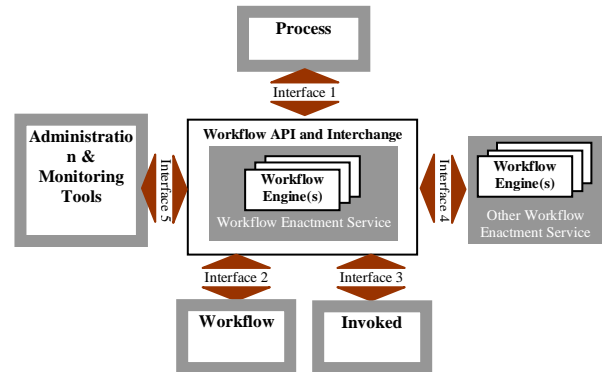


Fig. 1 WfMC's Workflow Reference Model

- 1) *Workflow Engine*. A software service that provides the run-time environment in order to create, manage and execute workflow instances.
- 2) *Process Definition*. The representation of a workflow process in a form which supports automated manipulation.
- 3) *Workflow Interoperability*. Interfaces to support interoperability between different workflow systems.
- 4) *Invoked Applications*. Interfaces to support interaction with a variety of IT applications.
- 5) *Workflow Client Applications*. Interfaces to support interaction with the user interface.
- 6) *Administration and Monitoring*. Interfaces to provide system monitoring and metric functions to facilitate the management of composite workflow application environments.

It can be seen that scheduling is a function module of the Workflow Engine(s), thus it is a significant part of workflow management systems.

The rest of the paper is structured as follows: Related work is discussed in Section II. Then section III describes our Proposed Work. The Implementation is presented in Section IV. And Section V will show the experimental details and simulation results. Finally Section VI includes the future scope of our research work.

## II. RELATED WORK

### A. Cloud Platforms

A comprehensive survey of cloud computing is defined by number of researchers. There are no. of definitions of cloud



computing. According to R. Buyya and S.Venugopal[5] Cloud computing is defined as “ a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements established through negotiation between the service provider and consumers”.

Sun Microsystems [3] takes an inclusive view that there are many different types of clouds like public cloud, private cloud, hybrid cloud .Many different applications that can be built by using these different clouds.

Recently, several academic and industrial organizations have started investigating and developing technologies and infrastructure for Cloud Computing.

**B. Workflow Management Systems**

Workflow is concerned with the automation of procedures whereby files and data are passed between Participants according to a defined set of rules to achieve an overall goal. A workflow management system defines, manages and executes workflows on computing resources. Workflow Scheduling: workflow scheduling is a kind of global task scheduling as it focuses on mapping and managing the execution of inter-dependent tasks on shared resources that are not directly under its control. Workflow management includes five dimensions: *time, cost, fidelity, reliability* and *security*.

The related work done in workflow management system is shown below in tabular form (see Table III):

TABLE III  
SURVEY ON WORKFLOW MANAGEMENT SYSTEM

S. No.	Citation	Brief Introduction About Paper
1.	<b>Authors:</b> S. Elnikety, E. Nahum, J. Tracey, W. Zwaenepoel <b>Year:</b> 2004	This paper [10] consider workflows that are invoked via web requests. The workflows are part of a web application that spans multiple resources in the grid.
2.	<b>Authors:</b> Jia Yu, Rajkumar Buyya & Chen Khong Tham <b>Year:</b> 2005	In this paper, a cost-based workflow scheduling algorithm is proposed [11] that minimizes execution cost while meeting the deadline for delivering results.
3.	<b>Author:</b> E. Deelman, G. Singh, D.S. Katz <b>Year:</b> 2005	Pegasus [12], is proposed which is a framework that maps complex scientific workflows onto distributed resources such as the Grid. DAGMan, together with Pegasus, schedules tasks to Condor system.
4.	<b>Author:</b> Jia Yu, Rajkumar Buyya <b>Year:</b> 2006	A budget constraint based scheduling is proposed [13], which minimizes execution time while meeting a specified budget for delivering results. A new type of genetic algorithm is developed to solve the scheduling optimization problem and the scheduling algorithm is tested in a simulated Grid tested.
5.	<b>Author:</b> Patel, Y. Darlaington <b>Year:</b> 2006	According to this paper, there are two categories of workflow scheduling [14]. The first category is based on the real time data such as waiting time in the queue or the shortest remaining execution length. The second category is based on average

		metrics such as mean arrival time, or mean execution length.
6.	<b>Author:</b> P. Patala, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singal <b>Year:</b> 2007	A control system is developed that adjusts the resource sharing among applications to ensure the desired QoS and maintains the high resource utilization [15].
7.	<b>Author:</b> J. Yu & R. Buyya <b>Year:</b> 2007	The work presented in this paper defines two major types of workflow scheduling [16], best-effort based and QoS constraint based scheduling, primarily for grid workflow management systems.
8.	<b>Author:</b> Zhifeng Yu & Weisong Shi <b>Year:</b> 2008	In this paper, a planner-guided strategy is presented for multiple workflows [17]. It ranks already tasks and decides which task should be scheduled.
9.	<b>Author:</b> Ke Liu, Jin Jun chin, Yun Yang & Hai Jin <b>Year:</b> 2008	A throughput maximization strategy is proposed [18] for scheduling transaction intensive workflows. But it is designed for transaction intensive workflows not for multiple workflows.
10.	<b>Author:</b> Meng Xu, Lizhen Cui, Haiyang Wang, Yanbing Bi <b>Year:</b> 2009	Authors of this paper worked on Multiple workflow and Multiple QOS.A strategy [19] is implemented for Multiple Workflow Management system with multiple QOS. The Scheduling access rate is increased by using this strategy.
11.	<b>Author:</b> Boris Mejias, Peter Van roy <b>Year:</b> 2010	In this paper, researchers proposed an architecture to organize a set of mini-clouds provided by different institutions, in order to provide a larger cloud that appears to its users as a single one[20].
12.	<b>Author:</b> J. Kosinska, J. Kosinski, K. Ziehnski <b>Year:</b> 2010	Purpose of this paper is to discuss various forms of mapping cluster topology requirements into cloud environments to achieve higher reliability & scalability of application executed within cloud resources[21].
13.	<b>Author:</b> M. Jensen, J.schwenk, J.M. Bohli, L.L. Iacono <b>Year:</b> 2011	This paper initiates discussion by contributing a concept which achieves security merits by making use of multiple distinct clouds at the same time[22].

**III. PROPOSED WORK**

This section presents a set of scheduling algorithms, based on Time management [23]. The aim of the algorithms is to optimize the makespan, which is defined as the maximum time taken for the completion of all the tasks in a given application. The proposed algorithms are implemented using a service based cloud and comparative results are shown.

The problem of scheduling a set of tasks to a set of processors can be divided into two categories:

- Job scheduling
- Job mapping and scheduling

In the former category, independent jobs are to be scheduled among the processors of a distributed computing system to optimize overall system performance. In contrast, the mapping and scheduling problems requires the allocation of multiple interacting tasks of a single parallel program in order to minimize the completion time on the parallel computer system.

To generate the schedule, our technique is based on the traditional list scheduling approach in which we construct a list and schedule the nodes on the list one by one to the processors.

**A. Algorithm 1**

The design of our algorithm 1 is basis on the following heuristics. It is based on the POSEC method [23]. POSEC is an acronym for Prioritize by Organizing, Streamlining, Economizing and Contributing. The objective of our algorithms is efficient time management and load balancing.

There are Four Quadarnrts of Descion Making :

- a) *Level 1: Low Urgency & Low Importance*
- b) *Level 2: Low Urgency & High Importance*
- c) *Level 3: High Urgency & Low Importance*
- d) *Level 4: High Urgency & High Importance*

There are Four Quadarnrts of Descion Making : It needs two types of Priority Scores to take descion , Urgency Score and Importance Score. Urgency Score given by Cluster Member of cloud. Importance Score is given by Cloud Resources Manager .

Urgency Score is Calculated on the scale of 10 on the basis of the following table.

TABLE IV  
CLASSIFICATION OF URGENCY LEVELS

Level of Severity	Description of Severity	Initial Respo nse Withi n	Score
Level I	Production application down or major malfunction causing business revenue loss resulting in majority of users unable to perform their normal functions	1 hour	$7.6 \leq \text{Score} \leq 10$
Level II	Critical loss of application functionality or performance resulting in high number of users unable to perform their normal functions	4 hours	$5.1 \leq \text{Score} \leq 7.5$
Level III	Moderate loss of application functionality or performance resulting in multiple users impacted in their normal functions	8 hours	$2.6 \leq \text{Score} \leq 5.0$
Level IV	Minor loss of application functionality or product feature question	24 hours	$1 \geq \text{Score} \leq 2.5$

Importance Score is Calculated by the Resource manager and its also on the scale of 10. The various parameter of resource cheking are CPU time. Threades etc. we have use resource monitor program to generate the importance score. High Importance means the Resources are available. Low Importance means based the Resources are Not available .

It is assumed that job consist of tasks. The cloud scheduler assigns these tasks to resources. Also it is assumed that each computational resource can run one application at a time, and must run that application to completion.

Let T be a set of n tasks and m is the number of computational resources in a cloud. We define a schedule of T as follows: A schedule S of T onto a cloud with m resources is a finite set of tuples  $\langle v, p, t \rangle$  where v is the schedule, t is the starting time, and p is the resource.

To generate the schedule, our technique is based on the traditional list scheduling approach in which we construct a list and schedule the nodes on the list one by one to the processor. The list is constructed by ordering the jobs according to their urgency score s. The list is static therefore the order of nodes on the list will not change during the resource allocation process.

We restrict ourselves to non-preemptive schedules where a job once started has to run to completion on the same machine.

Scheduler has information about all resources such as processing speed (in MIPS), processing cost per second, baud rate(communication rate) and resource load during peak hours and off peak hours.

After gathering the details of user jobs, the system calculated the importance score. The jobs are executed on the values of urgency and importance score.

The time management parameters used by the algorithms are:

- 1) *Total Execution Time: The total time consumed by the algorithm to execute all the jobs.*
- 2) *Execution Time of Algorithm: This is the time taken by the algorithm to execute.*
- 3) *Estimated Execution Time: Based on the average of total execution time parameters of previous jobs.*

The proposed algorithm comprises of two parts as explained below.

A. Task Ordering Procedure, to get the schedule list

B. Resource Allocation Procedure, which allocates resources to the jobs contained in scheduling list, generated by task ordering procedure.

a) *Task Ordering Procedure*

Begin

Step 1: The list is initialized to be an empty list. The cloud clients calculate the urgency score according to the severity of jobs.

Step 2: The urgency score is calculated. The urgency score is based on the scale of 10.

There are 4 cases to determine the urgency score of the job.

If Level 1:  $7.6 \leq \text{Score} \leq 10$

If Level II:  $5.1 \leq \text{Score} \leq 7.5$

If Level III:  $2.6 \leq \text{Score} \leq 5.0$

If Level IV:  $1 \geq \text{Score} \leq 2.5$

Step 3. According to the urgency score, the alert templates have set up

*b) Resource Allocation Procedure:*

Begin

Step 1: The Cloud Scheduler collects resources and their characteristics like processing speed, processing cost per second, resource load during peak hours and off peak hours.

Step 2: It generates importance score according to these characteristic on the scale of 10.

Step 3: After collecting the information about the job parameters like urgency and score, the jobs are executed according to the following case.

**If Urgency High, Importance High:** The email alert sent immediately.

**If Urgency High, Importance low:** whenever the resources are free, the email is sent on high priority basis.

**If Urgency Low, Importance High:** The email alert is sent after emptying the job Queue.

**If Urgency low, Importance low:** whenever the resources are free, and the job queue is empty, the email is sent with lower priority basis.

*B. Algorithm 2*

The second algorithm is based upon the Pareto Analysis [23].

**Pareto Analysis**

This is the idea that 80% of tasks can be completed in 20% of the disposable time. The remaining 20% of tasks will take up 80% of the time. This principle is used to sort tasks into two parts. According to this form of Pareto analysis it is recommended that tasks that fall into the first category be assigned a higher priority.

The **80-20-rule** can also be applied to increase productivity: it is assumed that 80% of the productivity can be achieved by doing 20% of the tasks. If productivity is the aim of time management, then these tasks should be prioritized higher.

For example, look at your to do list- if you have 10 tasks on there then two of those tasks will yield 80% of your results. Alternatively, 80% of income is owned by 20% of people - it works both ways!

The Pareto principle holds across business, academia, politics, and a number of other areas. The foundation of this time management skill is that: **20% of tasks yield 80% of results**

This algorithm is also comprised of two parts.

1. Task Ordering Procedure, to get the schedule list
2. Resource Allocation Procedure, which allocates resources to the jobs contained in scheduling list, generated by task ordering procedure.

*a) Task Ordering Procedure*

Begin

Step 1: The list is initialized to be an empty list. The cloud clients send the jobs to the cloud manager according to their priority.

Step 2: The urgency score is calculated. The urgency score is based on the scale of 10.

There are 4 cases to determine the urgency score of the job.

If Level I:  $7.6 \leq \text{Score} \leq 10$

If Level II:  $5.1 \leq \text{Score} \leq 7.5$

If Level III:  $2.6 \leq \text{Score} \leq 5.0$

If Level IV:  $1 \geq \text{Score} \leq 2.5$

Step 3. According to the urgency score, the alert templates have set up

*b) Resource Allocation Procedure:*

Step 1: From the previous set of jobs, importance score is calculated. The score is based upon the following method:

Let  $t = (\text{time to execute jobs} / \text{estimated time}) \%$

If  $t = 100$ , then the cloud resources are utilized properly. A high score of importance is sent by the algorithm.

If  $t < 100$  and  $t > 80$ , the cloud manages resources are overloaded; but according to 80:20 rule by paleto, a high importance score is generated.

If  $t < 80$ , a low importance score is generated.

If  $t > 100$ , the cloud manager resources are underutilization, a high importance score is generated by the cloud.

Step 2: According to the importance score and urgency score, calculated, the jobs are executed by the cloud manager.

**If Urgency High, Importance High:** The email alert sent immediately.

**If Urgency High, Importance low:** whenever the resources are free, the email is sent on high priority basis.

**If Urgency Low, Importance High:** The email alert is sent set after emptying the job Queue.

**If Urgency low, Importance low:** whenever the resources are free, and the job queue is empty, the email is sent with lower priority basis.

IV. IMPLEMENTATION

The management and scheduling of resources in Cloud environment is complex, and therefore demands sophisticated tools for analysis the algorithm before applying them to the real system. But there are no good tools are available that serve our needs. So we develop a service based cloud using Microsoft .Net Technologies. The proposed algorithms are implemented upon this real time cloud.

Feature of This cloud:

- a) To send real time email alerts to the cloud clients members like Bank, Insurance, and Hospital etc.
- b) The algorithms are tested on real time cloud.
- c) Google's SMTP server is used to send the mails.
- d) The Database is saved on the Web server.
- e) The cloud is working online. You need no special software to test the results.
- f) Visual studio 2008 is used as frontend and SQL 2005 is used as Backend.

Cloud Architecture:

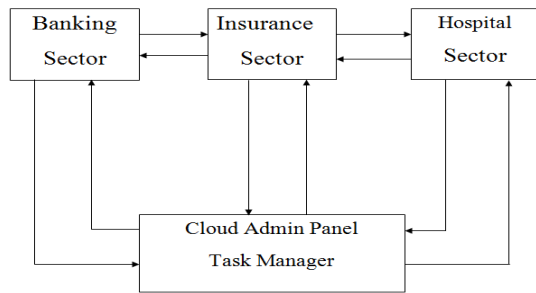


Fig. 2 Architecture of service based cloud

The cloud architecture is based upon the real time email alert system. It sends email alerts to its cluster client members.

Feature of the cloud architecture are:

- a) This scenario based cloud has real life application of sending email alerts to Bank clients, Hospital clients, and Insurance company clouds. This cloud takes jobs from the all other clients with their urgency score. The cloud manager executes the jobs according to the importance score based on cloud resources.
- b) The data Flow between all the clouds is using XML.
- c) XML is hardware and software free technology
- d) It's widely suited for cloud application.
- e) There are 4 domains used in this service based cloud.
- f) Moreover it's three tier architecture, the database is stored on the other server and web services execute on the other server.
- g) FileZilla client FT\P application is used to upload/download data from the server.

### V. EXPERIMENTAL RESULTS

This section describes the experiment results obtained after implementing the scheduling algorithms. The algorithms are implemented in Microsoft .Net framework using a service based cloud. It takes as input the required set of resources and a set of tasks. The algorithms are compares with each other on set of parameters like Total execution Time, Execution time for

Algorithm, Estimated Execution Time.

By Graphical Analysis of Experimental results, we analysed the simulation results using graphs. Graphical data consist of 3 cloud clients' data that are scheduled by main cloud manager. Each algorithm is run by 8 times to conclude the results:

#### Experiment Results of Algorithm1:

TABLE V  
EXPERIMENTAL RESULTS OF ALGORITHM I

Jobs	Total Execution Time	Total Algorithm Time	Estimated Time
1	22419.5038	432.2809	16624
2	17782.3016	575.8664	17204
3	18689.8762	868.9554	17248
4	48342.7388	19863.5854	17385
5	21839.9645	51.0341	19449
6	21720.2652	45.0113	19599
7	28639.296	329.7595	19723
8	22714.3025	123.876	20509

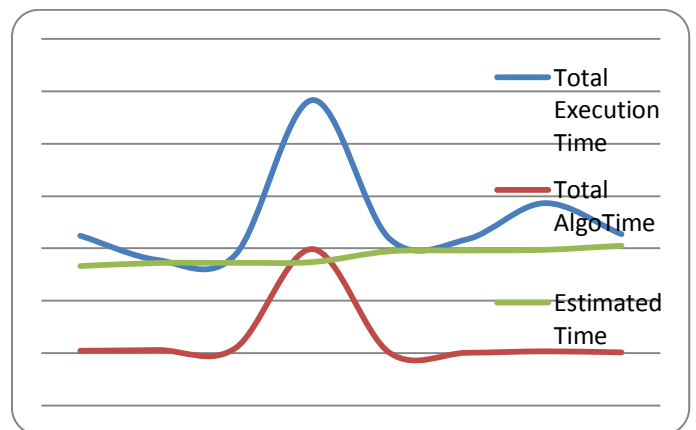


Fig. 3 Line Chart results of Algorithm 1

#### Experimental Results of Algorithm2:

TABLE VIII  
EXPERIMENTAL RESULTS OF ALGORITHM II

Jobs	Total Execution Time	Total Algorithm Time	Estimated Time
1	15048.9774	219.5119	17120
2	14730.2754	255.9244	16861
3	17153.3982	1454.7952	17256
4	17731.0295	319.7149	17359
5	23109.6739	489.1577	20219
6	19565.4316	76.3057	20371
7	22035.7285	286.8222	20331
8	22558.6223	230.3577	20412

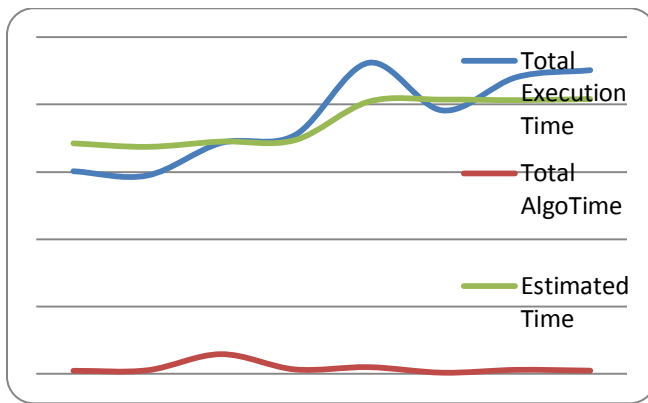


Fig. 4 Line Chart results of Algorithm 2

#### Makespan comparison of Algorithm 1 and Algorithm 2:

The following Graph shows the comparison of makespan time between the two proposed algorithms.

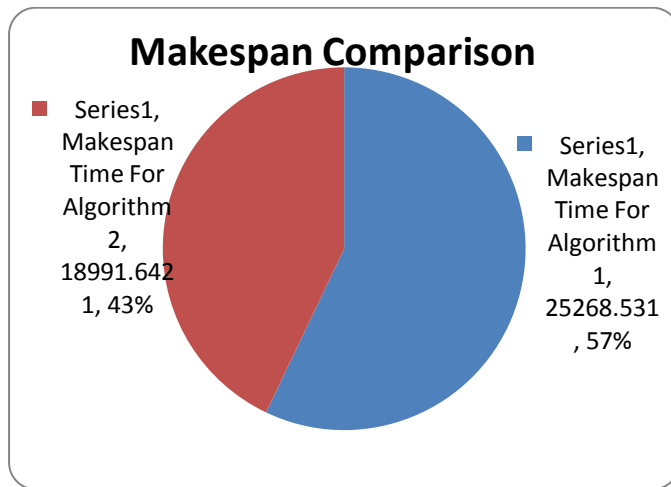


Fig. 5 Pie Chart with Makespan Comparison

#### VI. FUTURE SCOPE

We would like to extend these algorithms to include various parameters like options for advance reservation, preemptive jobs as well. Also, in the Future we can add more clouds to this main cloud, to distribute the load work. Currently this cloud provides services like email alerts, we can also extend to store online data and providing the synchronization mechanism in this.

