IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 3 Issue 8

August 2012

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

SAI

www.ijacsa.thesai.org


# INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS


**THE SCIENCE AND INFORMATION ORGANIZATION**

www.thesai.org | info@thesai.org












# Editorial Preface

*From the Desk of Managing Editor…*

IJACSA seems to have a cult following and was a humungous success during 2011. We at The Science and Information Organization are pleased to present the August 2012 Issue of IJACSA.

While it took the radio 38 years and the television a short 13 years, it took the World Wide Web only 4 years to reach 50 million users. This shows the richness of the pace at which the computer science moves. As 2012 progresses, we seem to be set for the rapid and intricate ramifications of new technology advancements.

With this issue we wish to reach out to a much larger number with an expectation that more and more researchers get interested in our mission of sharing wisdom. The Organization is committed to introduce to the research audience exactly what they are looking for and that is unique and novel. Guided by this mission, we continuously look for ways to collaborate with other educational institutions worldwide.

Well, as Steve Jobs once said, Innovation has nothing to do with how many R&D dollars you have, it's about the people you have. At IJACSA we believe in spreading the 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.

Throughout our archives, new ideas and technologies have been welcomed, carefully critiqued, and discarded or accepted by qualified reviewers and associate editors. Our efforts to improve the quality of the articles published and expand their reach to the interested audience will continue, and these efforts will require critical minds and careful consideration to assess the quality, relevance, and readability of individual articles.

To summarise, the journal has offered its readership thought provoking theoretical, philosophical, and empirical ideas from some of the finest minds worldwide. We thank all our readers for their continued support and goodwill for IJACSA. We will keep you posted on updates about the new programmes launched in collaboration.

We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJACSA provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

We regularly conduct surveys and receive extensive feedback which we take very seriously. We beseech valuable suggestions of all our readers for improving our publication.

**Thank you for Sharing Wisdom!**

**Managing Editor**
**IJACSA**
**Volume 3 Issue 8 August 2012**
**ISSN 2156-5570 (Online)**
**ISSN 2158-107X (Print)**

# Editorial Board

# Reviewer Board Members

- **A Kathirvel**
  Karpaga Vinayaka College of Engineering and Technology, India
- **A.V. Senthil Kumar**
  Hindusthan College of Arts and Science
- **Abbas Karimi**
  I.A.U_Arak Branch (Faculty Member) & Universiti Putra Malaysia
- **Abdel-Hameed A. Badawy**
  University of Maryland
- **Abdul Wahid**
  Gautam Buddha University
- **Abdul Hannan**
  Vivekanand College
- **Abdul Khader Jilani Saudagar**
  Al-Imam Muhammad Ibn Saud Islamic University
- **Abdur Rashid Khan**
  Gomal Unversity
- **Aderemi A. Atayero**
  Covenant University
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
  Menoufia University, Egypt
- **Ajantha Herath**
  University of Fiji
- **Ahmed Sabah AL-Jumaili**
  Ahlia University
- **Akbar Hossain**
- **Albert Alexander**
  Kongu Engineering College,India
- **Prof. Alcinia Zita Sampaio**
  Technical University of Lisbon
- **Amit Verma**
  Rayat & Bahra Engineering College, India
- **Ammar Mohammed Ammar**
  Department of Computer Science, University of Koblenz-Landau
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Anirban Sarkar**
  National Institute of Technology, Durgapur, India
- **Arash Habibi Lashakri**
  University Technology Malaysia (UTM), Malaysia
- **Aris Skander**
  Constantine University
- **Ashraf Mohammed Iqbal**
  Dalhousie University and Capital Health
- **Asoke Nath**
  St. Xaviers College, India
- **Aung Kyaw Oo**
  Defence Services Academy
- **B R SARATH KUMAR**
  Lenora College of Engineering, India
- **Babatunde Opeoluwa Akinkunmi**
  University of Ibadan
- **Badre Bossoufi**
  University of Liege
- **Balakrushna Tripathy**
  VIT University
- **Basil Hamed**
  Islamic University of Gaza
- **Bharat Bhushan Agarwal**
  I.F.T.M.UNIVERSITY
- **Bharti Waman Gawali**
  Department of Computer Science & information
- **Bremananth Ramachandran**
  School of EEE, Nanyang Technological University
- **Brij Gupta**
  University of New Brunswick
- **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**
  Chhattisgarh Swami Vivekananda Technical University
- **Chi-Hua Chen**
  National Chiao-Tung University
- **Constantin POPESCU**
  Department of Mathematics and Computer Science, University of Oradea
- **Prof. D. S. R. Murthy**
  SNIST, India.
- **Dana PETCU**
  West University of Timisoara
- **David Greenhalgh**
  University of Strathclyde