#### REFERENCES

[1] Workflow Management Coalition, Workflow Management Coalition Terminology & Glossary, February 1999.  
 [2] Ian Foster, Yong Zhao, Ioan Raicu and Shiyong Lu, "Cloud Computing and Grid Computing 360-Degree Compared", Grid Computing Environments Workshop 2008(GCE '08).  
 [3] Sun Microsystems, Inc. "Introduction to Cloud Computing Architecture" Whitepaper, 1st Edition, June 2009. [Online] Available: www.sun.com  
 [4] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility", Future Generation Computer Systems, Elsevier Science, Amsterdam, June 2009, Volume 25, Number 6, pp. 599-616

[5] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities", Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications (HPCC 2008, IEEE CS Press, Los Alamitos, CA, USA), Sept. 25-27, 2008, Dalian, China.  
 [6] Microsoft Azure. <http://www.microsoft.com/azure/> [30 Oct. 2010]  
 [7] Sun network.com (Sun Grid). <http://www.network.com> [18 July 2010]  
 [8] Amazon Elastic Compute Cloud (EC2). <http://www.amazon.com/ec2/> [18 July 2010]  
 [9] Google App Engine. <http://appengine.google.com> [18 July 2010]  
 [10] S. Elnikety, E. Nahum, J. Tracey and W. Zwaenepoel, "A method for transparent admission control and request scheduling in e-commerce web sites," in WWW '04: Proceedings of the 13th International Conference on World Wide Web, 2004, pp. 276-286.  
 [11] Jia Yu, Rajkumar Buyya and Chen Khong Tham, "Cost-based Scheduling of Scientific Workflow Applications on Utility Grids", In 1st IEEE International Conference on e-Science and Grid Computing, Melbourne, Australia, Dec. 5-8, 2005.  
 [12] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. Sci. Program., 13(3):219-237, 2005.  
 [13] Jia Yu and Rajkumar Buyya, "A Budget Constrained Scheduling of Workflow Applications on Utility Grids using Genetic Algorithms", Proceedings of the 15th IEEE International Symposium on High Performance Distributed Computing (HPDC 2006), IEEE CS Press, Los Alamitos, CA, USA, June 19-23, 2006, Paris, France.  
 [14] Patel, Y. Darlington, J., "A novel stochastic algorithm for scheduling QoS-constrained workflows in a web service-oriented grid," in Web Intelligence and International Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on, pp. 437-442.  
 [15] P. Padala, K. G. Shin, X. Zhu, M. Uysal, Z. Wang, S. Singhal, A. Merchant and K. Salem, "Adaptive control of virtualized resources in utility computing environments," SIGOPS Oper. Syst. Rev., vol. 41, pp. 289-302, 2007.  
 [16] J. Yu and R. Buyya, Workflow Scheduling Algorithms for Grid Computing, Technical Report, GRIDS-TR-2007-10, Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia, May 2007.  
 [17] Zhifeng Yu and Weisong Shi, "A Planner-Guided Scheduling Strategy for Multiple Workflow Applications," icppw, pp.1-8, International Conference on Parallel Processing - Workshops, 2008.  
 [18] Ke Liu, Jinjun Chen, Yun Yang and Hai Jin, "A throughput maximization strategy for scheduling transaction-intensive workflows on SwinDeW-G", Concurrency and Computation: Practice and Experience, Wiley, 20(15):1807-1820, Oct.2008.  
 [19] Meng Xu, Lizhen Cui, Haiyang Wang, Yanbing Bi, "A Multiple QoS Constrained Scheduling Strategy of Multiple Workflows for Cloud Computing," ispa, pp.629-634, 2009 IEEE International Symposium on Parallel and Distributed Processing with Applications, 2009.  
 [20] Boris Mejias, Peter Van Roy, "From Mini-clouds to Cloud Computing", 2010  
 [21] J. Kosinska, J. Kosinski, K. Zielinski, "The Concept of Application Clustering in Cloud Computing Environments", 2010  
 [22] M. Jensen, J. Schwenk, J.M. Bohli, L.L. Iacono, "Security prospects through cloud computing by adopting Multiple Clouds", 2011  
 [23] Wikipedia .[http://en.wikipedia.org/wiki/Time\\_management](http://en.wikipedia.org/wiki/Time_management) [25 march 2011].  
 [24] Rao, N. M. (2010). Cloud Computing Through Mobile-Learning. IJACSA - International Journal of Advanced Computer Science and Applications, 1(6).  
 [25] Goel, M. S., Kiran, R., & Garg, D. (2011). Impact of Cloud Computing on ERP implementations in Higher Education. IJACSA - International Journal of Advanced Computer Science and Applications, 2(6), 146-148.

# Study of Indian Banks Websites for Cyber Crime Safety Mechanism

Susheel Chandra Bhatt

Research Scholar, Computer Science department  
Kumaun University, Nainital, Uttarakhand, India

Durgesh Pant

Prof. & Director, School of Computer Science & IT  
Uttarakhand Open University  
Dehradun Campus, Dehradun, Uttarakhand, India

**Abstract**—The human society has undergone tremendous changes from time to time with rapid pace at social level from the beginning and technological level ever since the rise of technologies. This technology word changes the human life in every manner and every sector. Banking field is one of them. Banking in India originated in the last decades of the 18th century. Since that time the banking sector applying different ways to provide facilities and securities to a common man regarding to money. Security issues play extremely important role in the implementation of technologies specially in banking sector. Further on it becomes more critical when it comes to the cyber security which is at the core of banking sector. After the arrival of Internet and WWW this banking sector is totally change specially in terms of security because now money is in your hand on a single click. Now user has number of choices to manage his money with different kind of methods. In this paper an attempt has been made to put forward various issues of Indian banks websites for cyber-crime safety mechanism.

**Keywords**- Cyber; Encryption; Phishing; Secure Socket Layer.

## I. INTRODUCTION

Crime is a social and economic phenomenon and is old as the human society. Crime is a legal concept and has the sanction of the law. Crime or an offence is “a legal wrong that can be followed by criminal proceedings which may result into punishment”[1]. The hallmark of criminality is that, it is breach of the criminal law. Per Lord Atkin “the criminal quality of an act cannot be discovered by reference to any standard but one: is the act prohibited with penal consequences”[2]. Cyber-crime is a term used to broadly describe criminal activity in which computers or computer networks are a tool, a target, or a place of criminal activity and include everything from electronic cracking to denial of service attacks. It is also used to include traditional crimes in which computers or networks are used to enable the illegal activity. Cyber-crime is the latest and perhaps the most complicated problem in the cyber world. Any criminal activity that uses a computer either as an instrumentality, target or a means for perpetuating further crimes comes within the ambit of cyber-crime [3]. Cyber-crimes are computer related as well as computer generated crimes which are increasing day by day. Cyber-crime is a term used to broadly describe criminal activity in which computers or computer networks are a tool, a target, or a place of criminal activity and include everything from electronic cracking to denial of service attacks. It is also

used to include traditional crimes in which computers or networks are used to enable the illicit activity.

Phishing is a way of attempting to acquire sensitive information such as usernames, passwords and credit card details by illegally as a trustworthy entity in an electronic communication. Communications purporting to be from popular social web sites, auction sites, online payment processors or IT administrators are commonly used to lure the unsuspecting public. Phishing is typically carried out by e-mail spoofing or instant messaging [4], and it often directs users to enter details at a fake website whose look and feel are almost identical to the legitimate one. Phishing is an example of social engineering techniques used to deceive users [5] and exploits the poor usability of current web security technologies. Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures. A phishing technique was described in detail in 1987, and the first recorded use of the term “phishing” was made in 1996. The term is a variant of *fishing* [6] probably influenced by *phreaking* [7][8] and alludes to “baits” used in hopes that the potential victim will “bite” by clicking a malicious link or opening a malicious attachment, in which case their financial information and passwords may then be stolen. Not all phishing attacks require a fake website. Messages that claim to be from a bank, tell the users to dial a phone number regarding problems with their bank accounts [9]. Once the phone number (owned by the phisher, and provided by a Voice over IP service) was dialled, it prompts the user to enter the account number and PIN. Vishing (voice phishing) sometimes uses fake caller-ID data to give the appearance that calls is from a trusted organization [10].

## II. CYBER CRIME SAFETY MECHANISM

### A. Password encryption

One of the most important security features used today are passwords. It is important for users to have secure, strong passwords. Most of the more recent Linux distributions include password programs that do not allow user to set an easily guessable password. User has to make sure the password program is up to date and has these features. Encryption is very useful, possibly even necessary in this day and age. There are all sorts of methods of encrypting data, each with its own set of characteristics. Most Unixes (and



Linux is no exception) primarily use a one-way encryption algorithm, called DES (Data Encryption Standard) to encrypt the passwords. This encrypted password is then stored in database. When user attempt to login, the password user type in is encrypted again and compared with the entry in the file that stores the passwords. If they match, it must be the same password, it allowed access. Although DES is a two-way encryption algorithm (user can code and then decode a message, given the right keys), the variant that most Unixes use is one-way. This means that it should not be possible to reverse the encryption to get the password from the contents of database.

### B. Virtual Keyboard

A virtual keyboard is a computer keyboard that a user operates by typing on or within a wireless- or optical-detectable surface or area rather than by depressing physical keys. Such a system can enable the user of a small handheld device, such as a cellular telephone or a PDA (personal digital assistant) to have full keyboard capability. In one technology, the keyboard is projected optically on a flat surface and, as the user touches the image of a key, the optical device detects the stroke and sends it to the computer. In another technology, the keyboard is projected on an area and selected keys are transmitted as wireless signals using the short-range Bluetooth technology. Theoretically, with either approach, the keyboard could even be projected in space and the user could type by moving fingers through the air. The term virtual keyboard is sometimes used to mean a soft keyboard, which appears on a display screen as an image map. In some cases, a software-based keyboard can be customized. Depending on the host system and specific software, the user (who may be someone unable to use a regular keyboard) can use a touch screen or a mouse to select the keys. Virtual keyboard can be categorized as:

- virtual keyboards with touch screen keyboard layouts or sensing areas [11]
- optically projected keyboard layouts or similar arrangements of "keys" or sensing areas[12][13]
- optically detected human hand and finger motions[14]

### C. Secured Socket Layer

Secure Sockets Layer (SSL), are cryptographic protocols that provides communication security over the Internet [15]. SSL is the standard security technology for establishing an encrypted link between a web server and a browser. This link ensures that all data passed between the web server and browsers remain private and integral. SSL is an industry standard and is used by millions of websites in the protection of their online transactions with their customers. To be able to create an SSL connection a web server requires an SSL Certificate. When users choose to activate SSL on web server he will be prompted to complete a number of questions about the identity of the website and the company. The web server then creates two cryptographic keys - a Private Key and a Public Key. The Public Key does not need to be secret and is placed into a Certificate Signing Request (CSR) - a data file also containing the details. User should then submit the CSR. During the SSL Certificate application process, the Certification Authority will validate the details and issue an

SSL Certificate containing the details and allowing user to use SSL. The web server will match issued SSL Certificate to Private Key. The web server will then be able to establish an encrypted link between the website and the customer's web browser. The complexities of the SSL protocol remain invisible to the customers. Instead their browsers provide them with a key indicator to let them know they are currently protected by an SSL encrypted session - the lock icon in the lower right-hand corner, clicking on the lock icon displays the SSL Certificate and the details about it. All SSL Certificates are issued to either companies or legally accountable individuals. Typically an SSL Certificate will contain domain name, company name, address, city, state and country. It will also contain the expiration date of the Certificate and details of the Certification Authority responsible for the issuance of the Certificate. When a browser connects to a secure site it will retrieve the site's SSL Certificate and check that it has not expired, it has been issued by a Certification Authority the browser trusts, and that it is being used by the website for which it has been issued. If it fails on any one of these checks the browser will display a warning to the end user letting them know that the site is not secured by SSL.

### D. SMS Alerts

SMS as used on modern handsets was originated from radio telegraphy in radio memo pagers using standardized phone protocols and later defined as part of the Global System for Mobile Communications (GSM) series of standards in 1985 [16] as a means of sending messages of up to 160 characters[17] to and from GSM mobile handsets[18]. SMS stands for short message service. An SMS alert is a message sent to a cellular device, such as a phone, to notify the receiver of something. An SMS alert is received in much the same way as a phone call is received. There is normally a sound or vibration that will indicate that the message has come in. There are various types of SMS alerts that people may consent to. These include appointment reminders, banking transactions, and specials or sales offered by businesses they patronize. In many instances, an SMS alert is sent out to large numbers of people at once. This means that if two people are scheduled to receive the same SMS alert, they should receive them at about the same time. SMS alerts that contain personal information, such as banking transactions or requests for payment, are not usually handled this way. Sending an SMS alert is often viewed by the sender as a service. In many cases, the senders do not charge the receivers for such messages. There may, however, be a fee charged to both the sender and the receiver by their cellular companies. In other cases, an SMS alert can be part of a subscription. This is a service where a person pays a fee to receive certain types of notifications. These can include news, sports, and weather updates.

### E. User Awareness Programs

User or Customer is the key of any field. We can develop number of software or technology by which we can secure the things but these all software are waist if the end user is not getting the proper information regarding to these software. In Banking sector there are new gadgets and technologies are coming day by day by which the bank can provide secure transactions to end user. Using these gadgets the banks also

has to run some awareness programs for the end users by whom they can also understand the meaning of secure transactions as well as the user will be able to learn how to use these gadgets for secure transactions.

### III. STATUS OF INDIAN BANKS WEBSITES

In this paper we take five Indian banks and try to find out the security features using by the bank for online transactions. The data is collected by various reports from web, newspaper and media. For every security feature we provide 5 points. The banks are-

- State Bank of India (SBI)
- Punjab National Bank [PNB]
- Central Bank of India [CBI]
- Bank of Baroda [BOB]
- Allahabad Bank

TABLE 1. POINT TABLE

Bank	PE*	VK*	SSL*	SMS*	UAP*	Total
SBI	4	4	4	3	0	15
PNB	4	4	3	4	0	15
CBI	3	4	3	2	0	12
BOB	4	4	3	4	0	15
Allahabad	3	4	4	3	0	14

\*PE- Password Encryption, \*VK- Virtual Keyboard, \*SSL- Secure Socket Layer, \*SMS- Short message service alerts, \*UAP- User Awareness Program

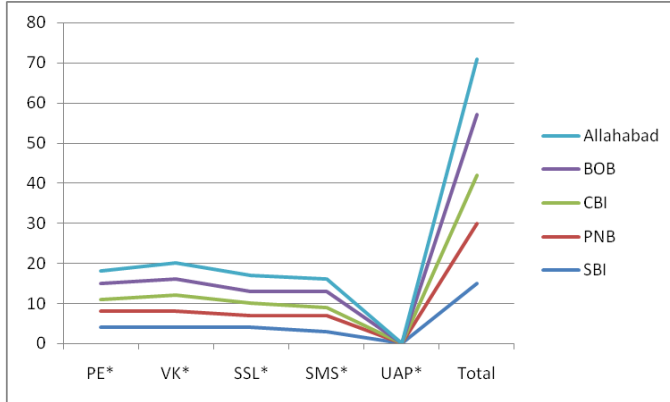


Figure 1. Graphical representation of security features

In this study we found that all banks having the same feature for his websites. They all are providing password encryption facilities with virtual keyboard. The websites of the banks are using secure socket layer and they also providing SMS alerts facilities to customer to know the information regarding to the money transaction. We provide 5 marks for each feature but no bank got full 5 mark for any feature and the aggregate total of every bank is vary 12 to 15 out of 25. The reason behind this is that they all have little bit loop wholes on all the security features but the biggest reason is User awareness feature. Sometime SMS alerts also come with viruses, so this is the responsibility of the bank to check either SMS alert is secure or not. Not all the bank user using the SMS alert facility because the user still in confusing mode

because of the lack of awareness. One another thing is that the most of the facilities started by the bank is still not using by the customer because they have no proper information regarding these facilities and if the customer is using few facilities then this information he came to know from the other customer. So in this area all banks have to focus properly. No bank is providing special training programme to aware the customer, who is the neediest person of world. Because many of the problems user face in bank sectors due to lack of his knowledge. So this is the responsibility of the bank to provide the proper information to end user. They have to tell the user

- What is the meaning of using virtual keyboard
- What is the meaning of strong password
- What is the meaning of SMS alerts
- Don't access net banking account from cyber café or public computer.
- Use a single computer as far as possible.
- Login net banking site by directly typing site name. Don't click any link, if that link takes you to login page, close the page, and start over.
- Bank or its representative never asks for password and username over telephone.
- Change the password after 6 months.
- Remember the id and password, don't write it anywhere.
- Don't give any of the personal information to any web site that does not use encryption or other secure methods to protect it.
- Don't share any information to any one regarding to account
- Install good antivirus programme on the system and regularly update the programme.

The need for user awareness program is continuous, in addition to being multi-disciplinary and multi-dimensional. It is imperative to first digest that information security is a process, not a product. An information security awareness training program thus, needs to maintain the equilibrium between usability, productivity and security. This paper describes the security feature using by the bank for money transaction on websites and focus what banks have to do for spreading awareness. Interesting thing is that it is very difficult to stop any kind of cyber-crime especially when we talk about crimes related to the banks because many of the problems occur in bank sector by the user due to lack of awareness. So if the banks will start to aware the user then definitely the scenario will change. We can't stop the crime, we have to face this. The only one thing which we can do is prevention. We have to learn how to prevent society by this kind of smart crime and banks sector has to play key role from the front to spread awareness.

### IV. CONCLUSION AND FUTURE SCOPE

In this paper we describe different safety features using by the banks for online transaction and examine where the problem is in the system. We found that all the banks use the latest technology for the online security feature but still they

have small loop wholes in this feature. As well as they don't have any user awareness program to spread information and this is one of the biggest reasons of this online security. All bank users do not use online facilities because they have no proper information and the reason behind all this is the same lack of awareness. Along the same line we people also have to increase our awareness level because this is not only the responsibility of the banks. One the interesting thing is that in future these technologies will increase rapidly. It means user will have to use these facilities therefore we need to make our system more secure regarding to the safety mechanism. We have to be aware of the technologies which we are using and also need to increase our awareness level to secure humankind.

#### REFERENCES

- [1] Granville Williams.
- [2] Proprietary Articles Trade Association v. A.G. for Canada (1932)
- [3] Duggal Pawan- Cyber Crime
- [4] Tan, Koontorm Center. "Phishing and Spamming via IM (SPIM)". Retrieved December 5, 2006.
- [5] Microsoft Corporation. "What is social engineering?" Retrieved August 22, 2007.
- [6] "Spam Slayer: Do You Speak Spam?". *PCWorld.com*. Retrieved August 16, 2006
- [7] "Phishing, n. OED Online, March 2006, Oxford University Press.". *Oxford English Dictionary Online*. Retrieved August 9, 2006
- [8] "Phishing". *Language Log, September 22, 2004*. Retrieved August 9, 2006.
- [9] Gonsalves, Antone (April 25, 2006). "Phishers Snare Victims with VoIP". Techweb.
- [10] "Identity thieves take advantage of VoIP". *Silicon.com*. March 21, 2005.

- [11] EP application 546704 Thomas H. Speeter/AT&T: "Intelligent work surfaces" priority date 13.12.1991
- [12] DE application 19734511 B. Kämmerer, C. Maggioni, H. Röttger/SIEMENS AG: "Kommunikationseinrichtung" filing date 08.08.1997
- [13] WO 0003348 C. Maggioni, B. Kämmerer/SIEMENS AG: "Projection Device / Vorrichtung zur Projektion" priority date 10.07.1998
- [14] EP 0554492 Hans E. Korth: "Method and device for optical input of commands or data" filing date 07.02.1992
- [15] T. Dierks, E. Rescorla (August 2008). "The Transport Layer Security (TLS) Protocol, Version 1.2"
- [16] GSM Doc 28/85 "Services and Facilities to be provided in the GSM System" rev2, June 1985
- [17] LA Times: Why text messages are limited to 160 characters
- [18] GSM 03.40 Technical realization of the Short Message Service (SMS)

#### AUTHORS PROFILE



Susheel Chandra Bhatt holds Post Graduate degree and pursuing Ph.D. in Computer Science. Presently, he is working as Research Scientist in Uttarakhand Open University, Campus Dehradun. He has five years of teaching experience. His areas of interest are Cyber security, computer network and database.



Prof. Durgesh Pant holds Post Graduate and Doctoral degree in Computer Science from BIT, Mesra, India. Presently He is working as Director, School of Computer Science & IT, Uttarakhand Open University, Haldwani Campus Dehradun, India. He has published more than 50 National and International papers in peer reviewed journals, and 3 books of his credit. 12 students have been completed their Ph.D. degree under his supervision. He is the member of various National and International academic, social and cultural associations and bodies.

# An Empirical Study of the Applications of Web Mining Techniques in Health Care

Dr. Varun Kumar

Department of Computer Science & Engineering  
ITM University,  
Gurgaon, India

MD. Ezaz Ahmed

Department of Computer Science & Engineering  
ITM University,  
Gurgaon, India

**Abstract-** Few years ago, the information flow in health care field was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. There has already been a long standing tradition for computer-based decision support, dealing with complex problems in medicine such as diagnosing disease, managerial decisions and assisting in the prescription of appropriate treatment. Today, one of the biggest challenges that health care system, face is the explosive growth of data, use this data to improve the quality of managerial decisions. Web mining and Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper addresses the applications of web mining and data mining in health care management system to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples. Further we propose the IDSS model for the health care so that exact and accurate decision can be taken for the removal of a particular disease.

**Keywords-** Web mining; Health care management system; Data mining; Knowledge discovery; Classification; Association rules; Prediction; Outlier analysis, IDSS.

## I. INTRODUCTION

In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Agriculture, Business and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data. Particular disease in a particular area and it will be related with the agriculture of the particular area so that we can predict prone of particular disease in a particular area due to excessive use of pesticides and fertilizer in a particular agricultural area. We have to collect clinical data for a particular disease and with the help of data mining and web mining we can predict a pattern.

Web mining as well as Data Mining (sometimes called data or knowledge discovery) has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information. There are increasing research interests in using web mining and data

mining in health care or web based health care management. This new emerging field, called health care Web mining, concerns with developing methods that discover knowledge from data come from medical environments [1].

The data can be collected from various medical institutes, doctor's clinic or hospital that resides in their databases. The data can be personal or institutional which can be used to understand patients' behavior, to assist doctor, to improve diagnosis, to evaluate and improve health care systems, to improve error free or zero error treatment and many other benefits.[1][2]

Health care data mining used many techniques such as decision trees, neural networks, k-nearest neighbor, naive bayes, support vector machines and many others. Using these methods many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for organization of syllabus, to predict how many patients will register for a particular disease, alienating traditional clinical model predicting a particular disease in a particular area.

This paper is organized as follows: Section 1 introduction and describes the data mining techniques adopted. Section 2 discusses the application areas of these techniques in a medical institute or clinic or any hospital. Section 3 concludes the paper. Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. [5]

### A. Data mining techniques

Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The steps identified in extracting knowledge from data are:

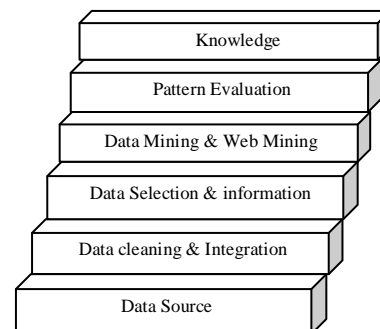


Figure1. The steps of extracting knowledge from data

### B. Association analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

More formally, association rules are of the form  $X \Rightarrow Y$ , i.e., " $A_i \wedge \dots \wedge A_m \rightarrow B_j \wedge \dots \wedge B_n$ ", where  $A_i$  (for  $i$  to  $m$ ) and  $B_j$  ( $j$  to  $n$ ) are attribute-value pairs. The association rule  $X \Rightarrow Y$  is interpreted as database tuples that satisfy the conditions in  $X$  are also likely to satisfy the conditions in  $Y$ ".

### C. Classification and Prediction

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction.

IF-THEN rules are specified as IF condition THEN conclusion

e.g. IF age=old and patient=diabetic then heart disease prone=yes

### D. Clustering Analysis

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity.

That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. [5]Application of clustering in medical can help medical institutes' group individual patient into classes of similar behavior. Partition the patient into clusters, so those patients within a cluster (e.g. healthy) are similar to each other while dissimilar to patient in other clusters (e.g. disease prone or Weak).

### E. Outlier Analysis

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in disease detection and predicting abnormal values or reason of disease.

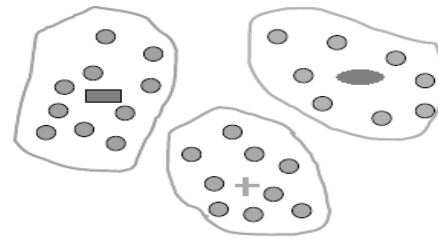


Figure 2 Picture showing the partition of patients in clusters

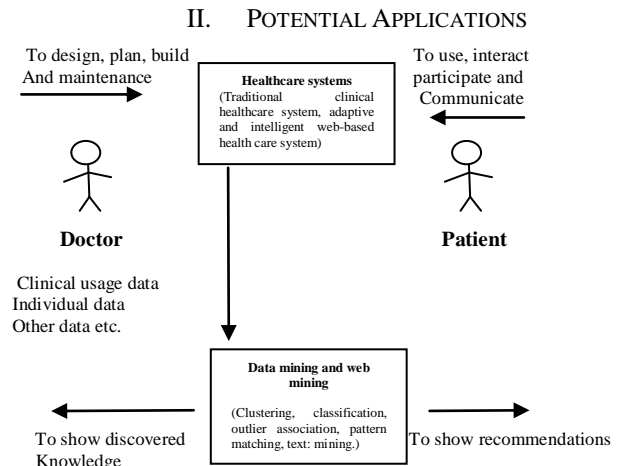


Figure 3 The cycle of applying web mining or data mining in health care system

The figure No. 3 illustrates how the data from the traditional clinic and web based health care systems can be used to extract knowledge by applying web mining and data mining techniques which further helps the doctors and patients to make decisions.

### A. Organization of DATA

It's important for medical institutes to maintain a high quality healthcare program. This will improve the patient's cure process. This will also use for the optimization of resources.

Presently, complication of disease is influenced by many factors such as life style, other disease, environment fertilizer, pesticides, genetic history, availability of best doctor, expert team of doctors and experiences.

One of the applications of data mining is to identify related disease in period of institutional programmes in a large healthcare institute.

A case study has been performed where the patient data collected over a period of time at healthcare Institute. The main of the study was to find the strongly related disease in a period offered by the institute. For this purpose following methodology was followed:

- 1) Identify the possible related disease.
- 2) Determine the strength of their relationships and determine strongly related disease.

In the first step, association rule mining was used to identify possibly related two disease combinations in the period which also reduces our search space. In the second step, Pearson Correlation Coefficient was applied to determine the strength of the relationships of disease combinations identified in the first step. [4]

TABLE 1 SHOWS THE PATIENT SUFFERING FROM THE DISEASE

Patient id	Disease 1	Disease 2	Disease 3
1	Blood Sugar	Blood Pressure	Heart Disease
2	Blood Sugar	Blood Pressure	Heart Disease
3	Blood Sugar	Blood Pressure	Heart Disease
4	Blood Sugar	Blood Pressure	Kidney Problem
5	Blood Sugar	Blood Pressure	Eye Problem

Association Rules that can be derived from Table 1 are of the form:

- $(X, disease1) \Rightarrow (X, disease2)$
- $(X, disease1) \wedge (X, disease2) \Rightarrow (X, disease3)$
- $(X, "Bloodsugar") \Rightarrow (X, "Bloodpressure")$  [support=2% and confidence=60%]
- $(X, "Bloodsugar") \wedge (X, "bloodpressure") \Rightarrow (X, "Heartdisease")$  [support=1% and confidence=50%]

Where support factor of the association rule shows that 1% of the patient suffering from the disease blood sugar and blood pressure, confidence factor shows that there is a chance that 50% of the patients who have "Blood sugar" will also have "Blood pressure"

This way we can find the strongly related disease and can optimize the database of a healthcare programme.

### B. Predicting the Registration of Patients in an Hospital

Now healthcare organizations are getting strong competition from itself, the need of better health by the government plans and by the people concern. It needs deep and enough knowledge for a better assessment, diagnosis, planning, and decision making.

Data Mining helps hospitals to identify the hidden patterns in databases; the extracted patterns are then used to build data mining models, and hence can be used to predict diagnosis and decision making with high accuracy. As a result of this hospitals are able to allocate resources more effectively. [6]

One of the applications of data mining can be for example, a hospital can take necessary actions before patient quit their treatment, or to efficiently assign resources with an accurate estimate of how many male or female will register for a particular disease by using the Prediction techniques.

In real scenario couple of other associated attributes like type of disease, hygiene environment, climate etc. can be used to predict the registration of patients in a particular disease.

### C. Predicting Patients Response Against Medicine

One of the questions, whose answer almost every doctor or

patient of a hospital would like to know "Can we predict patient performance against the medicine?" [6]

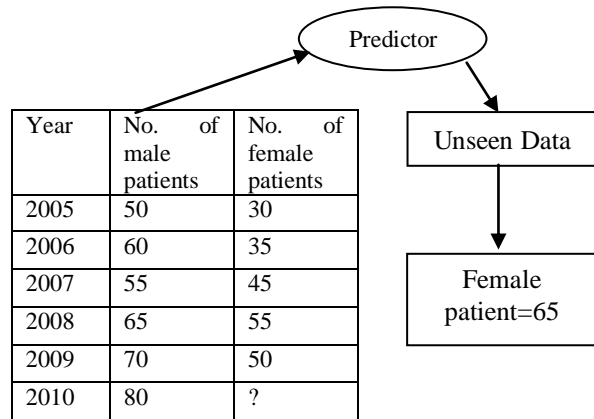


Figure 4 Prediction of female patient in the coming year

Over the years, many researchers applied various data mining techniques to answer this question.

In modern times, healthcare system is taking on a more important role in the development of our civilization by providing good health. Good health maintenance is an individual behavior as well as a social phenomenon.

It is a difficult task to deeply investigate and successfully develop models for evaluating healthcare system efforts with the combination of modernization of equipment's and practice. Healthcare organization's goals and outcomes clearly relate to "promoting good health through effective healthcare models and research in service to healthcare system." With the help of data mining techniques a decision making system can be developed which can help doctors and patients to know the weak points of the traditional clinical healthcare model. Also it will help them to face the rapidly developing real-life environment and adapt the current healthcare realities (such as IDSS etc.)

We use patients' response to medicine participation data as part of the grading policy. A doctor can assess the condition of patient by conducting an online discussion among a group of patients and use the possible indicators such as the time difference between medicine taken and time of response to the medicine etc.[6] With the help of this data, we can apply classification algorithms to classify the patients into possible levels of grading.

### D. Identifying Abnormal/ Erroneous Values

The data stored in a database may reflect outliers-noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. [5]

One of the applications of Outlier Analysis can be to detect the abnormal values in the response sheet of the patient. This may be due many factors like a software fault, data entry operator negligence or an extraordinary response of the patient for a particular disease.



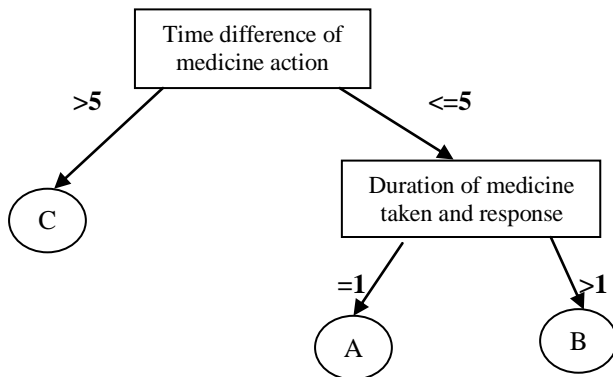


Figure 5 The Decision Tree built from the data in Table 2

TABLE 2 PATIENT RESPONSES DATA AND THEIR GRADES

Time difference between medicine taken action (in min)	Duration between medicine taken and response (in min)	Grade of the Patients
3	1	A
3	2	B
4	1	A
5	2	B
6	1	C
6	2	C

An outlier is an observation that is numerically distant from the rest of the data. Outliers, being the most extreme observations, may include the sample maximum or sample minimum or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

TABLE 3 THE RESPONSE OF PATIENT IN FOUR DISEASES

Patient ID	Level Blood Sugar		ECG Reading	Blood pressure reading	Heart stroke chance
	FF	PP			
101	130	231	35	145/89	30
102	167	270	75	128/90	67
103	189	310	90	178/100	77
104	230	450	35	190/105	99

In the table shown above the response of the patient in disease 4 heart strokes with patient ID 104 will be detected as an exceptional case and can be further analyzed for the cause.

### III. CONCLUSION

The focus of the study was to discuss the various data mining techniques which can support healthcare system via generating strategic information.

Since the application of data mining brings a lot of advantages in higher well equipped hospitals, it is

recommended to apply these techniques in the areas like optimization of resources, prediction of disease of a patient in the hospital, the disease response of the patient, number of them respond, number of the cured and number of them which are fully satisfied.

### REFERENCES

- [1] C. Romero, S. Ventura, E. Garcia, "Datamining in course management systems: Moodle case study and tutorial", *Computers & Education*, Vol. 51, No. 1, pp. 368-384, 2008
- [2] C. Romero, S. Ventura "Educational dataMining: A Survey from 1995 to 2005", *Expert Systems with Applications* (33), pp. 135-146, 2007
- [3] Shaela Ayesha, Tasleem Mustafa, AhsanRazaSattar, M. Inayat Khan, "Data Mining Model for Higher Education System", *European Journal of Scientific Research*, Vol.43, No.1, pp.24-29, 2010
- [4] W. A Sandham, E.D. Lehman "SIMULATING AND PREDICTING BLOOD GLUCOSE LEVEL FOR IMPROVED DIABETES HEALTHCARE", cmru,Imperial College, Royal Brompton London
- [5] Han Jiawei, MichelineKamber, *Data Mining: Concepts and Technique*. Morgan Kaufmann Publishers,2000
- [6] A. Hasman, R. Bindels and P. de Clercq "On the use of reminder Systems in Healthcare", Department of Medical Informatics, University of Maastricht, Netherlands. IEEE International Conference
- [7] Sri Lanka Institute of Information Technology, <http://www.sliit.lk/>
- [8] Sun Hongjie, "Research on Student Learning Result System based on Data Mining", *IJCSNS International Journal of Computer Science and Network Security*, Vol.10, No. 4, April 2010
- [9] Academy Connection – Training Resources Inhtml, <http://www.cisco.com/web/learning/netacad/index>, December 28th, 2005.
- [10] Wayne Smith, "Applying Data Mining to Scheduling Courses at a University", *Communications of the Association for Information Systems*, Vol. 16, Article 23.
- [11] Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(3), 80-84.
- [12] Thakur, M. (2011). Query based Personalization in Semantic Web Mining. *IJACSA - International Journal of Advanced Computer Science and Applications*, 2(2), 117-123.

### AUTHOR'S PROFILE



#### Prof. Varun Kumar

Ph.D. (Computer Science), Associate Head , CSE Deptt. School of Engineering and Technology, ITM University, Gurgaon Haryana ,India. Presently 3 Ph. D students are working under his supervision. Dr. Varun Kumar, completed his PhD in Computer Science. He received his M. Phil. in Computer Science and M. Tech. in Information Technology. He has 13 years of teaching experience. He is recipient of Gold Medal at his Master's degree. His area of interest includes Data Warehousing, Data Mining, and Object Oriented Languages like C++, JAVA, C# etc. He has published more than 35 research papers in Journals/Conferences/Seminars at international/national levels. He is working as an Editorial Board Member / Reviewer of various International Journals and Conferences. He has 3 books, 5 study materials and 3 lab manuals to his credit.



#### Md. Ezaz Ahmed

Pursuing Ph.D. under the supervision of Prof, Varun kumar. Currently, he is working with itm University as Asst. Professor. He did his M.E (CSE) in first division with honors. Before joining this Institute, He has more than 17 years of experience out of which 15 years teaching and 2 years industry experience. He has published 11 research papers, 1 in international Journal others in national conference and in departmental journal. His area of interest includes Web Development, software Engineering, Software verification validation and testing, Soft Computing, and Basics of computer and C programming. He is a member of Indian Society of Technical Education (ISTE). Co-author of one project book published in 1998.

# Quality EContent Design using Reusability approach

Senthil Kumar.J,

(Research Scholar, Vels University, India.) Asst.  
Professor, Dhanraj Baid Jain College, India

Dr S.K.Srivatsa, Ph.D.,

Professor, St.Josephs College of Engg. , India

**Abstract**— Technology is the one changing ever, and major technological innovations can make a paradigm shifts. The computer network known as the Internet is one such innovation. After affecting sweeping changes in the way people communicate and do business, the Internet is perched to bring about a paradigm shift in the way people learn. Consequently, a major change may also be coming in the way educational materials are designed, developed, and delivered to the learner. An instructional technology called “learning objects” currently leads other candidates for the position of technology of choice in the next generation of instructional design, development, and delivery. This paper aims to address the reusability, generating capability, adaptability, and scalability of the content designed using the learning objects. Object-orientation highly values the creation of components (called “objects”) that can be reused in multiple contexts. This is the fundamental idea behind learning objects.

**Keywords-** Technology; paradigm; network; learning objects.

## I. INTRODUCTION

The instructional technology communities have begun to grapple with mapping sound instructional principles to the technical attributes of learning object systems for education and training purposes [1]

Learning object systems are flexible, dynamic and highly engaging technology-based environments. These systems have great potential to capitalize on the goal-oriented nature of human learning processes as well as allowing learners to associate instructional content with their prior knowledge and individual experiences

The transition of the education and training communities to paperless, digital work and learning environments has important implications. They often involve the re-production of media and approaches that have been developed. Previously, tending to increase dramatically both the cost and the time required to develop training and education products and services. Simply re-hosting existing education and training approaches using digital media may optimize neither human nor technology’s capabilities.

## II. INSTRUCTIONAL DESIGN BASED ON OBJECT ORIENTED APPROACH-PEDAGOGICAL SHIFT

Before the industrial revolution, a craft based approach to product manufacture was prevalent, where one or two individuals create a completed product from the raw material available to them. After the industrial revaluation there were many changes in product manufacturing. The major developments were the division of labor, increased automation

and to development of the component based approach to manufacturing. The main benefit of a component based approach is reusability is a component used on product can be used to provide the same function for another product. Parallel to the industrial revolution has occurred within a shorter time frame in the software industry. It is only since the development of the idea of software engineering in the 1970’s that software development has begun to move from a craft to an industry.

The Idea of moving to a component model for development of courses and content has gained prominence move recently and been driven by the interest in the educational potential of the internet. There are number of initiatives which transfer the ideas and benefits of the component approach to developments and delivery of educational systems [2].

A learning component more commonly referred as a learning object is any discrete unit of learning material that can be extracted from one course and integrated into another.

Reusable learning object is an emerging paradigm shift in instructional system that promises to bring to education the same improvements in productivity that it has in software development. There are number of problems to be resolved before component manufactures becomes an established approach in educational system design. These include the issues of standards for learning objects and support for those educators making the transition to object based design.

## III. OBJECT ORIENTED APPROACH FOR CREATING REUSABLE AND TRANSPORTABLE LEARNING CONTENT

- a) *Transportable among applications and environments*
- b) *Re- purposable to different delivery structures.*

To be reusable and transportable an object needs to meet some technical coding standards and it must be instructionally designed for reuse. In addition each learning object must be labeled to make identification of content, topic, purpose, etc. readily apparent and to make the object easily retrievable. There are “two requisite components of a learning object: the object content and its metadata tag.” [3] Meta-tagging means linking or tagging objects and assets with specific metadata. “Metadata, literally ‘data about data’, is descriptive information about a resource...metadata allow you to locate an item very quickly without investigating all the individual items through which you are searching.”[4] Because they are stored in a database structure and managed through a Learning Content Management System via meta-tagging, learning objects make it easy to find and access content anywhere and anytime and they are easy to update and display.

#### IV. FEATURES OF LEARNING OBJECTS

The following is a list of some of the types of information that may be included in a learning object and its metadata:

- General Course Descriptive Data, including: course identifiers, language of content (English, Spanish, etc.), subject area (Maths, Reading, etc.), descriptive text, descriptive keywords
- Life Cycle, including: version, status
- Instructional Content, including: text, web pages, images, sound, video
- Glossary of Terms, including: terms, definition, acronyms
- Quizzes and Assessments, including: questions, answers
- Relationships to Other Courses, including prerequisite courses
- Educational Level, including: grade level, age range, typical learning time, and difficulty.

#### V. METADATA STANDARDS FOR LEARNING OBJECTS

Learning objects are indeed a good idea, but as long as they lack instructional value, we will be unable to use them effectively. From a practical and technical perspective, common metadata standards define what data needs to be collected and stored to provide descriptive information about a content object. The result is a content object metadata specification (e.g., showing title, author, and description for each object). Metadata standards theoretically should also enable the appropriate use of a content object as a learning object. In this case, the purpose is to enable learners to use one or more learning objects to achieve one or more instructional objectives.

The metadata on a library catalog card provides information commonly used for finding a book or other media form, but has little instructional information concerning the reader's instructional use of the item. If our sole purpose is to provide metadata for describing content objects, the descriptive information commonly included by most standards today is sufficient. However, learning objects have important embedded instructional objectives and, if we are not providing instructional information in metadata, all we have is a content object. If we ignore key instructional issues, how can we successfully use learning objects for learning?

Many groups are working together to define common international standards that the world can adopt for describing learning objects that can be interoperable, reusable, repurposable, and effectively managed and presented. Their common interest is to find a minimum set of metadata standards that will support the worldwide deployment of learning objects for multiple purposes. Just a few of the groups participating in these worldwide standards-making efforts through the IEEE Learning Technology Standards Committee [5] are:

- Alliance of Remote Instructional Authoring and Distribution Networks for Europe (ARIADNE, 2000)
- Instructional Management Systems (IMS, 2000a) Project
- Dublin Core Education Working Group (DC-Ed, 2000)
- Advanced Distributed Learning Initiative (ADL, 2000)

#### VI. MUTABILITY OF LEARNING OBJECT

A mutated learning object is, according to Michael Shaw, a learning object that has been "re-purposed and/or re-engineered, changed or simply re-used in some way different from its original intended design". Shaw also introduces the term "contextual learning object", to describe a learning object that has been "designed to have specific meaning and purpose to an intended learner". [6]

#### VII. PORTABILITY OF LEARNING OBJECT

Before any institution invests a great deal of time and energy into building high-quality e-learning content (which can cost over \$10,000 per classroom hour), it needs to consider how this content can be easily loaded into a Learning Management System. It is possible for example, to package learning objects with SCORM specification and load it at Moodle Learning Management System. If all of the properties of a course can be precisely defined in a common format, the content can be serialized into a standard format such as XML and loaded into other systems. When you consider that some e-learning course need to include video mathematical equations using MathML, chemistry equations using CML and other complex structures the issues become very complex, especially if the systems needs to understand and validate each structure and then place it correctly in a database.

#### VIII. LEARNING OBJECT PROJECTS

Some examples of learning object projects include:

- AGORA, a publicly accessible online learning environment at the Virtual Museum of Canada. Content is created and produced by Canadian museum educators.
- eduSource, a Canada-wide project to create the infrastructure for a network of inter-operable learning object repositories. The eduSource project is based on national and international standards; it is bilingual (French and English) and designed to be fully accessible.
- MERLOT (Multimedia Educational Resource for Learning and Online Teaching), a free and open resource designed primarily for faculty and students of higher education.
- IQity Reactor is a learning object repository that allows educators to create and share custom curriculum, organized by state educational standards. Reactor is integrated with a learning management system.
- Wisc-Online is a web-based repository of learning objects, developed primarily by faculty members from the Wisconsin Technical College System.