# CONTENTS

# Instruction Design Model for Self-Paced ICT System E-Learning in an Organization

Ridi Ferdiana
Electrical Engineering and IT Department
Universitas Gadjah Mada
Yogyakarta, Indonesia

Obert Hoseanto
Partners in Learning Department
Microsoft Indonesia
Jakarta, Indonesia

***Abstract*— Adopting an Information Communication and Technology (ICT) system in an organization is somewhat challenging. User diversity, heavy workload, and different skill gap make the ICT adoption process slower. This research starts from a condition that a conventional ICT learning through short workshop and guidance book is not working well. This research proposes a model called ICT instruction design model (ICT-IDM). This model provides fast track learning through integration between multimedia learning and self-paced hands-on E-learning. Through this case study, we discovered that the proposed model provides 27% rapid learning adoption rather than conventional learning model.**

***Keywords- ICT System Adoption; Learning Model; Multimedia Learning; E-learning; Instruction Design Model; Learning Plan.***

## I. INTRODUCTION

Good ICT system is not only about good software or a good hardware but it also need skilled users. After an ICT system is developed, it needs additional time for the system to be used by the user. The implementation phase in software engineering should make sure that the user feels comfortable to use the system.

The implementation phase in ICT system is done by doing several socialization activities such as training, hands-on workshop, coaching or even giving a grant for the users who use the system correctly. Tsui and Karam [16] mention several characteristics of good ICT system implementation which are:

- Readability: The software can be easily read and understood by the programmers.
- Maintainability: The code can be easily modified and maintained.
- Performance: All other things being equal, the implementation should produce code that performs as fast as possible.
- Traceability: All code elements should correspond to a design element. Code can be traced back to design.
- Correctness: The implementation should do what it is intended to do.
- Completeness: All of the system requirements are met.

However, these good implementation characteristics are only eligible when the users are digital literate and can use the system well. If the users lack digital literacy or unable to use the system, the good implementation only happened within the software team and not in the user's perspective.

An organization invests an ICT system to be used by the employees. It will not be a matter if the user of the ICT system is only a few (less than 10 users). However, the problem arises when the ICT system should be adopted for tens or hundreds of users.

This research starts from a real implementation problem in an organization called Alpha. Alpha is a government organization that works to enhance mathematic educators in Indonesia. Alpha implements an ICT collaboration system called Live@edu. Live@edu is a collaboration system developed by Microsoft for education institutions. It contains software which is:

- Email software.
- Online storage software.
- Instant messaging software.
- Personal Information Management software.
- Office Productivity software.

The software has more than five features. According to the Live@edu services page at http://bit.ly/liveateduservices , it shows that the services have 21 primary features. Additionally, half of the features are essential for business solution in the organization. These essential features should be acknowledged and used by the 50 users within two weeks.

The implementation team has done socialization activities through face-to-face discussion, full-day workshop, and delivered a user manual. The socialization activities received a good response with satisfaction index 8.5 (of scale 10). However, the organization management did not see a usage progress of the system and found several issues which are:

- User manual exists but more than 93% people does not read the manual or try by themself.
- User feels comfortable with the legacy system or without the system at all.

- User has little reason to use the system since they are too busy with the others job.
- Different user skill makes the adoption sluggish for the entire organization.
- User has not much time to learn and to explore the system when they are inside of working hours.

Based on these issues, this research will make an effort to improve the ICT system adoption by doing a several activities such as:

- Engage the user personally using self-paced multimedia learning.
- Creating an Instruction design model (IDM) for the organization.
- Creating several learning plan for several different people based in their existing skill.

In this research, we found that our proposed model can improve the learning curve of the user. It shows that the Live@edu usages increase by 27 % after we applied the Instructions Design Model (IDM) through multimedia learning. In the next Section, we will discuss about the previous researches that related with our findings.

## I. PREVIOUS RESEARCHES

Carliner [5] proposes an instructions design model (IDM) process that contains several phases that are definition phase, design phase, development phase, and implementation phase. These phases are designed to ensure that the e-learning is well tested and can be used independently. Table I shows the phase purposes.