#### IX. CONCLUSION

Object-orientation highly values the creation of components (called "objects") that can be reused in multiple contexts. This is the fundamental idea behind learning objects. Instructional designers can build small instructional components that can be reused a number of times in different learning contexts. Additionally, learning objects are generally understood to be digital entities deliverable over the Internet, meaning that any number of people can access and use them simultaneously.

Moreover, those who incorporate learning objects can collaborate on and benefit immediately from new versions. These are significant differences between learning objects and other instructional media that have existed previously.

#### REFERENCES

- [1] Merrill, M. D. (1999a). Instructional transaction theory (ITT): Instructional design based on knowledge objects. In C. M. Reigeluth (Ed.), *Instructional-Design Theories and Models*:
- [2] Roschelle, J., Kaput, J. Stroup, W. and Kahn, T.M., "Scalable integration of educational software: exploring the promise of component architectures: *Journal of Interactive media in education*, volume 6, 1998, [www-jime.open.ac.uk/98/6](http://www-jime.open.ac.uk/98/6)
- [3] Longmire, Warren. (2000). *A Primer on Learning Objects*.
- [4] Wiley David A (2000). *Connecting learning objects to Instructional Design Theory*
- [5] Learning Technology Standards Committee (2002) (PDF), *Draft Standard for Learning Object Metadata. IEEE Standard 1484.12.1*, New York: Institute of Electrical and Electronics Engineers, retrieved 2008-0429
- [6] [http://www.shawmultimedia.com/edtech\\_oct\\_03.html](http://www.shawmultimedia.com/edtech_oct_03.html).

# Retrieval of Images Using DCT and DCT Wavelet Over Image Blocks

H. B. kekre

Professor Department of Computer Engineering  
MPSTME,  
NMIMS University,  
Vileparle (W), Mumbai, India

Kavita Sonawane

Ph. D Research Scholar Department of Computer  
Engineering MPSTME,  
NMIMS University,  
Vileparle (W), Mumbai, India

**Abstract**— This paper introduces a new CBIR system based on two different approaches in order to achieve the retrieval efficiency and accuracy. Color and texture information is extracted and used in this work to form the feature vector. To do the texture feature extraction this system uses DCT and DCT Wavelet transform to generate the feature vectors of the query and database images. Color information extraction process includes separation of image into R, G and B planes. Further each plane is divided into 4 blocks and for each block row mean vectors are calculated. DCT and DCT wavelet is applied over row mean vector of each block separately and 4 sets of DCT and DCT wavelet coefficients are obtained respectively. Out of these few coefficients are selected from each block and arranged in consecutive order to form the feature vector of the image. Variable size feature vectors are formed by changing the no of coefficients selected from each row vector. Total 18 different sets are obtained by changing the no of coefficients selected from each block. These two different feature databases obtained using DCT and DCT wavelet are then tested using 100 query images from 10 different categories. Euclidean distance is used as similarity measure to compare the image features. Euclidean distance calculated is sorted into ascending order and cluster of first 100 images is selected to count the images which are relevant to the query image. Results are further refined using second level thresholding which uses three criteria which can be applied to first level results. Results obtained are showing the better performance by DCT wavelet as compare to DCT transform.

**Keywords**-component; DCT; DCT wavelet; Euclidean distance.

## I. INTRODUCTION

Large amount of images are being generated, stored and used daily in various real life applications through various fields like engineering, medical sciences, biometrics, architectural designs and drawings and many other areas. Although various techniques are being designed and used to store the images efficiently, still it demands to search new effective and accurate techniques to retrieve these images easily from large volume of databases. Text based image retrieval techniques have tried in this direction which has got many constraints and drawbacks associated with it which is continuously encouraging the researchers to come up with the new techniques to retrieve the images based on contents instead of text annotations. Image contents are broadly classified into global and local contents. Local contents define the local attributes of the image like color, shape and texture

information. These attributes can be used and processed to represent the image feature to make them comparable for similarity. Many techniques are being developed in this field to retrieve the images from large volume of database more precisely [1], [2], [3], [11], [12], [13] [32] [33]. This paper contributes in same direction by introducing the novel techniques which are giving favorable performance which is analyzed through different aspects of the behavior of the proposed CBIR system.

In this work many variations are introduced which are not used in the previous work in the same direction. We are focusing on color and texture information of image. First we are separating the image into R, G, B planes and then decomposing the image plane into 4 blocks and applying DCT transform over row mean vectors of each block of it to obtain the texture information of the image. The logic behind that DCT is a good approximation of principal component extraction, which helps to process and highlight the signal frequency features [21], [24], [26], [27], [29], [31]. Same process is repeated with DCT wavelet transform over row mean vectors of each block of each plane. As Wavelets can be combined, using a "shift, multiply and sum" technique called convolution, with portions of an unknown signal to extract information from the unknown signal. They have advantages over traditional fourier methods in analyzing physical situations where the signal contains discontinuities and sharp spikes [10], [22], [23], [28]. This paper is organized as follows. Section II will introduce transforms applied to form the feature vectors. Section III gives the algorithmic flow of the system that explains how to extract the image contents and formation of the feature vector databases [4], [5], [6], [16]. Section IV explains the experimental results with performance analysis of the system and Section V delineate the conclusion of the work done.

## II. DISCRETE COSINE TRANSFORM – AND DCT WAVELET

Discrete cosine transform is made up of cosine functions taken over half the interval and dividing this interval into N equal parts and sampling each function at the center of these parts [8], the DCT matrix is formed by arranging these sequences row wise. This paper uses DCT transform to generate the feature vectors which is explained in section III.

Wavelets are mathematical functions that cut up the data or signal into different frequency components by providing a way to do a time frequency analysis. Analysis of the signals containing the discontinuities and sharp spikes is possible with help of wavelet transforms [7], [10], [17]. Kekre's generalized algorithm which generates the wavelet from any orthogonal transform is used to generate DCT wavelet as DCT is an orthogonal transform [10], [15]. To take advantage of this property of wavelet, this paper has proposed a new algorithm to represent the feature vectors in the form of discrete cosine wavelet transform coefficients for the CBIR.

The DCT definition of 2D sequence of Length N is given in equation (1) using which the DCT matrix is generated [15] [24]. The generalized algorithm which can generate wavelet transform of size  $N^2 \times N^2$  from any orthogonal transform of size  $N \times N$  is applied to DCT matrix and DCT Wavelet is developed which satisfies the condition of orthogonal transforms given in equation (2). Once the Discrete Cosine Transform Wavelet is generated following steps are followed to form the feature vectors of the images.

$$F[k, l] = \sum_{m=0}^{M-1} \sum_{n=0}^{nN-1} f[m, n] \alpha(k) \alpha(l) \cos \left[ \frac{(2m+1)k\pi}{2M} \right] \cos \left[ \frac{(2n+1)l\pi}{2N} \right] \quad (1)$$

Where  $\alpha(k) = \begin{cases} \sqrt{\frac{1}{N}}, & \text{for } k = 0 \\ \sqrt{\frac{2}{N}}, & \text{for } k = 1, 2, \dots, N-1 \end{cases} \quad (2)$

**Orthogonal:** DCT Wavelet transform is said to be orthogonal if the following condition is satisfied.

$$[DCTW][DCTW]^T = [D] \quad (3)$$

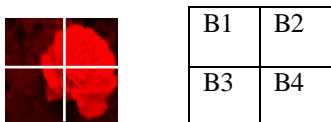
### III. ALGORITHMIC VIEW OF CBIR USING DCT AND DCT WAVELET

In following algorithm step1 to step3 is same for both the approaches of CBIR

**Step1.** Separate the image into R, G and B planes.



**Step2.** Divide each plane of image into four blocks B1, B2, B3 and B4 of all equal sizes. [35]



**Step3.** For each block calculate the row mean vectors.

122	168	.....	145	→	(122 + 168 + ... 145) / n
188				→	.
..				→	.
199	220	.....	160	→	(199 + 220 + ... 160) / n

**Step4.** In First approach we Apply Discrete Cosine Transform over all row mean vectors of each block of each plane of the all the database images and DCT feature database is prepared [35]. Similarly, for second approach we applied DCT wavelet over all row mean vectors of all four blocks of each plane of all database images and new DCT Wavelet feature database is prepared for the second approach.

**Step5.** Representation of feature vectors for both the approaches is explained as follows:

Select few DCT and DCT wavelet coefficients from each row vector of all four blocks of each plane and arrange them in single vector in consecutive order. It gives the feature vector of that particular plane. Similar procedure is followed to get the feature vector for all three planes R, G, B.

This feature vector consist of four components for each plane for example red plane these components are named as RB1, RB2, RB3 and RB4 where suppose each component has 64 coefficients. Arrangement of these four components in single row vector gives the final feature vector for red plane of size  $64 \times 4 = 256$  coefficients.

This CBIR system is experimented with various different size feature vectors for both the approaches. Details of how the coefficients are selected are given in following manner.

DCT and DCT Wavelet Feature vectors in Variable Size	
No .of Coefficients Selected From Each Block	1, 2, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64
Total coefficients in the Final Feature Vector in Feature Database DCT and DCT wavelet	4, 8, 16, 32, 48, 64, 80, 96, 112, 128, 144, 160, 176, 192, 208, 224, 240, 256

Feature vectors for red, green and blue plane are obtained using above procedure and two feature vector databases are created for all the database images using DCT and DCT wavelet.

**Step6.** Once the feature databases are prepared system is tested with query image. Feature extraction of query image will be done in same manner as it does for the database images.

Similarity measure Euclidean distance given in equation (5) is applied to compare the query image with the database images for similarity [4], [5], [6], [19] [37].



$$D_{QI} = \sqrt{\sum_{i=1}^n (FQ_i - FI_i)^2} \quad (5)$$

**Step7.** Retrieval results are based on the criterion of sorting the Euclidean distances in ascending order and selecting first 100 images with respect to first 100 minimum distances from 1000 distances sorted in ascending order for all database images.

### III. EXPERIMENTAL RESULTS AND DISCUSSIONS

#### A. Database and Query Image

Algorithms discussed above in section III is experimented with database of 1000 images which includes 100 images from each of the following categories; that are Flower, Sunset, Mountain, Building, Bus, Dinosaur, Elephant, Barbie, Mickey and Horse images. Feature vectors for all these 1000 images are extracted using above procedure based on DCT and DCT wavelet transforms. This CBIR system is tested with query by example image. Whenever system receives the query image it extracts the feature vector for it in the same way as it extracts for database images. By means of similarity measure Euclidean distance, it will compare the query with database images for the exact match. Ten queries from each of the 10 classes are given as query to the proposed algorithms and Euclidean distance is calculated for all of them. Sample Images from all classes are shown in Figure.1

#### B. Retrieval of Similar Images from Database of 1000 Images

Once the Query is entered it is processed as explained above to extract its contents to form the feature vector. As given in step 1 in section III that each image is separated into R, G, and B planes, we are having 3 sets of feature databases for each approach that is features for R plane, G plane and Blue Plane. Query image along with this 3 features R, G and B plane features will be compared with R, G and B plane features of all database images respectively. This gives us the 3 sets of retrieval results with respect to each plane [9], [14], [15], [20]. During the experiments of this system some variation are made in the selection of coefficients to form the feature vector. When we work in transform domain to utilize and analyze the energy compaction property of them we have selected the starting few coefficients which are carrying most of the information of the image to represent the feature vector. Here we have tried different size feature vectors by changing the no of coefficients [36]. First we took all coefficients and then we went on reducing their count to reduce the size of the feature vector. Total 18 different sets we tried with the range of feature vector size from 256 to 4 coefficients for each plane and each approach.

One more variation we made in the coefficients is while selecting the first coefficient we have scaled down it to the range of its succeeding coefficients in that list. Because the first coefficient is high energy coefficient as compare to all successive coefficients. Two different scale down factors 10 and 5 are selected to just scale down the first coefficient of each sequence. Based on these two factors, two sets of feature databases are obtained per plane. Total 3x2x18 feature vectors are obtained, 3 planes 2 Scale down Factors and 18 different sizes. In turn 108 x 2 executions are made for both the

approaches for 1 query image and 216 results obtained for that query. Like this 100 queries are tried for both the approaches based on DCT and DCT wavelet. Table I shows the average values of 100 query images from 10 different classes. Each value in table is representing the average out of 10,000.

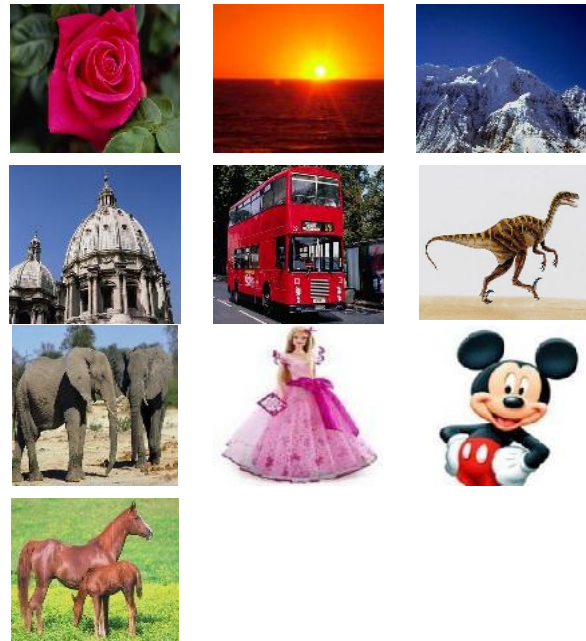


Figure 1. Sample Images from 10 different classes

Further, to reduce these results obtained in Table-I we have combined the results obtained separately for each plane using following criteria.

**Criterion 1:** Image will take into final retrieval only if it is being retrieved in result set of all 3 planes R, G and B.

**Criterion 2:** Image will be retrieved into final retrieval only if it is being retrieved in at least any 2 of the three planes R, G and B.

**Criterion 3:** Image will be retrieved into final retrieval if it is being retrieved in at least one of the three planes R, G and B.

All Criteria are repeated with 2 factors (10 and 5) for 100 query images. And total 3x2x100 results are obtained for each of the two approaches based on DCT and DCT wavelet. Each value in Table II is representing the average number of similar images retrieved out of 10,000.

TABLE I. AVERAGE VALUES OF 100 QUERIES FOR EACH OF THE 18SETS OF VARIABLE COEFFICIENTS.

Retrieval Results for DCT and DCWT for Two Scale Down Factors				
No. of Coefficients	Scale Down Factor 10		Scale Down Factor 5	
	DCT	DCWT	DCT	DCWT
4	8726	8728	8740	8744
8	7270	7359	8713	8612

Retrieval Results for DCT and DCWT for Two Scale Down Factors				
No. of Coefficients	Scale Down Factor 10		Scale Down Factor 5	
	DCT	DCWT	DCT	DCWT
16	8123	8115	9350	9331
32	7938	7987	9139	9150
48	7888	7958	9065	9113
64	7686	7823	8754	8964
80	7851	7800	9006	8936
96	7831	7846	8988	9002
112	7828	7852	8993	8989
128	7828	7837	8985	8992
144	7829	7834	8983	8988
160	7685	7826	8983	8813
176	7923	7917	8979	8975
192	7644	7811	8745	8979
208	7813	7820	8973	8982
224	7810	7814	8977	8981
240	7824	7758	8976	8912
256	7812	7815	8976	8912

**Observation.** Scale down factor 5 gives far better performance as compare to factor 10 and DCWT results are better than DCT.

### C. Results and analysis of CBIR using DCT and DCT wavelet.

Proposed algorithm is experimented with 100 queries, 10 images from each category and results are obtained by applying the similarity measure Euclidean distance. Retrieval results of all 18 sets of feature vector of different sizes are obtained for each plane separately along with two scale down factors 10 and 5. Table I is showing the average values for execution of 100 queries for each feature vector set from 18 variations. When we observed these results obtained using the scale down factor 5 are giving best performance in all the sets and for all the planes R, G, and B. It can also be noticed in the chart 1 and 2 that performance of factor 5 is having good accuracy as compare to factor 10.

When we observed the these results obtained for each one 18 sets of different size feature vectors for all 3 planes it has been noticed that for first few coefficients sets selection system is performing well. It has been observed that when feature vector size was 16 coefficients for factor 5 and 4 coefficients for factor 10, system has given its best

performance as shown in Chart 1 and 2. One more observation made that red plane is proving its best in terms of the average values of retrieval set. To refine the results obtained for 3 planes we have applied the above mentioned 3 criteria and the results obtained are shown in Table II for the both the approaches with reference to both factors 5 and 10. In these results we can notice that criterion 3 is giving best performance among all three sets of results where the image similar to query will be retrieved in final set if it is being retrieved in at least one of three planes. Chart 3, 4, and Chart 5, 6 are displaying the results for all criteria for DCT and DCWT for factor 10 and 5 respectively, where we can notice the behavioral difference of the system for these 3 criteria as mentioned above.

### D. Performance Evaluation of CBIR using DCT and DCT wavelet.

Results obtained in this work using DCT and DCT Wavelet, is indirectly compared with the traditional parameters Precision and Recall. Here when system generates the retrieval result in terms of 1000 Euclidean distances between the given query image and 1000 database images which are sorted in ascending order; out of which first 100 images are selected as retrieval set of similar images which carries images belong to same category of query and even other category images as well [18]. When we talk in terms of precision, it is in the range of 30% to 70% for most of the query images. At the same time very good results are obtained for most of the query images for both the approaches in terms of recall parameter which is in the range of 40% to 90 % for many query images.

We compare these results with the other work done using DCT or other wavelets [15], [29], [30], [35]. It can be observed and noticed that the database we are using includes images from different classes and each class has got 100 images of its own category which has got images with different background also which has impact on the feature extraction and even on the retrieval process. It still performs better in terms of precision and recall. We have tried 100 query images and the cumulative result which is average of 100 queries is summarized in the above tables. If we consider the result of each query separately in most of the queries can say for 50 % of the queries we have got very good values for precision which is around 0.7 to 0.8. and at same time for the same query are getting good results in terms of recall which is around 0.6 to 0.7 which can be considered as best results for CBIR system. But at the same time if we have to consider the overall performance of these approaches they should perform well or in same manner for all 100 or more queries which again triggering us to make future improvements. This is explained in brief in the last section after conclusion.

TABLE II. AVERAGE VALUES OF 100 QUERIES FOR EACH OF THE 18SETS OF VARIABLE COEFFICIENTS FOR ALL 3 CRITERIA FOR FACTOR 10

Size of Feature vector	DCT with Scale down factor 10			DCWT with Scale down factor 10		
	Criterion1	Criterion2	Criterion3	Criterion1	Criterion2	Criterion3
4	1293	2653	4900	1294	2653	4901
8	941	2199	4284	833	2076	4410
16	1237	2522	4412	1228	2535	4419
32	1216	2477	4298	1238	2493	4318
48	1233	2456	4233	1248	2473	4285
64	1152	2395	4203	1154	2367	4346
80	1245	2468	4200	1156	2372	4329
96	1244	2458	4193	1244	2464	4197
112	1242	2455	4187	1245	2464	4199
128	1244	2461	4189	1245	2564	4188
144	1244	2457	4191	1247	2461	4185
160	1248	2452	4181	1245	2388	4206
176	1248	2452	4174	1245	2455	4178
192	1090	2344	4264	1251	2453	4169
208	1248	2451	4167	1250	2454	4171
224	1249	2450	4169	1253	2454	4171
240	1246	2455	4170	1163	2360	4290
256	1249	2454	4160	1247	2453	4162

TABLE III. AVERAGE VALUES OF 100 QUERIES FOR EACH OF THE 18SETS OF VARIABLE COEFFICIENTS FOR ALL 3 CRITERIA FOR FACTOR 5

Size of Feature vector	DCT with Scale down factor 5			DCWT with Scale down factor 5		
	Criterion1	Criterion2	Criterion3	Criterion1	Criterion2	Criterion3
4	1285	2648	4902	1286	2646	4898
8	1272	2692	4817	1138	2586	4957
16	1569	2946	4889	1574	2951	4888
32	1527	2861	4768	1555	2901	4820
48	1546	2852	4716	1548	2871	4762
64	1430	2754	4636	1438	2760	4834
80	1547	2760	4687	1441	2838	4817
96	1551	2839	4678	1550	2839	4667
112	1555	2836	4660	1553	2837	4669
128	1553	2837	4659	1556	2836	4662
144	1554	2839	4653	1553	2844	4658
160	1553	2839	4652	1558	2775	4683
176	1553	2839	4648	1558	2836	4650
192	1364	2716	4741	1558	2834	4653
208	1556	2837	4649	1557	2834	4650
224	1554	2836	4653	1555	2836	4655
240	1561	2838	4652	1449	2758	4766
256	1560	2832	4652	1559	2833	4644

*Observation:* DCWT with scale down factor 5 gives better performance under all three criteria. No of images retrieved increases from criterion 1 to criterion 3