TABLE I. IDM PURPOSES

| Phase | Purposes |
|---|---|
| Definition | • Conducting a need analysis<br>• Settings goal<br>• Preparing the need of analysis report |
| Design | • Choosing the form of E-learning<br>• Developing the learning strategy<br>• Establishing guidelines |
| Development | • Drafting the E-learning program<br>• Receiving feedback of E-learning<br>• Revising the E-learning |
| Implementation and Maintenance | • Producing and distributing the E-learning program<br>• Promoting the E-learning<br>• Maintaining the E-learning program |

In Carliner research, it is found that the IDM provides a sufficient guidance to build generic E-learning system. Generic E-learning is usually used for schools and universities. It contains curriculum, course contents, practices, and course evaluation and student profile. However, the IDM model is rarely used in an organization. This is because the company needs is different with the needs of the academic world. Company focused e-learning as a tool to increase not only the knowledge but also their productivity.

An implementation of IDM model in organization is done by O'Brien and Hall [8] by constructing a model called Training Need Analysis (TNA). It is shown that before an organization creates the e-learning they should enable companies to identify areas where their employees require training. The research provides generic TNA tools that are dedicated for SMEs allowing them to identify training requirements and assisting them to specify their own e-learning content in a structure. The tool implicitly shows that to create an effective e-learning, it needs to align with the organization need through modeling rule analysis and design e-learning content.

The research about modeling rule analysis and design in E-learning content is done by Kim and Choi [10]. In the research, it shows that the usage a SCORM model as content aggregation and design model that can be the foundation to design and develop learning management system and contents.

The development of learning management and e-learning focused in several initiatives that are related in content, type of e-learning and deliverable model. Table II display the previous researches that related with E-learning development that related with an organization.

TABLE II. RELEVANT RESEARCH IN E-LEARNING DEVELOPMENT IN AN ORGANIZATION

| Authors | Research Topic |
|---|---|
| Teo and Gay [6] | A prototype system that performs a subset of functions (learner profiling, knowledge visualization, and learning route mapping) is being developed |
| Stephenson, et al [9] | Building a specific e-learning for an organization called KaryoLab. KaryoLab contains background, tutorial, practice, and assessment |
| Schiaffino, et al [14] | an adaptive hypermedia system that uses the adaptability concept with the aim of providing the same content to different students groups |
| Ferretti, et al [13] | Building an E-learning system called We-Lcome. We-LCoME aim is to mashup compound multimedia potentials with the so called "collective intelligence" which the new Web 2.0 has revealed. Final e-learning media by using SMIL |
| Moller, et al [7] | An integrative concept for information and communication technology (ICT) supported education in modeling and simulation (M&S). The implementation of the M&S program uses ICT as an indispensable part of the modern education system. |

Based on the previous researches that described in Table II, it shows that the development E-learning in an organization should focus on:

- The specific need of the user in an organization. The e-learning initiative is done by creating custom application that delivers specific learning experiences.
- Building the system with multimedia and collaboration standard such as web 2.0 or SMIL.
- Profiling the user based on their skill, knowledge gap, and learning plan.

The last step on the IDM model is the implementation and maintenance phase of E-learning, also called as an adoption phase. This phase is critical path of the E-learning adoption. Table III Shows The Related Research That Focuses In E-Learning Adoption.

TABLE III. RELEVANT RESEARCH IN E-LEARNING ADOPTION

| Authors | Research Topic |
|---|---|
| Mesomela and Villiers [12] | The study that covers usability evaluation or an examination of the user interface of an e-learning product and usage analysis of an e-learning application designed to support learning in a cognitive domain |
| Luojus and Vilkki [13] | The pedagogical starting point for developing instruction in digital media was to produce new competence. The outcome of the development work was a teaching model that follows the user-driven design process, with the aim of providing students with the ability to act as developers of product development and innovation processes in their fields |
| Bang [19] | Integrate Multimedia Digital AP Server to provide immediate self-help and strengthening materials to students so that students can focus on the part they needed to be strengthened among extensive pool of books, thus helping learners to enhance the fun and efficiency of learning |
| Mehlenbacher, et al [3] | The research reviews 300 e-learning journal and one of the related point of view is about how Web Based Interface provide sufficient interface to learn effectively |

Based on the research that described in Table III, it shows that the adoption of the E-learning depends on:

- The timeframe of the adoption. It is shown that the quick adoption make the participant should learn more intensively. The technology such as web provides a flexible way to learn.
- The usage of a skill that delivered in E-learning. It is shown that a skill that related with the daily work of an organization lead to a quick adoption on the organization E-learning.
- Contextual and learning plan selection. Several researches show that effective e learning need align with content design based on the user need.