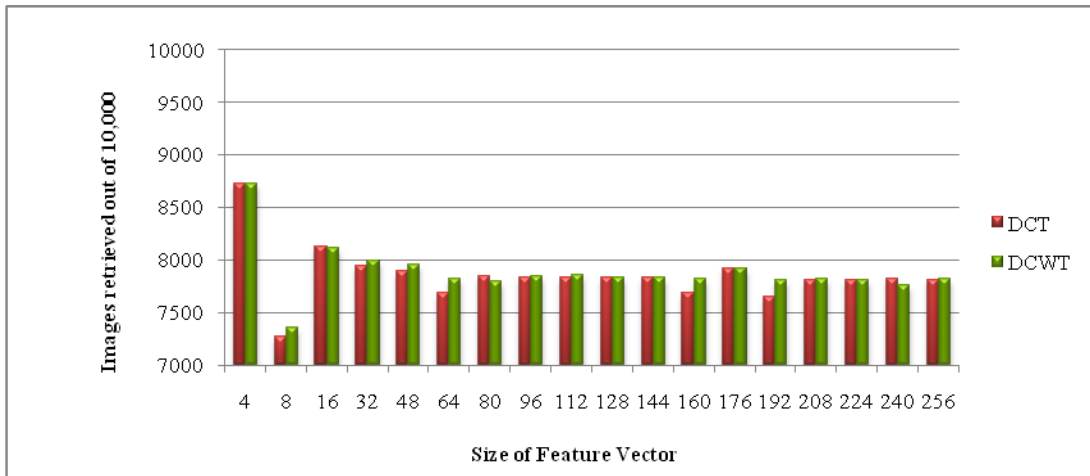


Figure 2. Plot for DCT and DCWT Using Scale Down Factor 10 for

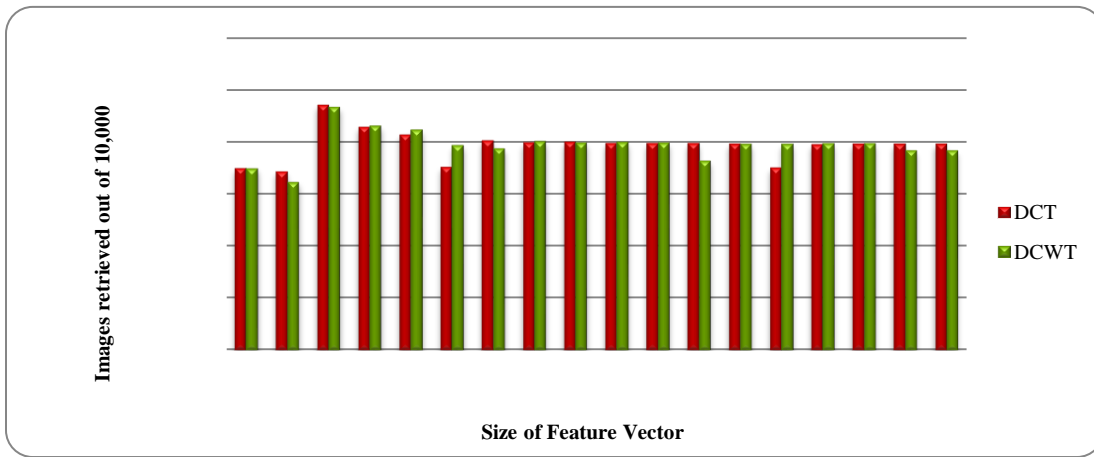


Figure 3. Plot for DCT and DCWT using Scale Down Factor 5

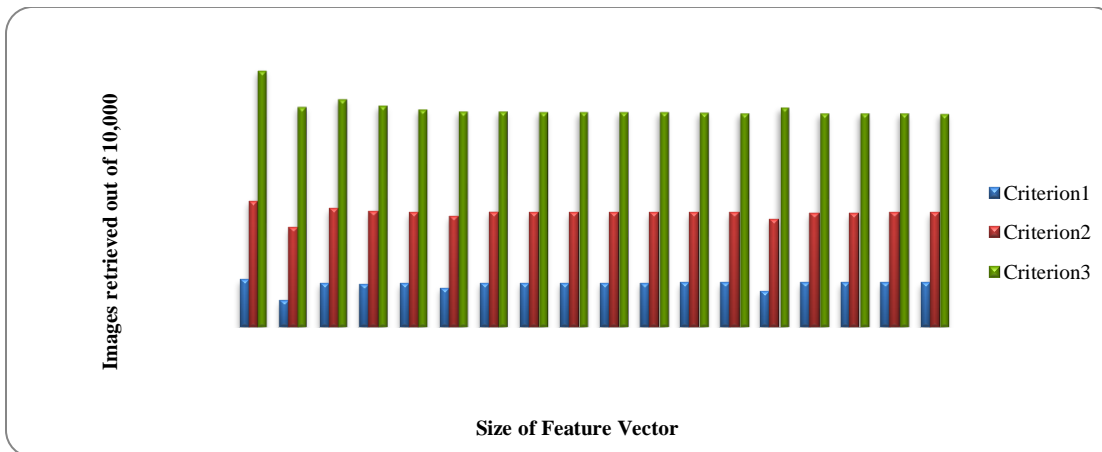


Figure 4. Plot for All 3 Criteria Using Scale Down Factor 10 for DCWT

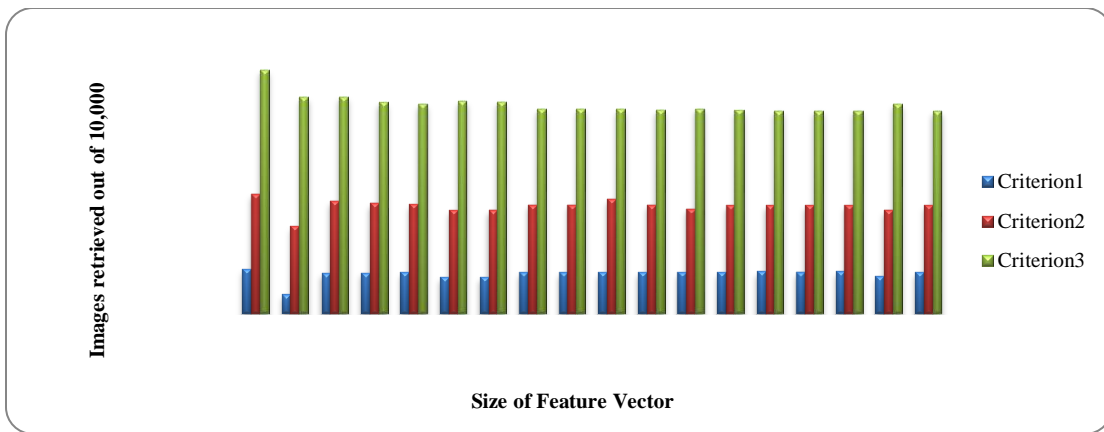


Figure 5. Plot for All 3 Criteria Using Scale Down Factor 10 for DCT

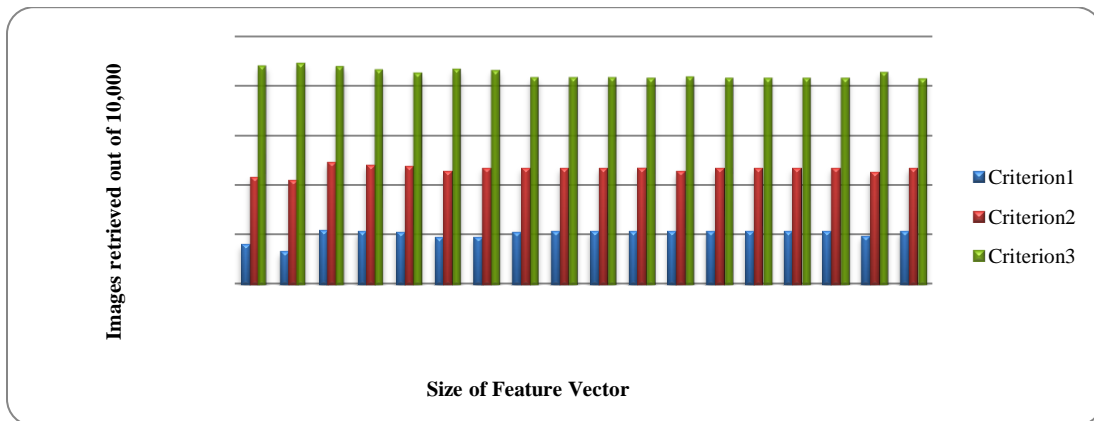


Figure 6. Plot for All 3 Criteria Using Scale Down Factor 5 for DCWT

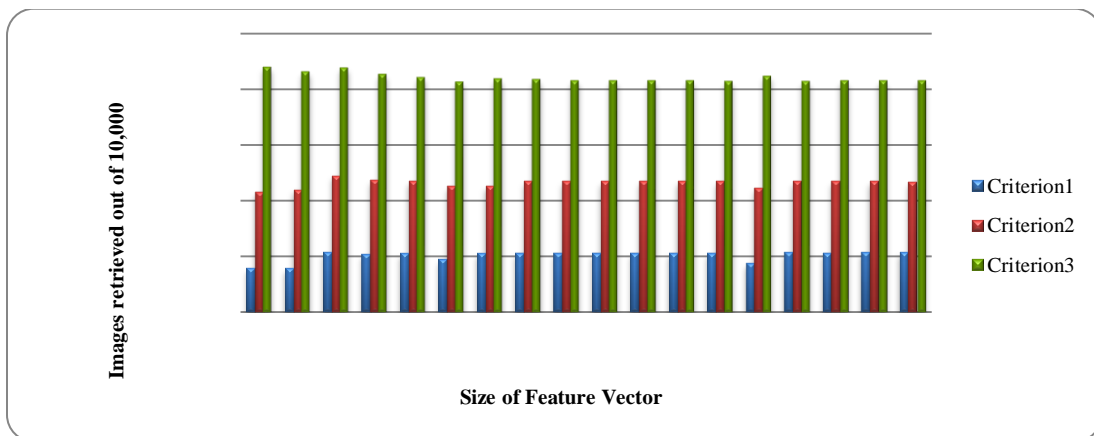


Figure 7. Plot for All 3 Criteria Using Scale Down Factor 5 for DCT

#### IV. CONCLUSION

CBIR system based on DCT and DCWT has been studied through many different aspects of its behavior in this paper. It mainly focuses on application of two transforms DCT and DCWT, their performance analysis and comparative study. This includes many things within it. In both the algorithms each image is divided into 3 planes that mean color information is handled separately to form the feature vectors. As each plane is divided into 4 blocks and transforms are

applied to row mean vectors of each block which tells that the texture of the image is taken into consideration while forming the feature vectors.

By changing the size of the feature vectors using 18 different sets computational time complexity is analyzed and it can be defined that computational time can be saved with smaller size feature vectors which are performing better as compared to the larger ones.

Along with the different size of feature vectors System's performance is also checked using the scale down factors 10 and 5 which actually stabilizes the high energy of first DCT or DCWT coefficient, brings it into the same range of remaining low energy coefficients. This gives the strong improvement in the retrieval results as shown in chart 1 and 2.

As three planes are handled separately each time 3 results sets are obtained which are further combined using three criteria to prove the best out of it where we can notice that criteria 3 is giving best performance among all 3.

Finally when we compare DCT and DCWT it can be noticed that DCWT is performing better. The best performance is given by DCWT with factor 5 at 16 coefficients as shown in figure 2. If properties of wavelet taken into consideration we can say that all small details of the image can be extracted to form the feature vectors and also maximum computational time can be saved as compare to normal DCT transform. Required multiplications using DCT for  $N \times N$  blocks are  $N^2$  where DCWT requires only  $N(2N-1)$  multiplications which saves considerable computational time of the system and gives better performance as well.

On the basis of comparison of this work with existing systems many places we found our results are better in terms of similarity retrieval and also in terms of computational time required. But as there is scope for further improvement so that these approaches can be used for variable image sizes and along with color and texture shape feature can also be considered for the comparisons and also the overall average precision and recall can further be improved for all 100 or more queries towards the ideal value.

#### REFERENCES

- [1] Remco C. Veltkamp, mirela tanase department of computing science, utrecht university, "content-based image retrieval systems: a survey" Revised and extended version of technical report uu-cs- 2000-34, october 28, 2002.
- [2] Yixin chen, member IEEE, james z. Wang, member IEEE, and robertkrovetz clue: "Cluster-Based Retrieval Of Images By Unsupervised Learning" IEEE Transactions On Image Processing, Vol.14, No. 8, August 2005.
- [3] Qasim Iqbal And J. K. Aggarwal, "Cires: A System For Content-Based Retrieval In Digital Image Libraries" Seventh International Conference On Control, Automation, Robotics And Vision (Icarcv'02), Dec 2002, Singapore.
- [4] Guoping Qiu "Color Image Indexing Using Btc" IEEE Transactions On Image Processing, Vol. 12, No. 1, January 2003.
- [5] R. W. Picard and T. P. Minka, "Vision texture for annotation," J. Multimedia Syst., vol. 3, no. 1, pp. 3-14, 1995.
- [6] S. Santini and R. Jain, "Similarity measures," IEEE Trans. Pattern Anal. Mach. Intell., vol., 2005, in press.
- [7] E. de Ves, A. Ruedin, D. Acevedo, X. Benavent, and L. Seijas, "A New Wavelet-Based Texture Descriptor for Image Retrieval", CAIP 2007, LNCS 4673, pp. 895-902, 2007, Springer-Verlag Berlin Heidelberg 2007.
- [8] H.B.Kekre, Dharendra Mishra, "DCT-DST Plane sectorization of Rowwise Transformed color Images in CBIR" International Journal of Engineering Science and Technology, Vol. 2 (12), 2010, 7234-7244.
- [9] H. B. Kekre, Kavita Patil, "Standard Deviation of Mean and Variance of Rows and Columns of Images for CBIR", International Journal of Computer and Information Engineering 3:1 2009.
- [10] H. B. Kekre, Archana Athawale, Dipali Sadavarti "Algorithm to Generate Wavelet Transform from an Orthogonal Transform", International Journal Of Image Processing (IJIP), Volume (4): Issue (4).
- [11] H.B.Kekre, Sudeep D. Thepade, Priyadarshini Mukherjee, Shobhit Wadhwa, Miti Kakaiya, Satyajit Singh, "Image Retrieval with Shape Features Extracted using Gradient Operators and Slope Magnitude Technique with BTC", International Journal of Computer Applications, September 2010 issue. (0975 – 8887) Volume 7– No.10, October 2010.
- [12] Samy Ait-Aoudia1, Ramdane Mahiou1, Billel Benzaid, "Yet Another Content Based Image Retrieval system", 1550-6037/10 \$26.00 © 2010 IEEE, DOI 10.1109/IV.2010.83
- [13] P. S. Hiremath, Jagadeesh Pujari, "Content Based Image Retrieval using Color, Texture and Shape features", -7695-3059-1/07\$25.00© 2007 IEEE, 10.1109/ADCOM.2007.21.
- [14] H. B. Kekre Kavita Sonawane, "CBIR Using Kekre's Transform over Row column Mean and Variance Vectors", International Journal on Computer Science and Engineering, Vol. 02, No. 05, 2010, 1609-1614.
- [15] H. B. Kekre, Kavita Patil, "DCT over Color Distribution of Rows and Columns of Image for CBIR" Sanshodhan – A Technical Magazine of SFIT No. 4 pp. 45-51. Dec.2008.
- [16] H.B.Kekre, Sudeep D. Thepade, "Image Retrieval using Augmented Block Truncation Coding Techniques", ACM International Conference on Advances in Computing, Communication and Control (ICAC3-2009), pp.: 384-390, 23-24 Jan 2009, Fr. Conceicao Rodrigous College of Engg., Mumbai. Available online at ACM portal.
- [17] H.B.Kekre, Dharendra Mishra, "Performance comparison of four, eight and twelve Walsh transform sectors feature vectors for image retrieval from image databases", International journal of Engineering, science and technology (IJEST) Vol.2(5) 2010, 1370-1374 ISSN 0975-5462.
- [18] H.B.Kekre, Tanuja Sarode, Sudeep D. Thepade, "Color-Texture Feature based Image Retrieval using DCT applied on Kekre's Median Codebook", International Journal on Imaging (IJI), Volume 2, Number A09, Autumn 2009, pp. 55-65. Available online at www.ceser.res.in/iji.html (ISSN: 0974-0627).
- [19] Subrahmanyam Murala, Anil Balaji Gonde, R. P. Maheshwari, "Color and Texture Features for Image Indexing and Retrieval", 2009 IEEE International Advance Computing Conference (IACC 2009), Patiala, India, 6-7 March 2009.
- [20] H. B. Kekre Ms. Kavita Sonawane, "Feature Extraction in Bins Using Global and Local thresholding of Images for CBIR", Published in International Journal of Computer, Information and System Science and Engineering. (IJCISSE), Vol. 3, No. 1, Winter 2009 pp.1- 4).
- [21] Lu, Z.-M.; Burkhardt, H., "Colour image retrieval based on DCT domain vector quantization index histograms" Electronics Letters Volume 41, Issue 17, 18 Aug. 2005 pp: 956 – 957.
- [22] Combes J, Grossmann A, Tchamitchian P. "Wavelets: Time-Frequency Methods and Phase Space". 2. Springer-Verlag; 1989.
- [23] M.K. Mandal, T. Aboulnasr, S. Panchanathan, "Image indexing using moments and wavelets", IEEE Trans. Consumer Electron. Vol. 42 No 3, 1996, pp.557-565.
- [24] Yung-Gi Wu1, Je-Hung Liu, "Image Indexing in DCT Domain" , Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)
- [25] C.C Chang, J.C Chuang, Y.S Hu, "Retrieving digital images from a JPEG compressed image database", Digital Image Vision Computing Vol. 22 2004, pp. 471-484.
- [26] K. Ait saadi, A.Zemouri, Z. Brahim, H.Meraoui. "Indexing and Retrieval Medical images based on 2X2 DCT and IDS Compression", Proceedings of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05).
- [27] Ramesh Babu Durai, V.Duraisamy "A generic approach to content based image retrieval using dct and classification techniques", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 2022-2024.
- [28] H.B.Kekre, Tanuja Sarode, Sudeep D. Thepade, "Image Retrieval using Color-Texture Features from DCT on VQ Codevectors obtained by



- Kekre's Fast Codebook Generation", ICGST-GVIP Journal, Volume 9, Issue 5, September 2009, ISSN: 1687-398X.
- [29] M.Babu Rao, B.Prabhakara Rao, A.Govardhan, "Content based image retrieval using Dominant color and Texture features", International Journal of Computer science and information security, Vol.9 issue No:2, February 2011, pp:41-46.
- [30] Kishore Kumar et al. "Content based image retrieval - extraction by objects of user interest", International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011.
- [31] Gwangwon Kang, Junguk Beak "Features Defined by Median Filtering on RGB Segments for Image Retrieval", Second UKSIM European Symposium on Computer Modeling and Simulation, 978-0-7695-3325-4/08, 2008 IEEE.
- [32] Yu-Len Huang and Ruey-Feng Chang, "Texture features for dct-coded image Retrieval and classification", 0-7803-5041 -3/99, 1999 IEEE.
- [33] Chong-Wah Ngo, Ting-Chuen Pong, "Exploiting image indexing techniques in DCT domain", Pattern Recognition 34 (2001) 1841-1851 Published by Elsevier Science Ltd.
- [34] S.Cheng, W. Huang, Y. Liao and D. Wu, "A Parallel CBIR Implementation Using Perceptual Grouping Of Block-based Visual Patterns", IEEE International Conference on Image Processing – ICIP, 2007, pp. V -161 - V – 164,
- [35] Mann-Jung Hsiao, Yo-Ping Huang, Te-Wei Chiang, "A Region-Based Image Retrieval Approach Using Block DCT", 0-7695-2882-1/07, 2007 IEEE
- [36] Kekre Transform over Row Mean, Column Mean and Both Using Image Tiling for Image Retrieval", International Journal of Computer and Electrical Engineering, Vol.2, No.6, December, 2010, 1793-8163.
- [37] M. Saadatmand-Tarzjan and H. A. Moghaddam, "A Novel Evolutionary Approach for Optimizing Content-Based Image Indexing Algorithms", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 37, No. 1, February 2007, pp. 139-153.

#### AUTHORS PROFILE



Dr. H. B. Kekre has received B.E. (Hons.) in Telecomm. Engg. from Jabalpur University in 1958, M.Tech (Industrial Electronics) from IIT Bombay in 1960, M.S. Engg. (Electrical Engg.) from University of Ottawa in 1965 and Ph.D. (System Identification) from IIT Bombay in 1970. He has worked Over 35 years as Faculty of Electrical Engineering and then HOD Computer Science and Engg. at IIT Bombay. For last 13 years worked as a Professor in Department of Computer Engg. at Thadomal Shahani Engineering College, Mumbai. He is currently Senior Professor working with Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS University, Vile Parle(w), Mumbai, INDIA. He has guided 17 Ph.D.s, 150 M.E./M.Tech Projects and several B.E./B.Tech Projects. His areas of interest are Digital Signal processing, Image Processing and Computer Networks. He has more than 350 papers in National / International Conferences / Journals to his credit. Recently twelve students working under his guidance have received best paper awards. Five of his students have been awarded Ph. D. of NMIMS University. Currently he is guiding eight Ph.D. students. He is member of ISTE and IETE.



Ms. Kavita V. Sonawane has received M.E (Computer Engineering) degree from Mumbai University in 2008, currently Pursuing Ph.D. from Mukesh Patel School of Technology, Management and Engg, SVKM's NMIMS University, Vile-Parle (w), Mumbai, INDIA. She has more than 8 years of experience in teaching. Currently working as a Assistant professor in Department of Computer Engineering at St. Francis Institute of Technology Mumbai. Her area of interest is Image Processing, Data structures and Computer Architecture. She has 7 papers in National/ International conferences / Journals to her credit.

# Cryptanalysis of an Advanced Authentication Scheme

Sattar J Aboud

Department of Information Technology  
Iraqi Council of Representatives  
Baghdad-Iraq

Abid T. Al Ajeeli

Department of Information Technology  
Iraqi Council of Representatives  
Baghdad-Iraq

**Abstract**—In this paper we study a scheme for making cryptanalysis and security improvement. This protocol by Song, is a password authentication protocol using smart card. We note that this protocol has been shown to be prone to the offline password guessing attack. We perform an additional cryptanalysis on this scheme and detect that it is vulnerable to the clogging attack, a type of denial-of-service attack. We notice that all smart card typed authentication schemes which lead the scheme by Song, and need the server to find the computationally exhaustive modular exponentiation, similar to the scheme by Xu et al., and it is vulnerable to the clogging attack. Then we propose an enhancement in the scheme to avoid the clogging attack.

**Keywords**-authentication protocol; offline password guessing attack; clogging attack.

## I. INTRODUCTION

The idea behind improving password authentication protocol is to help authorized users obtain services from an authorized server. When an entity needs a service from a server, it has to identify itself to the server in a certain way. Password authentication has been one of the most suitable techniques for a user ID throughout the years. Currently, millions of providers utilize password authentication schemes to identify authorized users. General cases include private web service, Internet shopping, e-mail service, e-trade service, and other services. Fundamentally, each password authentication method has two steps:

- Registration step: In this step, the user enters a user ID and password in the computer. The password is saved by the server and kept confidential between the entity and the computer.
- Authentication step: In this step, the entity needs a service from the computer. It passes its identity and password to the computer to get a service. The computer then decides if the user is authorized by checking the received information of user ID password with the saved details. The server extends the preferred service to the entity, if found authorized

In authentication processes, the password is sent cross an insecure communication channel. A hacker can intercept the message by listening to the communication. It can imitate the entity by re-using the password acquired from the communication. These issues compromise the entity confidentiality. The service providers normally keep passwords of users in a database for potential verification and authentication. The passwords are kept in a password index in

the computer database. It does not provide any security against unprivileged insiders of the computer, and it does not protect the passwords when the computer database is somehow hacked. To reduce the problem of password list disclosure, the computer can encrypt the passwords and protect them. Nevertheless, the communication interception remains a threat to the organization security. Another problem is recalling a user identity and password. A compromise in the password for any entity can be like losing a credit card. The hacker can take advantage before the corporation is informed about the loss. Therefore we have to improve efficient password authentication protocol to finish the authentication process in a secured way.

Taking this into consideration, and to produce a more secure scheme, various smart cards typed password authentication schemes have been proposed throughout the last decade [4, 6, 7]. In such a system, the entity is given with a smart card. When the user needs a service, it gives its smart card with a password that remains private. The smart card then employs this password to build a login message that is passed to the computer. The computer then authenticates this message and gives the preferred service if the password is found legitimate.

In this paper, we study such a scheme under smart card by Song [1]. Song presented the scheme as an enhancement of another scheme set is by Xu et al. [3]. The scheme has been shown to be prone to the imitation attack. Song then considered this competitive scheme in the same study, to avoid the imitation attack on the scheme. In this paper, we will show that the Song protocol is prone to the clogging attack, a type of the denial-of-service-attack. We noted that the scheme in [3] was vulnerable to clogging attack and also prone to clogging attack. In this attack, the hacker can easily prevent the computer from giving any service without having any information related to the identity or password of the entity. The hacker wants to make any complex computations to launch a clogging attack on Song scheme. We have come to know a cryptanalysis of the Song scheme by Tapiador et al [2]. However, they did not study the clogging attack in the scheme in their research. It also worth mentioning that the chain of smart card typed authentication schemes that keeps the server computationally exhaustive modular exponentiation [5] are all prone to the clogging attack.

The hacker takes the benefit of the calculation intensiveness of the modular exponentiation computation in initiating this attack. To avoid clogging attack on these schemes key agreement schemes are required in different

types of communication. Keys want to be securely exchanged before a channel can be recognized. There are few security threats to this that are intruder-in-the-middle so that the hacker pretends to be someone else than connecting participants. Replay of old keys is one more attack which is common in this viewpoint. Therefore it is necessary to improve secure key exchanging schemes to create secure channel. The key agreement protocols commonly have two steps:

- Determine the public and private keys: In this step, both the participants compute a pair of keys: the secret key, which is kept private, and the public key, which is made public. In some schemes, a part of this step is completed by a key distribution center that keeps the public keys of the entities in a directory.
- Determine the secret session key: In this step, the users swap their public keys with a certain integer. The secret session key for message is computed using those integers, and the secret and public keys. Several schemes also let a certain participant determine the private key or session key and pass it to the other participant by encrypting it with the message. In this paper we will not study any scheme regarding the key agreement protocols.

The rest of the paper is organized as follows: in section 2, we study the Song scheme including some an enhancement we suggest concerning public and private keys ; then we give a toy example in section 3. Then we have a cryptanalysis of Song scheme .including clogging attack and offline password guessing attack in section 4. We then propose a solution to this attack in section 5. In section 6, we conclude this work.

## II. REVIEW OF SONG SCHEME

We summarize here the password authentication protocol of Song [1]. This authentication protocol uses a smart card. In section 3, we consider the achievement and security vulnerability of this protocol. In section 4, we introduce a clogging attack on the scheme. We also consider a possible solution against the attack. Finally, we show that similar protocols of [3], that let the server calculate modular exponentiation, are vulnerable to the same type of attack and has the same achievement dependencies as this protocol. The Song protocol contains three steps: registration, login and authentication. Prior to starting with the registration phase, the server performs the following steps:

1. two primes  $p$  and  $q$  are chosen where  $p = 2 * q + 1$ .
2. an integer  $i \in Z_q^*$  is chosen.
3. a hash function  $h$  is chosen.
4. an integer  $e$  is chosen as encryption key such that  $\gcd(e, p - 1) = 1$ .
5. the decryption key  $d$  is computed, where  $d = e^{-1} \bmod p - 1$ .
6.  $i$  and  $d$  are determined as both are private keys.

### A. Registration Phase

This phase comprises the following steps:

1. Entity  $A$  selects  $(id_A, w_A)$
2. Entity  $A$  passes  $(id_A, w_A)$  to the server over a secure way, such that  $id_A$  and  $w_A$  are the user-id and password of entity  $A$  respectively.
3. The server calculates  $z_A = h(id_A^i \bmod p) \oplus h(w_A)$ .
4. The server saves the parameter  $(id_A, z_A, h, e)$  into a smart card.
5. The server sends the smart card to the entity  $A$ .

### B. Logical Phase

Entity  $A$  performs the following steps:

1. Provides its  $(id_A, w_A)$ .
2. Chooses an arbitrary integer  $r_A$ .
3. Sets  $T_A \leftarrow$  system present time.
4. Finds  $y_A = z_A \oplus h(w_A)$ .
5. Computes  $x_A = e_{y_A}(r_A \oplus T_A)$ .
6. Computes  $u_A = h(T_A || r_A || x_A || id_A)$ .
7. Sends the message  $(id_A, u_A, x_A, T_A)$  to the server.

### C. Authentication Phase

The following steps are performed by the server:

1. the server performs the following after receiving the login request from an entity:
  - Verifies that  $id_A$ , and  $T_A$ . If not, rejects the login message.
  - Finds  $y_A = h(id_A^i \bmod p)$ .
  - Computes  $r'_A = d_{y_A}(x_A) \oplus T_A$
  - Verifies that  $u_A \equiv h(T_A || r'_A || x_A || id_A)$ . Otherwise rejects the login message.
  - Finds  $u_S = h(id_A || r'_A || T_S)$ ,
  - Passes the message  $(id_A, u_S, T_S)$  to an entity  $A$ .
2. The following steps are performed by an entity  $A$  after the receipt of  $(id_A, u_S, T_S)$  from a server:
  - authenticates  $id_A$  and  $T_S$
  - Checks that  $u_S \equiv h(id_A || r_A || T_S)$ . If identical, the server is validated.
3. The entity  $A$  and the server then find the session key  $s$  as follows:
  - Entity  $A$  compute the session key as follows  $s = h(id_A || T_S || T_A || r_A)$
  - The server computes the session key as follows  $s = h(id_A || T_S || T_A || r'_A)$ .
  - Finally, a session key  $s$  is agreed by the two participants.

### D. Comments

When the entity wants to alter the password  $w_A$  to a new password  $w'_A$ , the smart card will first verify the validity of

$w_A$  through interacting with the server, and if it succeeds, it resets  $w_A$  to  $w'_A$ , and substitutes  $z_A$  with  $z'_A = z_A \oplus w_A \oplus w'_A$ .

### III. TOY EXAMPLE YOUR

Suppose that the server selects  $p=23$ , then  $q=11$ . Suppose that  $i=2$ , and for simplicity ignore in this phase the hash function  $h(\cdot)$ . Suppose the encryption key  $e=7$ , and then compute the decryption key  $d$  as follows  $e * d \equiv 1 \pmod{p-1} = 7 * 63 \pmod{22} = 1$ .

**Registration Phase:** Suppose that entity  $A$  selects  $(id_A = 29, w_A = 24)$ , then entity  $A$  passes  $(id_A = 29, w_A = 24)$  to the server over a secure way so that  $id_A$  and  $w_A$  are the user-id and password of entity  $A$ . Then, the server calculates  $z_A = h(id_A^i \pmod{p}) \oplus h(w_A)$   
 $= h(29^2 \pmod{23}) \oplus h(24) = 21$ . The server then saves the parameter  $(id_A = 29, z_A = 21, h(\cdot), e = 7)$  into a smart card. Finally, the server sends the smart card to the entity  $A$ .

**Login Phase:** First, the smart card is attached to the card reader by an entity  $A$  and it provides  $(id = 29, w_A = 24)$ . The smart card then chooses a random integer  $(r_A = 12)$  and then sets  $(T_A = 4)$ . Then entity  $A$  computes  $y_A = z_A \oplus h(w_A) = 21 \oplus h(24) = 13$ . Then compute  $x_A = e_{y_A} (r_A \oplus T_A) = 13(12 \oplus 4) = 12$ . compute  $x_A = 12^7 \pmod{23} = 16$ . Then compute  $u_A = h(T_A || r_A || x_A || id_A)$   
 $= h(4 + 12 + 16 + 29) = 61 \pmod{23} = 15$ . Then entity  $A$  sends the message  $(id_A = 29, u_A = 15, x_A = 16, T_A = 4)$  to the server.

**Authentication Phase:** First, the server verifies that  $id_A$  is valid. If not, it rejects the login message. Then,  $y_A = h(id_A^i \pmod{p}) = h(29^2 \pmod{23}) = 13$  is computed. Then  $r'_A = d_{y_A} ((x_A^d \pmod{p}) \oplus T_A) = 13((16^{63} \pmod{23}) \oplus 4) = (12 \oplus 4) = 8 * 13 \pmod{23} = 12$  is computed. Then, the server verifies that  $u_A \equiv h(T_A || r'_A || x_A || id_A) = h(4 + 12 + 16 + 29) = 61 \pmod{23} = 15$ . If the verification of the previous step succeeds, set  $T_s$  as current server time. Suppose that  $T_s = 4$ , and then compute  $u_s = h(id_A || r'_A || T_s) = h(29 + 12 + 4) = 45 \pmod{23} = 22$ . Finally, the server passes the message  $(id_A = 29, u_s = 22, T_s = 4)$  to entity  $A$ . An entity  $A$  authenticates  $id_A$  and  $T_s$  by Checking that  $u_s \equiv h(id_A || r_A || T_s) = h(29 + 12 + 4) = 22$ . In this example, it is identical; it means that the server is validated. However, the entity  $A$  and the server then find the session key  $s$  as follows: first, entity  $A$  computes the session key as follow  $s = h(id_A || T_s || r_A) = h(29 + 4 + 12) = 49 \pmod{23} = 3$ . Then, the server computes the session key as follows  $s = h(id_A || T_s || r'_A) = h(29 + 4 + 12) = 49 \pmod{23} = 3$ .

### IV. SONG SCHEME CRYPAYLISIS

The clogging attack and offline password guessing attack will be discussed in this section.

#### A. The Clogging Attack

This scheme has a large dependency on the computer and entity clocks. For a connection-oriented use, this may be unwieldy. The scheme should be designed to care for time synchronization between clocks of different entities and servers. This scheme should be made fault tolerant to deal with complex network faults and also with different types of attacks. Despite that the communication is secure, a chance of an attack can occur and a hacker may intercept a message and alter its timestamp  $T_A$ . In this way the hacker successfully repudiates the authorized entity because the server will refuse the login message on the basis of timestamp dissimilarity. Thus this type of attack is probable even if the scheme avoids replay attacks. Also, interacting delays can prepare the timestamp to go beyond the threshold thus making the entire service inherently decelerate. Therefore, we illustrate that Song scheme is vulnerable to the clogging attack. The clogging attack is a type of attack by which the hacker  $H$  constantly passes messages to a server and clogs it with those messages [8]. Suppose this could occur with the Song scheme. The following is done by a hacker  $H$ :

1.  $H$  intercepts the message  $(id_A = 29, u_A = 11, x_A = 12, T_A = 4)$  passed by an entity to a server in the login phase.
2.  $H$  can alter the timestamp  $T_A = 4$  to some  $T_u = 10$ , because the message is unencrypted. The change satisfies the criterion  $T^* - T_u \leq g$ .
3.  $H$  alters  $u_A = 11$  to arbitrary nonsense value  $u_u = 9$ .
4.  $H$  passes  $(id_A = 29, u_u = 9, x_A = 12, T_u = 10)$  to the server.

The following is performed by the server:

1. Verify if  $id_A = 29$  is valid. At this point it is valid.
2. Find  $y_A = h(id_A^i \pmod{p}) = h(29^2 \pmod{23}) = 13$ .
3. Find  $r'_A = d_{y_A} (x_A \oplus T_u) = 13(12 \oplus 10) = 9$
4. verify  $u_u = h(T_u || r'_A || x_A || id_A) = h(10 + 9 + 12 + 29) = 60 \pmod{23} = 14$ . This does not succeed, thus the message is rejected.

Then, the hacker  $H$  will continue repeating the steps many times and let the server calculate the modular exponentiation repeatedly. Essentially,  $H$  can potentially alter all the entering login requests from the authorized entity to the server. As modular exponentiation is computationally exhaustive, the victimized server spends large processing resources doing ineffective modular exponentiation rather than any actual work. Therefore the hacker  $H$  clog the server with ineffective work and so repudiates any authorized entity. The hacker only wants an  $id$  of a valid entity to achieve the clogging attack many times.

### B. Offline Password Guessing Attack

In [1] it is claimed that the hacker should not be able to attack and get access to the server by extracting the information kept on the smart card. But, a hacker who gets the result of  $z_A = h(id_A^i \text{ mod } p) \oplus h(w_A)$  can simply increase an offline password guessing attack by easily viewing one right authentication session and obtaining access to the results of  $x_A$ , and  $u_A$ . The attack is performed as follows: for every password  $w_A^*$ , the hacker calculates the uncertain encryption key  $y_A^* = z_A \oplus h(w_A^*)$ . Such a key is then employed to decrypt the nonce result  $r_A^*$  by first recovering  $x_A$  with  $y_A^*$  and then  $\oplus$  the value with  $T_A$  (both of which are public); namely,  $r_A^* = d_{y_A^*}^*(x_A) \oplus T_A$ . Note that when an attempt of password  $w_A^*$  is right (i.e.,  $w_A^* = w_A$ ), then it is obtained encryption key  $y_A^*$  and so, the nonce  $r_A^*$ . Now,  $u_A$  can be used to verify when that is the case. The hacker finds  $u_A^* \equiv h(T_A || r_A^* || x_A || id_A)$  and, when it coincides with  $u_A$ , it can conclude that  $r_A^*$  is right and thus the password tried. In this reasoning we suppose that  $h$  has no collisions. Yet, even when  $h$  is not perfect, extra eavesdropping sessions can be employed to exclude false positives and find the right password. Briefly, in addition to what is claimed in [1], the messages exchanged during the scheme certainly decrease the entropy of the password, at least for a hacker with access to the values stored in the card. Also, once the password is guessed, the scheme provides no protection against other attacks.

### C. Comments

It can be observed that the attack showed can also be made on the schemes by Xu et al. [3] and by Tsaur et al. [5]. Thus we notice that the clogging attack can be executed on all the smart card typed authentication schemes using computing modular exponentiation.

## V. THE POSSIBLE SOLUTION

We will discuss the possible solution for the problem raised.

### A. Prevent The Clogging Attack

At the start of the authentication phase, the server will check if the IP address of the entity is valid. It has to identify the IP addresses of any registered authorized users. Despite that, hacker  $H$  might spoof the IP address of an authorized entity and replay the login request. To stop it, we may add a cookie exchange step at the start of the login phase of Song scheme. This step has been presented as in the familiar Oakley key exchange scheme [9].

1. The entity  $A$  selects an arbitrary number  $m_1$  and passes it with the message  $(id_A, u_A, x_A, T_A)$  to the server.
2. The server accepts the message and passes its own cookie  $m_2$  to the entity  $A$ .

### B. Solution Discussion

When the hacker  $H$  spoofs the entity IP address,  $H$  will not obtain  $m_2$  back from the server. But  $H$  just succeeds to have the server return an acknowledgement, not to calculate the computationally modular exponentiation. Thus the clogging attack is prevented by these extra steps. We note that this process does not avoid the clogging attack but only frustrates it to a certain extent.

## VI. CONCLUSION

We have studied a scheme in this paper. The scheme is a password authentication protocol; which we have shown to be vulnerable to the clogging attack. We demonstrated that the attack on this scheme could be prevented by using an extra step of exchanging numbers. We demonstrated that it is prone to man-in-the-middle attack. Then we showed how to prevent this attack by using an encryption and decryption algorithm. We indicated a security get-out as Song proposed, which is that the hacker can execute modular exponentiation on both sides of the authentication scheme. In addition, after intercepting the retrieved information, the hacker can start new logon information and successfully log into the server system. Thus, Song proposition cannot give adequate security and it is not appropriate for practical implementation of the proposed scheme.

## REFERENCES

- [1] Ronggong Song, "Advanced smart card based password authentication protocol", Computer Standards & Interfaces, Volume 32, Issue 4, pp. 321-325, June 2010.
- [2] Juan E. Tapiador, Julio C. Hernandez-Castro, Pedro Peris-Lopez, John A. Clark, "Cryptanalysis of Song's advanced smart card based password authentication protocol", Unpublished manuscript, June 2010.
- [3] Jing Xu, Wen-Tao Zhu, and Deng-Guo Feng, "An improved smart card based password authentication scheme with provable security", Computer Standards & Interfaces, Volume 31, Issue 4, pp. 723-728, June 2009.
- [4] Chang. C.C., Wu T.C., "Remote password authentication scheme with smart cards", IEEE Proceedings Computers and Digital Techniques, Volume 138, Issue 3, pp.165-168, 1991.
- [5] Woei-Jiunn Tsaur, Chia-Chun Wu, and Wei-Bin Lee, "A smart card based remote scheme for password authentication in multi-server Internet services", Computer Standards & Interfaces, Volume 27, pp. 39-51, June 2004.
- [6] Wen-Sheng Juang, "Efficient password authenticated key agreement using smart cards", Computers & Security, Volume 23, Issue 2, pp. 167-173, March 2004.
- [7] Chien H.Y., Jan J.K., and Tseng Y.M., "An efficient and practical solution to remote authentication: smart card", Computers and Security, vol.21, no.4, pp.372-375, 2002.
- [8] Tseng Y.M., "An efficient two-party identity-based key exchange protocol", Informatica 18 (1) pp. 125-136, 2007.
- [9] Hsi-Chang Shih, "Cryptanalysis on Two Password Authentication Schemes", Master Thesis, Laboratory of Cryptography and Information Security Department of Computer Science and Information Engineering, National Central University, Chung-Li, Taiwan 320, Republic of China, 2006.
- [10] Meshram, C. (2010). Modified ID-Based Public key Cryptosystem using Double Discrete Logarithm Problem. IJACSA - International Journal of Advanced Computer Science and Applications, 1(6).
- [11] Meshram, C. (2011). Some Modification in ID-Based Public key Cryptosystem using IFP and DDLP. IJACSA - International Journal of Advanced Computer Science and Applications, 2(8).

### AUTHORS PROFILE

**Sattar J Aboud** is a Professor and advisor for Science and Technology at Iraqi Council of Representatives. He received his education from United

Kingdom. Dr. Aboud has served his profession in many universities and he awarded the Quality Assurance Certificate of Philadelphia University, Faculty of Information Technology in 2002. Also, he awarded the Medal of Iraqi Council of Representatives for his conducting the first international conference of Iraqi Experts in 2008. His research interests include the areas of both symmetric and asymmetric cryptography, area of verification and validation, performance evaluation and e-payment schemes.

**Abid T. Al Ajeeli** is a Professor at Iraqi Council of Representatives. He received his education from United Kingdom. Dr. Ajeeli has served his profession in many universities and he awarded the Medal of Iraqi Council of Representatives for his conducting the first international conference of Iraqi Experts in 2008. His research interests include the areas of software engineering, verification and validation, information security and simulation.



# Conceptual Level Design of Semi-structured Database System: Graph-semantic Based Approach

Anirban Sarkar

Department of Computer Applications  
National Institute of Technology, Durgapur  
West Bengal, India

**Abstract**—This paper has proposed a Graph – semantic based conceptual model for semi-structured database system, called GOOSSDM, to conceptualize the different facets of such system in object oriented paradigm. The model defines a set of graph based formal constructs, variety of relationship types with participation constraints and rich set of graphical notations to specify the conceptual level design of semi-structured database system. The proposed design approach facilitates modeling of irregular, heterogeneous, hierarchical and non-hierarchical semi-structured data at the conceptual level. Moreover, the proposed GOOSSDM is capable to model XML document at conceptual level with the facility of document-centric design, ordering and disjunction characteristic. A rule based transformation mechanism of GOOSSDM schema into the equivalent XML Schema Definition (XSD) also has been proposed in this paper. The concepts of the proposed conceptual model have been implemented using Generic Modeling Environment (GME).

**Keywords**- *Semi-structured Data; XML; XSD; Conceptual Modeling; Semi-structured Data Modeling; XML Modeling.*

## I. INTRODUCTION

The increasingly large amount of data processing on the web based applications has led a crucial role of semi-structured database system. In recent days, semi-structured data has become prevalent with the growing demand of such web based software systems. Semi-structured data though is organized in semantic entity but does not strictly conform to the formal structure to strict types. Rather it possess irregular and partial organization [1]. Further semi-structured data evolve rapidly and thus the schema for such data is large, dynamic, is not strict to type and also is not considered the participation of instances very strictly.