It is found that the implementation E-learning for an organization to learn something or to adopt anything is different with the implementation E-learning in the university or academic learning. Based on the previous researches, it shows that the E-learning in organization should have unique approach to achieve the benefits.

## II. THEORETICAL BACKGROUND

As mentioned before, the purposes of the research is to create a learning plan for any organization who wants to adopt ICT system through E-learning model. Based on previous researches, it shows various approaches to learn and adopt E-learning in an organization. In this research, it is chosen that the research will use the IDM model. The IDM model is chosen because:

- IDM model is mature enough and has been proven in academic and industry. [1] [4] [5] [11]
- IDM model provides complete engineering process for E-learning lifecycle.
- IDM can be pattern matched with a software development lifecycle (SDLC). It has similar phases such as requirement, analysis-design, development, and deployment [17]. Table IV shows the pattern matching between IDM phase and SDLC.

TABLE IV. IDM AND SDLC PATTERN MATCHING

| IDM Phases | SDLC Phases (V&V) | Similarity |
|---|---|---|
| Definition | Requirements Engineering | Providing a detail step to refine the purposes, to select the technology, to collect the need of a system |
| Design | Design | Providing a several action to design the system |
| Development | Development | Modification, construct, and installing the system |
| Implementation and Maintenance | Testing and Implementation | User acceptance testing and system socialization. |

Table IV shows that both IDM and SDLC has several similarities. The similarity leads the research to create a research phases that combined both phases. The great combination focuses in the last phase that is implementation.

In the implementation phase, the SDLC phase focuses in User Acceptance Test (UAT). UAT in software engineering provides several activities that dedicated for the properness of the created system with the user. The proposed system should meet its business requirements and to provide confidence that the system works correctly and is usable before it is formally "delivered" [18]. UAT should also address the testing of system documentation.

The IDM also bring the UAT in the first class testing with addition system socialization. When the proposed system meets the business need, the next challenge is to adopt the system to the existing environment. In this step, IDM provides step to promote E-learning. Based on that fact, we strongly believe that the IDM and SDLC give sufficient lifecycle to adopt ICT system. Both phases provide complete lifecycle in ICT development and adoption. In this research, we focus in the adoption only because several reasons such as:

- The system is already built by a software vendor. Therefore, it has no need to do full lifecycle step in SDLC. It just needs focus in IDM phase.
- The system is a common system for ICT Communication. In this research, the system is a full suite of communication platform such as email, online storage, and PIM software. Therefore, it will no need special training or long workshop.
- The adopted system has a previous similar system. Therefore, some user feels comfortable with the

existing system and need a good reason to use the new system.

Based on that reasons, it will start the research by creating adoption plan, constructing the learning model, and evaluating the learning model.

## III. ADOPTION PLANNING

The adoption plan starts with the implementation of the system. Since, the ICT system is Software as a Services solution the implementation simply as activating the services, enrolling the user and testing the services availability. Figure 1 show the research phase that designed based on the combination of both.

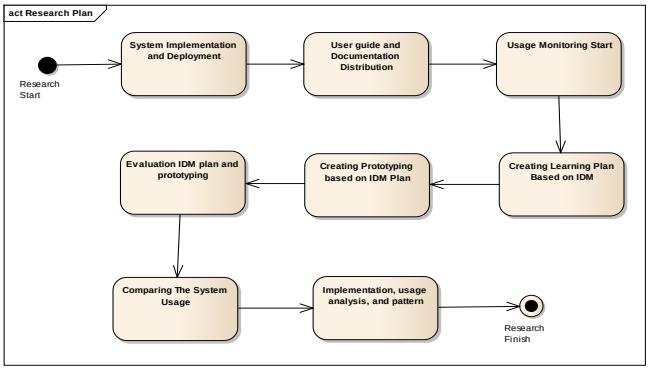



Figure 1. Research Phase

After doing implementation in the new system, the research starts from conventional adoption strategies which are the distribution of documentation and guidance. At this point, the organization stakeholder creates a memo to the employees to use the system. The memo enriches with manual and user guide for the system usage. In this case, the research prepares the instrumentation system to evaluate the usage of the system. Fortunately, the proposed system has out of the box-reporting feature called account trend. Account trend covers provisioned and active account. Provisioned account is the numbers of enrollment account. Active account is the numbers of active user that use the services. By dividing the active account with the provisioned account the research declares the active usage index. The active usage index will have a value between 0.00 – 1.00. The closer to the 1.00 the better value to the for active usage index. Figure 2 shows the reporting features in the proposed system.