The eXtensible Markup Language (XML) is increasingly finding acceptance as a standard for storing and exchanging structured and semi-structured information over internet [12]. The Document Type Definition (DTD) or XML Schema Definition (XSD) language can be used to define the schema which describes the syntax and structure of XML documents [9]. However, the XML schemas provide the logical representation of the semi-structured data and it is hard to realize the semantic characteristics of such data. Thus it is

important to devise a conceptual representation of semi-structured data for designing the information system based on such data more effectively. A conceptual model of semi-structured data deals with high level representation of the candidate application domain in order to capture the user ideas using rich set of semantic constructs and interrelationship thereof. Such conceptual model will separate the intention of designer from the implementation and also will provide a better insight about the effective design of semi-structured database system. The conceptual design of such system further can be implemented in XML based logical model.

To adopt the rapidly data evolving characteristics, the conceptual model of semi-structured data must support several properties like, representation of irregular and heterogeneous structure, hierarchical relations along with the non – hierarchical relationship types, cardinality, n – array relation, ordering, representation of mixed content etc. [13]. Beside these, it is also important to realize the participation constraints of the instances in association with some relation or semi-structured entity type. The participation of instances in semi-structured data model is not strict. In early years, Object Exchange Model has been proposed to model semi-structured data [2], where data are represented using directed labeled graph. The schema information is maintained in the labels of the graph and the data instances are represented using nodes. However, the separation of the structural semantic and content of the schema is not possible in this approach. In recent past, several researches have been made on conceptual modeling of semi-structured data as well as XML. Many of these approaches [3, 4, 5, 6, 7, 8] have been extended the concepts of Entity Relationship (ER) model to accommodate the facet of semi-structured data at conceptual level. The major drawbacks of these proposals are in representation of hierarchical structure of semi-structured data. Moreover, only two ER based proposals [7, 8] support the representation of mixed content in conceptual schema. In [7], a two level approach has been taken to represent the hierarchical relations. In first level the conceptual schema is based on extended concept of ER model and in second level, hierarchical organizations of parts of the overall conceptual schema are designed. In general, ER model are flat in nature [14] and thus unable to facilitate the reuse capability and representation of hierarchical relationship very efficiently. On the other hand, ORA-SS [11] proposed to realize the semi-structured data at conceptual level starting from its hierarchical structure. But the approach does not support directly the representation of

no-hierarchical relationships and mixed content in conceptual level semi-structured data model.

Very few attempts have been made to model the semi-structured data using Object Oriented (OO) paradigm. ORASS [11] support the object oriented characteristic partially. The approaches proposed in [9, 10, 12] are based on UML. These approaches support object oriented paradigm comprehensively and bridge the gap between OO software design and semi-structured data schemata. However, the UML and extensions to UML represent software elements using a set of language elements with fixed implementation semantics (e.g. methods, classes). Henceforth, the proposed approaches using extension of UML, in general, are logically inclined towards implementation of semi-structured database system. This may not reflect the facet of such system with high level of abstraction to the user. In other word, semi-structured data model with UML extension cannot be considered as semantically rich conceptual level model. In [16] a graph semantic based web data model has been proposed and is appropriate for modeling structured web database system. The approach has not considered semi-structured characteristics of web databases.

In this paper, a graph semantic based conceptual model for semi-structured database system, called Graph Object Oriented Semi-Structured Data Model (GOOSSDM), has been proposed. The model is comprehensively based on object oriented paradigm. Among others, the proposed model supports the representation of hierarchical structure along with non-hierarchical relationships, mixed content, ordering, participation constraints etc. The proposed GOOSSDM reveals a set of concepts to the conceptual level design phase of semi-structured database system, which are understandable to the users, independent of implementation issues and provide a set of graphical constructs to facilitate the designers of such system. The schema in GOOSSDM is organized in layered approach to provide different level of abstraction to the users and designers. In this approach a rule based transformation mechanism also has been proposed to represent the equivalent XML Schema Definitions (XSD) from GOOSSDM schemata. The correctness of such transformation has been verified using the structural correlation mechanism described in [15]. Moreover, the concepts of proposed GOOSSDM have been implemented using Generic Modeling Environment (GME) [14] which is a meta-configurable modeling environment. The GME implementation can be used as prototype CASE tools for modeling semi-structured databases using GOOSSDM.

The preliminary version of this work has been published in [17] which has been now enriched and completed with comprehensive formalization and CASE tools.

## II. GOOSSDM: THE PROPOSED MODEL

The GOOSSDM extends the object oriented paradigm to model semi-structured data. It contains all the details those are necessary to specify the irregular and heterogeneous structure, hierarchical and non-hierarchical relations, n - array relationships, cardinality and participation constraint of instances. The proposed data model allows the entire semi-structured database to be viewed as a Graph (V, E) in layered organization. At the lowest layer, each vertex represents an

occurrence of an attribute or a data item, e.g. name, day, city etc. Each such basic attribute is to be represented as separate vertex. A set of vertices semantically related is grouped together to construct an *Elementary Semantic Group (ESG)*. So an ESG is a set of all possible instances for a particular attribute or data item. On next, several related ESGs are group together to form a *Contextual Semantic Group (CSG)*. Even the related ESGs with non-strict participations or loosely related ESGs are also constituent of related CSG – the constructs of related data items or attributes to represent one semi-structured entity or object. The edges within CSG are to represent the containment relation between different ESG in the said CSG. The most inner layer of CSG is the construct of highest level of abstraction or deeper level of the hierarchy in semi-structured schema formation. This layered structure may be further organized by combination of one or more CSGs as well as ESGs to represent next upper level layers and to achieve further lower level abstraction or higher level in the semi-structure data schema hierarchy. From the topmost layer the entire database appears to be a graph with CSGs as vertices and edges between CSGs as the association amongst them. The CSGs of topmost layer will act as roots of semi-structured data model schemata.

### A. Modeling Constructs in GOOSSDM

Since from the topmost layer, a set of vertices V is decided on the basis of level of data abstraction whereas the set of edges E is decided on basis of the association between different semantic groups. The basic components for the model are as follows,

A set of  $t$  distinct attributes  $A = \{a_1, a_2, \dots, a_t\}$  where, each  $a_i$  is an attribute or a data item semantically distinct.

(a) *Elementary Semantic Group (ESG)*: An elementary semantic group is an encapsulation of all possible instances or occurrences of an attribute, that can be expressed as graph  $ESG(V, E)$ , where the set of edges E is a null set  $\emptyset$  and the set of vertices V represent the set of all possible instances of an attribute  $x_i \in A$ . ESG is a construct to realize the elementary property, parameter, kind etc. of some related concern. Henceforth there will be set of  $t$  ESGs and can be represented as  $E_G = \{ESG_1, ESG_2, \dots, ESG_t\}$ . The graphical notation for the any ESG is *Circle*.

(b) *Contextual Semantic Group (CSG)*: A lowest layer contextual semantic group is an encapsulation of set of ESGs or references of one or more related ESGs to represent the context of one entity of semi-structured data. Let, the set of  $n$  CSGs can be represented as  $C_G = \{CSG_1, CSG_2, \dots, CSG_n\}$ . Then any lowest layer  $CSG_i \in \subseteq C_G$  can be represented as a graph  $(V_{C_i}, E_{C_i})$  where vertices  $V_{C_i} \in E_G$  and the set of edges  $E_{C_i}$  represents the association amongst the vertices. For any CSG, it is also possible to designate one or more encapsulated ESGs as *determinant vertex* which may determine an unordered or ordered set of instances of constituent ESGs or CSGs. The graphical notation for any CSG is *square* and determinant vertex is *Solid Circle*.

Composition of multiple CSGs can be realized in two ways. *Firstly*, the simple *Association* (Discussed in subsection II.B) may be drawn between two or more associated CSGs






either of same layer or of adjacent layers to represent the non-hierarchical and hierarchical data structure respectively. The associated CSG will be connected using *Association Connector*. Those CSGs may share a common set of ESGs or referred ESGs.

Secondly, lower layer CSGs may maintain an *Inheritance* or *Containment* relationship (Discussed in subsection 2.B) with the adjacent upper layer CSG to represent the different level of abstraction. Thus, the upper layer CSG can be formed by inheritance or composition of one or more lower layer CSGs along with encapsulation of zero or more related ESGs or reference of ESGs. Then any upper layer  $CSG_i \in C_G$  can be represented as a graph  $(V_{C_i}, E_{C_i})$  where vertices  $V_{C_i} \in C_G \cup E_G \cup Reference (E_G)$  and the set of edges  $E_{C_i}$  represents the association amongst the vertices.

(c) *Annotation*: Annotation is a specialized form of CSG and can be expressed as  $G(V, E)$ , where  $|V| = I$  and  $E = \emptyset$ . Annotation can contain only text content as tagged value. Annotation can be containment in or associated with any other CSG. Further the cardinality constraint for Annotation construct is always 1:1 and ordering option can be 1 or 0, where 1 means content will be in orderly form with other constituent ESG and 0 means text content can be mixed with other constituent ESGs. The annotation construct will realize the document-centric semi-structured data possibly with mixed content. This concept is extremely important for mapping semi-structured data model in XML. Graphically Annotation can be expressed using *Square with Folded Corner*.

The summary of GOOSSDM constructs and their graphical notations have been given in Table I.

TABLE I. SUMMARY OF GOOSSDM CONSTRUCTS AND THEIR GRAPHICAL NOTATIONS

GOOSSDM Constructs	Description	Graphical Notation
ESG	Elementary Semantic Group	
Determinant ESG	Determinant vertex of any CSG which will determine the other member vertices in the CSG	
CSG	Contextual Semantic Group	
Annotation	Specialized form of CSG. Contain only Text Content.	
Association Connector	Connect multiple associated CSGs	

### B. Relationship Types in GOOSSDM

The proposed GOOSSDM provides a graph structure to represent semi-structured data. The edges of the graph represent relationships between or within the constructs of the model. In the proposed model, *four* types of edges have been used to represent different relationships. The type of edges and their corresponding meanings are as follows,

(a) *Containment*: Containments are defined between encapsulated ESGs including determinant ESG and parent CSG, or between two constituent CSGs and parent CSG, or

between CSG and referential constructs. The Containment relationship is constrained by the parameters tuple  $\langle p, \theta \rangle$ , where  $p$  determines the participation of instances in containment and  $\theta$  determines the ordering option of constituent ESGs or CSGs. With any CSG, this represents a bijective mapping between determinant ESG and other ESGs or composed CSG with participation constraint. Graphically association can be expressed using *Solid Directed Edge* from the constituent constructs to its parent labeled by constraint specifications. The possible values for  $p$  are as follows,

- (i) 1:1 – Represents ESG with mandatory one instantiation or total participation in the relationship and is default value of  $p$ .
- (ii) 0:1 – Represents ESG with optional one instantiation in the relationship.
- (iii) 1:M – Represents ESG with mandatory multiple instantiation in the relationship.
- (iv) 0:M – Represents ESG with optional multiple instantiation in the relationship.
- (v) 0:X – Represents ESGs with optional exclusive instantiation in the relationship. If a CSG contain single such ESG then it will act like 0:1 option. Otherwise one such ESG will optionally instantiate among all ESGs with  $p$  value 0:X.
- (vi) 1:X – Represents ESG with mandatory exclusive instantiation in the relationship. If a CSG contain single such ESG then it will act like 1:1 value option. Otherwise it is mandatory that one such ESG will instantiate among all ESGs with  $p$  value 1:X.

The possible values for  $\theta$  are as follows,

- (i) 1 – Represents that for any CSG, the constituent ESGs and CSGs are ordered and order must be maintained from left to right in the list of ESGs with  $\theta$  value 1.
- (ii) 0 – This is default value of  $\theta$  and represents that for any CSG, the constituent ESGs and CSGs are not ordered.

(b) *Association*: Associations are defined between related CSGs of same layer of adjacent layers. The Association relationship is constrained by the parameters tuple  $\langle P, \theta \rangle$ , where  $P$  determines the cardinality of Association and  $\theta$  determines the ordering option of associated CSGs. Graphically association can be expressed using *Solid Undirected Edge*. Any CSG wish to participate in association will be connected with association relationship. On next, multiple associated CSGs will be connected through *Association Connector*. For semi-structured it is sometime important to have specific context of some association. Such context can be represented using *Associated CSG* defined on *Association Connector*. Association Connector facilitates the  $n$  – array relationship. Graphically association can be expressed using *Solid Undirected Edge*, Association connector can be expressed using *Solid Diamond* and Associated CSG can be connected with Association Connector using *Dotted Undirected Edges* with Participation constraint specifications. The values for  $P$  can be 1:1 or 0:1 or 1:N or 0:N or 0:X or 1:X with corresponding meaning and possible values for  $\theta$  can be 1 or 0 with corresponding meaning.

(c) *Link*: Links are used to represent the inheritance relationships between two CSGs. Graphically link can be expressed using *Solid Directed Edge with Bold Head* from the generalized CSG to the specialized one.

(d) Reference: In semi-structured data model, it is important to represent the symmetric relationship between ESGs or CSGs. Reference can be used to model such concepts. Reference relations are defined either between ESG and referred ESG or between CSG and referred CSG. Graphically reference can be expressed using *Dotted Directed Edge*.

The summary of GOOSSDM relationship types and their graphical notations have been given in Table II.

TABLE II. SUMMARY OF GOOSSDM RELATIONSHIP TYPES AND THEIR GRAPHICAL NOTATIONS

GOOSSDM Relationships	Description	Graphical Notation
Containment	Defined between Parent CSG and constituent ESGs and CSGs	$\langle p, \theta \rangle$ 
Association	Defined between CSGs of same layer or adjacent layers.	$\langle P, \theta \rangle$ 
CSG Association	Defined between association and associated CSG	$\langle P, \theta \rangle$ 
Link	Defined between two adjacent layer parent CSG and inherited CSG	
Reference	Defined either between ESG and referred ESG or between CSG and referred CSG.	

### III. TRANSFORMATION OF GOOSSD INTO XSD

In general, the proposed GOOSSDM can be useful to realize the semi-structured data schema at conceptual level. The logical structure of such schema can be represented using the artifacts of XSD. Moreover, XSD is currently the de facto standard for describing XML documents. An XSD schema itself can be considered as an XML document. *Elements* are the main building block of any XML document. They contain the data and determine the elementary structures within the document. Otherwise, XSD also may contain sub-element, attributes, complex types, and simple types. XSD schema elements exhibit hierarchical structure with single root element.

A systematic rule based transformation of GOOSSD schemata to XSD is essential to express the semi-structured data at logical level more effectively. For the purpose, a set of rules have been proposed to generate the equivalent XSD from the semantic constructs and relationship types of a given GOOSSDM schemata. Based on the concepts of GOOSSDM constructs and relationship types the transformation rules are as follows,

**Rule 1:** An ESG will be expressed as an *element* in XSD. For example,  $ESG_{City}$  can be defined on attribute *City* to realize Customer city. Any DESG construct must be expressed with typed *ID* in XSD. The equivalent representation in XSD can be as given in Figure 1.

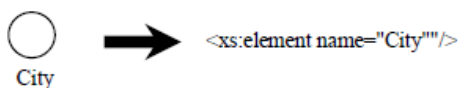


Figure1. Representation of ESG

**Rule 2:** A CSG will be expressed as a *complexType* in XSD. For example,  $CSG_{Customer}$  can be defined to realize the detail of

Customer. The equivalent representation in XSD can be as given in Figure 2.

**Rule 3:** Any Annotation construct will be expressed as a *complexType* in XSD with suitable *mixed* value. On containment to other CSG, if  $\theta$  value is 0 then it will be treated as mixed content in the resulted XML document. Otherwise if  $\theta$  value is 1 then it will be treated as annotation text in resultant XML document in orderly form. An example of annotation constructs has been shown in Figure 3.

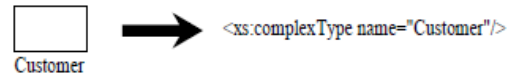


Figure 2. Representation of CSG

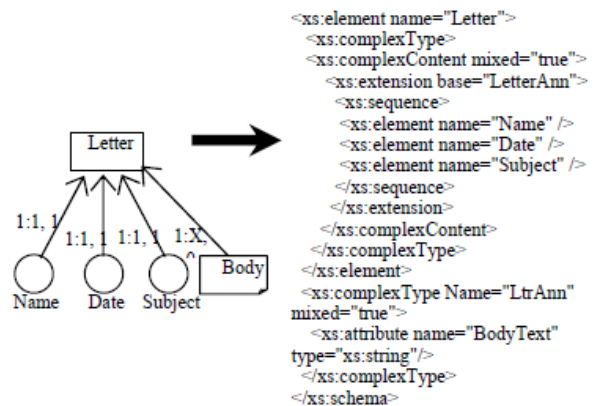


Figure 3. Representation of Annotation Construct

**Rule 4:** A Reference of ESG and CSG will be expressed as a *complexType* in XSD. For example, a reference of  $ESG_{City}$  can be defined on attribute *City* to realize a referential attribute on Customer city. The equivalent representation in XSD can be as given in Figure 4.

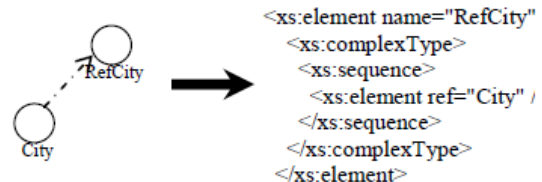


Figure 4. Representation of Reference of ESG.

**Rule 5:** CSGs of topmost layer will be treated as root in XSD declaration.

**Rule 6:** Any lowest layer CSG with containment of some ESGs will be expressed as a *complexType* with *elements* declaration in XSD. Further the participation constraint ( $p$  value in GOOSSDM concept) can be expressed using *minOccurs* and *maxOccurs* attribute in XSD. The ordering constraint ( $\theta$  value in GOOSSDM concept) can be expressed using compositor type of XSD. If  $\theta$  value is 1 then compositor type will be *sequence* otherwise *all*. For ordered set ESGs, the order will be maintained from *left to right*. If any subset of ESGs contains the  $p$  value  $X:1$  then those ESGs will be composite using *choice* compositor type in XSD. An example of XSD representation of lowest layer CSG has been shown in Figure 5.

**Rule 7:** Any upper layer CSG with containment of ESGs, reference of ESGs and adjacent lower layer CSGs will be

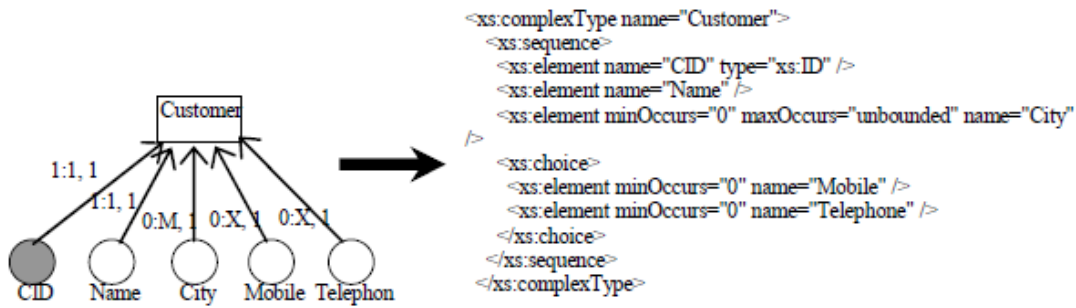


Figure 5. Representation of Lower layer CSG

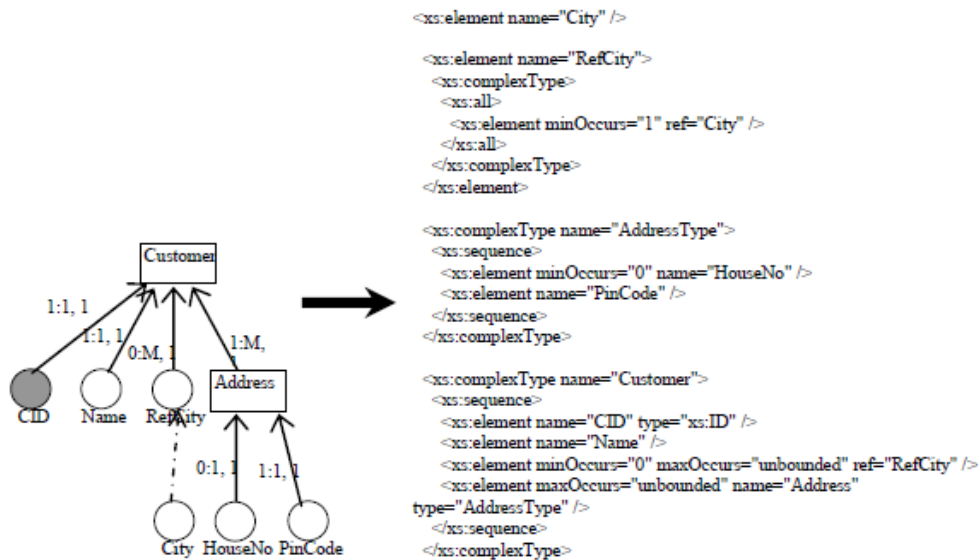


Figure 6. Representation of Upper layer CSG.

expressed as a *complexType* in XSD. An example of XSD representation of upper layer CSG with containment relation has been shown in Figure 6.

**Rule 8:** Any upper layer CSG with *Link* relationship with adjacent lower layer CSGs will be expressed as a *complexType* with inheritance in XSD. Upper layer CSG will be the child of lower layer CSG. An example of XSD representation of upper layer CSG with inheritance with adjacent lower layer CSG has been shown in Figure 7.

**Rule 9:** Any upper layer CSG with *Association* relationship with adjacent lower layer CSGs will be expressed as a *complexType* with nesting in XSD. Upper layer CSG will be treated as root element.

**Rule 10:** *Association* relationship between any two CSGs in the same layer will be expressed as a *complexType* with nesting in XSD. Rightmost CSG will be treated as the root element and on next nesting should be done in *right to left* order of the CSG in the same layer.

**Rule 11:** N – array *Association* relationship within a set of CSGs spread over several layer will be expressed as a *complexType* with nesting in XSD. Topmost layer CSG will be treated as the root element in XSD. Then, in the adjacent lower layer the rightmost CSG should be treated as nested element within the root element. Further the nesting should be done in

right to left order of the CSG in the same layer and on next moving on the adjacent lower layers.

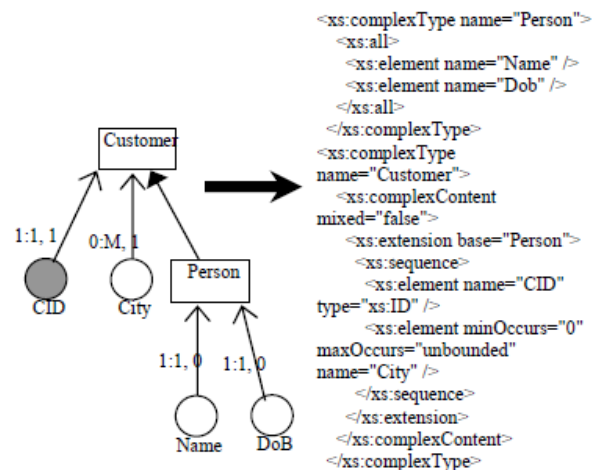


Figure 7. Representation of Link Relationship

**Rule 12:** With several *Association* relationships (composition of n – array and simple relationship) within a set of CSGs spread over several layer will be expressed as a *complexType* with nesting in XSD. Topmost layer CSG will be treated as the root element in XSD. Then, if available, the directly associated CSGs in each adjacent lower layer will be nested till it reaches



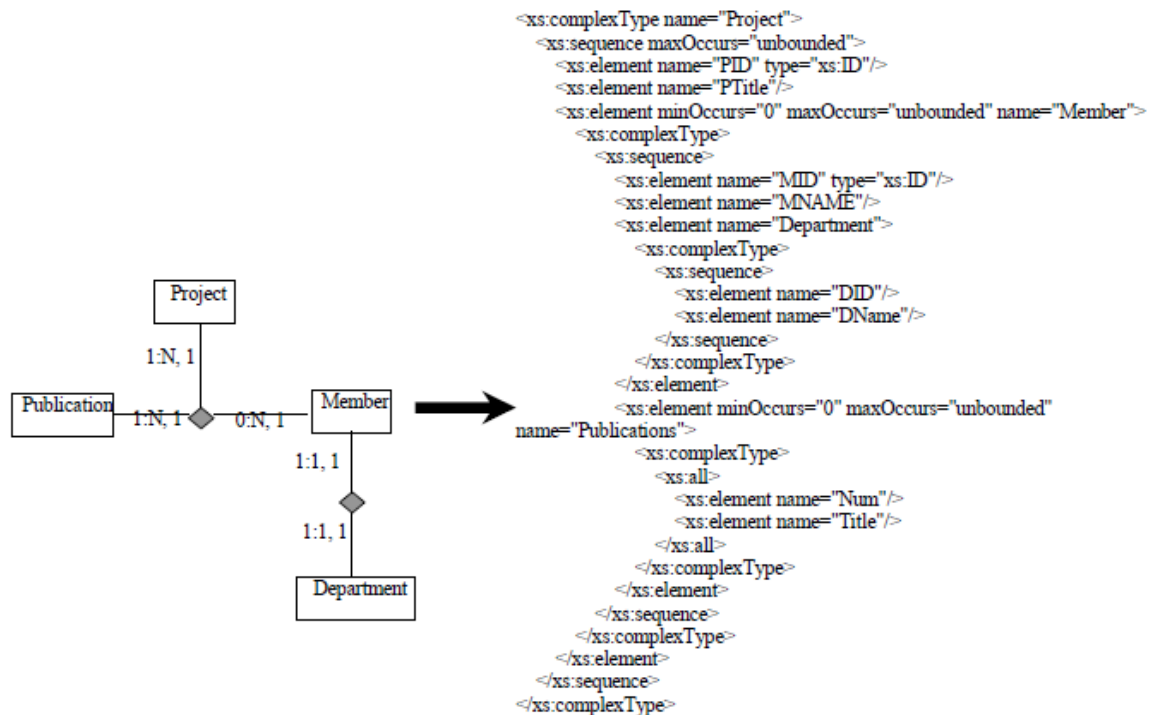


Figure 8. Representation of Associated CSGs spread over several layers (ESG layer is hidden).

to the lowermost layer of available associated CSG. On next, the CSGs of adjacent lower layer of the root element will be nested from right to left order in the same layer along with the nesting of directly associated CSGs (if available) in each corresponding adjacent layers. An example has been shown in Figure 9 for XSD representation of GOOSSDM schemata where associated CSGs are spread over three layers and contain both n – array and simple associations.

```

<Patient>
  <name>Patient 1</name>
  <Visit>
    <Date>10-JAN-2009</Date>
    <Doctor>
      <RegID>1234</RegID>
      <DName>Dr. P. Roy</DName>
    </Doctor>
    <Dept><DID>1</DID><DeptName>General</DeptName>
      <Hospital><Name>Hospital A</Name> </Hospital>
    </Dept>

    <Date>15-MAR-2010</Date>
    <Doctor>
      <RegID>4321</RegID>
      <DName>Dr. T. De</DName>
    </Doctor>
    <Clinic><Name>Clinic B</Name></Clinic>
  </Visit>

  <name>Patient 2</name>
  <Visit>
    <Date>12-SEPT-2009</Date>
    <Doctor>
      <RegID>4321</RegID>
      <DName>Dr. T. De</DName>
    </Doctor>
    <Clinic><Name>Clinic D</Name></Clinic>
  </Visit>
</Patient>
  
```

Figure 9. Irregular Structure in Visit Records XML

#### IV. CASE STUDY

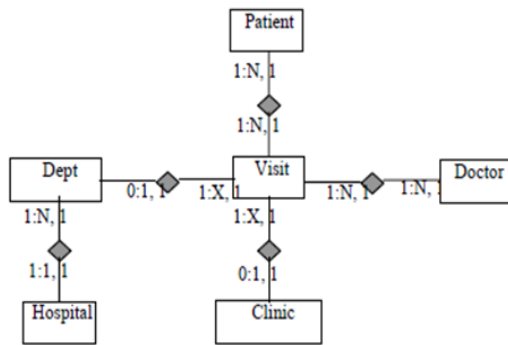
Let consider an example of *Visit Record* of *Patient* where a *Patient* can visit to a *Doctor* either at *Hospital Department* or at *Clinic* [7]. Any patient can visit several times to different doctors. The Figure 9 shows an irregularly structured XML representation of visit records of two patients. *Patient 1* visited twice to two different *Doctors*, one at *Hospital Department* and another at *Clinic*. *Patient 2* visited once to one *Doctor* common to *Patient 1* but at different *Clinic*. All though in the document, the *Date of Visit*, *Doctor* and option of *Hospital Department* and *Clinic* are in order. The XML document of Figure 9 represents the semi-structured data for such *Visit Record* database. The suitable GOOSSDM schemata for such data and its equivalent XSD have been shown in Figure 10. The equivalent XSD of GOOSSDM schemata of Figure 10 can be generated using the rules described in Section III.

#### V. CORRECTNESS OF GOOSSDM TRANSFORMATION

The set of proposed transformation rules described in Section III facilitates the systematic transformation of conceptual level semi-structured data model like GOOSSDM to the equivalent XSD in logical level. The correctness of the model transformation can be proved using the structural correspondence approach described in Narayanan *et al* [15]. In every model transformation, there is a correlation or correspondence between parts of the input model and parts of the output model. One can specify these correlations in terms of the abstract semantics of the source and target model constructs. The approach of Narayanan *et al.* describes that, if a transformation has resulted in the desired output models, there will be a verifiable structural correspondence between the source and target model instances that is decidable. Moreover, the transformation can be accepted as correct, if a node in the



source model and its corresponding node in the target model satisfy some correspondence conditions.



```

<xs:element name="Patient">
  <xs:complexType>
    <xs:sequence maxOccurs="unbounded">
      <xs:element name="name" />
      <xs:element maxOccurs="unbounded" name="Visit">
        <xs:complexType>
          <xs:sequence maxOccurs="unbounded">
            <xs:element name="Date" />
            <xs:element name="Doctor">
              <xs:complexType>
                <xs:sequence>
                  <xs:element name="RegID" />
                  <xs:element name="DName" />
                </xs:sequence>
              </xs:complexType>
            </xs:element>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:choice>
  <xs:element minOccurs="0" name="Dept">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="DID" type="xs:ID" />
        <xs:element name="DeptName" />
        <xs:element name="Hospital">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="Name" />
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element minOccurs="0" name="Clinic">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Name" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:choice>
</xs:complexType>
</xs:element>

```

Figure 10. Corresponding GOOSSDM schemata and Equivalent XSD of Figure 9

In case of GOOSSDM, the meta-model level identifiable correspondence structures are listed in Table III and the table can be treated as the *look-up table* for the cross links between the *source model* (GOOSSDM) and *target model* (XSD).

Further, the proposed set of rules will realize the correctness condition in model transformation. In our proposed approach, the correspondence rules must ensure that semantics and syntax for every constructs in the GOOSSDM model and its instance being transformed into the XSD model elements.

Several examples have been illustrated for the proposed transformation rules to verify the correctness of correspondence mapping of GOOSSDM schemata to the equivalent XSD.

TABLE III. LOOK-UP TABLE FOR STRUCTURAL CORRESPONDANCE

GOOSSDM Constructs	Graphical Notation	Equivalent XSD Representation
ESG	○	<i>xs:Element</i>
Determinant ESG	●	<i>xs:ID</i>
CSG	□	<i>xs:complexType</i>
Annotation	□	<i>xs:complexType</i> with suitable <i>Mixed</i> value
Association Connector	◆	<i>xs:complexType</i> with nesting from right to left order
Containment	<p, θ> →	<i>xs:complexType</i> with <i>element</i> declaration
Association	<P, θ> —	<i>xs:complexType</i> with nesting
CSG Association	<P, θ> - - - - -	<i>xs:complexType</i> with nesting from right to left order
Link	→	<i>xs:extension</i> declaration
Reference	- - - - ->	<i>ref</i> declaration in <i>xs:element</i>
P value	1:1 or 0:1 or 1:N or 0:N or 0:X or 1:X	<i>minOccurs</i> and <i>maxOccurs</i> delcarations
θ value	1 or 0	<i>Compositor</i> type : <i>all</i> or <i>sequence</i>

## VI. FEATURES OF GOOSSDM

The proposed GOOSSDM is an extension of comprehensive object oriented model for Semi-structured Database System and which can be viewed as a Graph (V, E) in layered organization. It contains set of semantically enriched constructs and relationship types to describe all the details those are necessary to specify the artifacts of the system containing semi-structured data. Moreover, using proposed set of rules, the proposed model schemata can be systematically transformed into equivalent XSD, which represents the logical schema for semi-structured data. Apart from these, one of the major advantages of the model is that it defines each level of structural detail on the constructs which are independent of the implementation issues. Moreover, the graph structure maintains the referential integrity inherently. The features of the proposed model are as follows,

(a) *Explicit Separation of structure and Content:* The model provides a unique design framework to specify the design for the semi-structured database system using semantic definitions of different levels (from elementary to composite) of data structure through graph. The model reveals a set of structures like ESG, CSG, Annotation, Association Connector etc. along with a set of relationships like Containment, Association, Link, Reference etc. between the structures, which are not instance based or value based. So, the nature of contents that corresponded with the instances and the functional constraint on the instances has been separated from the system's structural descriptions.

(b) *Abstraction:* In the proposed GOOSSDM, the concepts of layers deploy the abstraction in semi-structured data schema. The upper layer views will hide the detail structural complexity from the users. Such a representation is highly flexible for the user to understand the basic structure of semi-structure database system and to formulate the alternative design options.

(c) *Reuse Potential:* The proposed model is based on Object oriented paradigm. It is supported with inheritance mechanism using the *Link* relationship. Henceforth, there is no binding in the model to reuse some CSG constructs of any layer. On reuse of CSG, the specialized CSG must be shown in adjacent upper layer of the parent CSG. Moreover, lowest layer ESG or lower layer can be shared and reused with different CSGs of the upper layers using *Containment* relationship.

(d) *Disjunction Characteristic:* The instances of semi-structured data schema are likely to be less homogeneous than structured data. Disjunction relationships facilitate the possibility of non-homogeneous instances. The proposed GOOSSDM supports disjunction relationship using the participation constraint attribute  $p$  or  $P$  (by setting  $p$  or  $P$  value either  $0:X$  or  $1:X$ ). The Containment relationships between constituent ESGs or CSGs with the parent CSG can be disjunctive or Association relationships between two or more CSGs can be disjunctive. Figure 5 and Figure 10 respective explain such disjunctions.

(e) *Hierarchical and Non-hierarchical Structure:* The proposed model explicitly supports both hierarchical and non-hierarchical representation in semi-structure data modeling at conceptual level. Associated CSGs of different or same layers form the hierarchical or non-hierarchical structure in semi-structured data model. At the logical level modeling of semi-structured data using XSD supports only hierarchical structure. For the purpose, the set of rules have been proposed to transform more generous conceptual level schema to hierarchical logical schema.

(f) *Ordering:* Ordering is one important concept in modeling of semi-structured data. One or more attributes or relationships in semi-structured data schema can be ordered. Our proposed model supports ordering in two ways using the relationship ordering constraint attribute  $\theta$ . Firstly, the ordering may be enforced between parent CSG and any set of constituent ESGs and CSGs by specifying the  $\theta$  value on containment relationship. Secondly, the ordering can be enforced on the any set of Association relationships within CSG.

(g) *Irregular and Heterogeneous structure:* By characteristic the semi-structured data is irregular and heterogeneous. The proposed GOOSSDM supports disjunction characteristic, ordering and representation of both hierarchical and non-hierarchical structure in the same schema. With all these facets, the proposed model can efficiently model the irregular and heterogeneous semi-structured data. Modeling of irregular structure using GOOSSDM has been shown in Figure 10.

(h) *Participation constraint:* Instances participations in the semi-structured data schema are not followed strictly. Participations of instances can be optional or mandatory or even exclusive for such schema. This can affect the participation of constituent ESGs and CSGs in the parent CSG or may affect the participation of CSGs in some association relationship either of simply type or n-array type. All these participation constraint can be modeled in proposed GOOSSDM by specifying the value for participation constraint attribute  $p$  or  $P$ .

(i) *Document-centric and Mixed Content:* In real world, document texts are mixed with semi-structured data. The feature is more important and frequent in XML documents. Thus it is an essence that, the conceptual model for semi-structured data must support modeling of such feature. In the proposed model, the *Annotation* construct facilitates to model document centric design of semi-structured data at conceptual level. Moreover, the modeling of the *Annotation* construct in the GOOSSDM schema allows the instances of CSG and ESG to be mixed with the text content. The presence of this construct along with the other defined constructs and relationships, the proposed GOOSSDM is also capable to model XML document at conceptual level.

## VII. IMPLEMENTATION OF GOOSSDM USING GME

The Generic Modeling Environment (GME) provides meta-modeling capabilities and where a domain model can be configured and adapted from meta-level specifications (representing the Conceptual modeling) that describe the domain concept. It is common for a model in the GME to contain several numbers of different modeling elements with hierarchies that can be in many levels deep. The GME supports the concept of a viewpoint as a first-class modeling construct, which describes a partitioning that selects a subset of conceptual modeling components as being visible.

Moreover, GME support the programmatic access of the metadata of GME models. Most usual techniques for such programmatic access is to write GME interpreter for some metamodel. The interpreter will be able to interpret any domain model based on that predefined metamodel. GME interpreters are not standalone programs, they are components (usually Dynamic Link Libraries) that are loaded and executed by GME upon a user's request. Most GME components are built for the Builder Object Network (BON), an inbuilt framework in GME and provide a network of C++ objects. Each of these represents an object in the GME model database. C++ methods provide convenient read/write access to the objects' properties, attributes, and relations described in GME metamodel.

In the context of GOOSSDM, the lower layers can be conceptualized using levels in GME. The semi-structured data



## VIII. CONCLUSION

In this paper, a model has been introduced for the conceptual level design of semi-structured data using graph based semantics. This is a comprehensive object oriented conceptual model and the entire semi-structure database can be viewed as a Graph (V, E) in layered organization. The graph based semantics in GOOSSDM model extracts the positive features of both Object and Relational data models and also it maintains the referential integrity inherently. Further the layered organization of the model facilitates to view the semi-structured data schema from different level of abstraction.

The proposed GOOSSDM contains detailed set of semantically enriched constructs and relationships those are necessary to specify the facets of semi-structured database system at conceptual level. Moreover, a set of rules also have been proposed to systematically transform any GOOSSDM schema to its equivalent XSD structure. The expressive powers of the set of transformation rules have been illustrated with suitable examples and case study. Moreover, the proposed model also facilitates the designer to provide alternative design of same schema by changing the ordering scheme, which in result can be transformed in different XSDs with different nesting patterns. It provides better understandability to the users and high flexibility to the designers for creation and / or modification of semi-structured data as well as XML document at conceptual level. The proposed approach is also independent from any implementation issues.

It is also important to note that with the concept of Annotation construct the proposed GOOSSDM facilitate the document – centric design of semi-structured data at conceptual level. Also the proposed model supports irregular, heterogeneous, hierarchical and no-hierarchical structure in data. Moreover, the set of proposed rules are capable to transform systematically the GOOSSDM schema into hierarchical XSD schema. Due to these features, the proposed approach is also capable to design XML document at conceptual level.

The proposed approach also has been automated through the GME based meta-model configuration of GOOSSDM. The meta-level specification of GOOSSDM along with interpreter can be used as a CASE tool for the model by the semi-structured database designer. The tools facilitates the automatic generation of XML Schema Definitions from the conceptual level graphical model, using the set of proposed rule set.

Future studies will concentrate on developing a graphical query language for the proposed approach.

## REFERENCES

[1] S. Abiteboul, P. Buneman, and D. Suciu, "Data on the Web: From Relations to Semistructured Data and XML", Morgan Kaufman, 1999.

- [2] Jason McHugh, Serge Abiteboul, Roy Goldman, Dallas Quass, Jennifer Widom "Lore: a database management system for semistructured data", Vol. 26, Issue 3, PP: 54 - 66, 1997
- [3] A. Badia, "Conceptual modeling for semistructured data", In Proc. of the Third International Conference on Web Information Systems Engineering, PP: 170 – 177, 2002.
- [4] M. Mani. "EReX: A Conceptual Model for XML", In Proc. of the Second International XML Database Symposium, PP 128-142, 2004.
- [5] G. Psaila, "ERX: A Conceptual Model for XML Documents", In Proc. of the ACM Symposium on Applied Computing, PP: 898-903, 2000.
- [6] A. Sengupta, S. Mohan, R. Doshi, "XER - Extensible Entity Relationship Modeling", In Proc. of the XML 2003 Conference, PP: 140-154, 2003.
- [7] Martin Necasky, "XSEM: a conceptual model for XML", In Proc. of 4<sup>th</sup> ACM International Asia-Pacific conference on Conceptual Modeling, Vol. 67, PP: 37 - 48, 2007.
- [8] B. F. Lócio, A. C. Salgado, L. R. Galvão, "Conceptual modeling of XML schemas", In Proc. of the 5<sup>th</sup> ACM International Workshop on Web Information and Data Management, PP: 102 – 105, 2003.
- [9] H. X. Liu, Y. S. Lu, Qing Yang, "XML conceptual modeling with XUML", In Proc. of the 28<sup>th</sup> International Conference On Software Engineering, PP: 973 – 976, 2006.
- [10] C. Combi, B. Oliboni, "Conceptual modeling of XML data", In Proc. of the ACM Symposium On Applied Computing, PP: 467 – 473, 2006.
- [11] X. Wu, T. W. Ling, M. L. Lee, G. Dobbie, "Designing semistructured databases using ORA-SS model", In Proc. of the 2<sup>nd</sup> International Conference on Web Information Systems Engineering, Vol. 1, PP: 171 – 180, 2001.
- [12] R. Conrad, D. Scheffner, J. C. Freytag, "XML conceptual modeling using UML", In Proc. of the 19<sup>th</sup> International Conference On Conceptual Modeling, PP: 558-574, 2000.
- [13] M. Necasky, "Conceptual Modeling for XML: A Survey", Tech. Report No. 2006-3, Dep. of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, 2006.
- [14] Á. Lédeczi, A. Bakay, M. Maroti, P. Volgyesi, G. Nordstrom, J. Sprinkle, G. Karsai, "Composing Domain-Specific Design Environments", IEEE Computer, pp. 44-51, November 2001.
- [15] A. Narayanan, G. Karsai, "Specifying the correctness properties of model transformations", Proc. of 3<sup>rd</sup> Int. workshop on Graph and model transformations (Int. Conf. on Software Engineering), PP 45-52, 2008.
- [16] Abhijit Sanyal, Anirban Sarkar, Sankhayan Choudhury, "Automating Web Data Model: Conceptual Design to Logical Representation", 19<sup>th</sup> Intl. Conf. on Software Engineering and Data Engineering (SEDE 2010), PP 94 – 99, 2010.
- [17] Anirban Sarkar, Sesa Singha Roy, "Graph Semantic Based Conceptual Model of Semi-structured Data: An Object Oriented Approach", 11<sup>th</sup> International Conference on Software Engineering Research and Practice (SERP 11), Vol. 1, PP 24 – 30, USA, July 18 – 21, 2011.

## AUTHORS PROFILE



**Anirban Sarkar** is presently a faculty member in the Department of Computer Applications, National Institute of Technology, Durgapur, India. He received his PhD degree from National Institute of Technology, Durgapur, India in 2010. His areas of research interests are Database Systems and Software Engineering. His total numbers of publications in various international platforms are about 25.