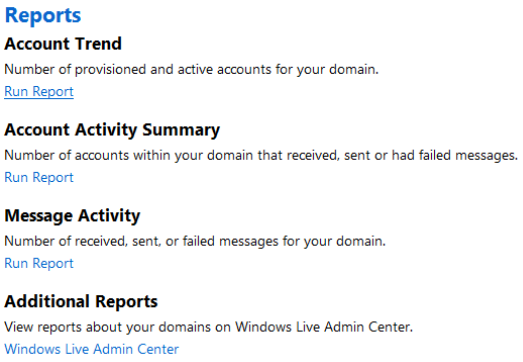



Figure 2. Active ID Excerpt Sample

The active usage index will be the main evidence for the research result. The active usage index will be calculated manually on this research. The active usage index will be measured before and after the treatment. As mentioned before, the treatment step is started by creating a learning model that described in Section V. The learning model implementation report will be discussed in Section VI.

## IV. CONTSRUCTING A LEARNING MODEL

In this research, it is assumed that the learning model is an approach to deliver good e-learning experience. It covers behavior of user, content learning plan, and socialization technique. With a scope that the main focus of the training is quickly adopt a new ICT system. The proposed learning model is constructed by doing several activities such as:

- Observing to the organization and see what they use and like in the existing system.
- Meeting with the stakeholder what they think and they hope with the new system.
- Classifying the learning item based on the need of the organization
- Creating the learning plan based on the system features classification.
- Creating learning content based on the learning plan.
- Distributing learning content based on the discussion within the stakeholder.

The observation activity covers a set of action that engages between implementation team with the client. In this step, it is done by using formal discussion and quick pool. Formal discussion did a quick observation about daily activity of the employee organization. Fifty employees join the session. The research does patterns matching between their daily activity with the proposed ICT system. Table V shows the system features and the usage scenarios that related with the organization business process. The system features is ranked by the result of quick pool.

TABLE V. SYSTEM FEATURES AND USAGE SCENARIOS PATTERNS MATCHING

| System Features | Usage Scenarios | Priority |
|---|---|---|
| Email | Business Communication, Internal memo, business letter | High |
| Calendar | internal meeting schedule | High |
| Task | Assignment memo from supervisor to emplooye | Medium |
| Online Storage | Storing digital multimedia content such as document and e-learning content | Medium |
| Public Address Book | Organization contact repository | Medium |
| Instant Messaging | Quick chat | Low |

The priority is concluded from a quick pool in the discussion session. Each participant will select the priority (high medium and low). If 70% participants select the same priority, the feature will have the selected priority. It is shown that email and calendar are the most demanded features in the organization. The result of the quick pool and discussion is

summarized and presented to the stakeholders in stakeholder gathering session.

Stakeholder gathering session establishes a shared vision between implementation team and stakeholder. In this research, the stakeholder gathering sessions discuss several items that are:

- An initial report that is obtained from the observation session.
- A feedback sharing session that is come from the stakeholder about the new system.
- The current activity and problem that might be solved by the new system
- Short discussion about the challenge that might be or already happen when adopting the new system.
- The milestone plan that will be adopted to implement new system.

Table VI summarizes the result of the gathering session with the stakeholder. It shows the problems, approaches, and the milestone when the problems and approaches is executed.

TABLE VI. MILESTONE, PROBLEM, AND APPROACH PLANNING

| Milestone | Problem Addressed | Approaches |
|---|---|---|
| Piloting Phase | Not all full time employees already enrolled to the system | Mass registration to the system |
| | Employees do test drive the new system without manual | User guide and quick reference card is distributed using online storage channel |
| | The user guide and quick manual won't be read during the business of the employee | Building a short tutorial rather than user guide |
| | Employees are confused to import their existing work into the new system | Building a tutorial how to migrate the current work into the new system |
| Adopting Phase | The new system adoption should be also adopted by the client and customer. | Enrolling and distributing key access of the system |
| | The impossibility to do socialization and workshop for more than 1000 clients and customers | Building an e-learning portal that can be accessed online |

Table VI shows that the learning model should covers three main approaches for learning which are building short tutorial for daily usage, building short tutorial for migrating a system, and building a learning portal for remote consumer. The three activities contents are structured by the team in an internal meeting.

The internal meeting classifies the main structure of the leaning model. The structure is composed by considering several items which are:

- The employees have several information assets in the existing system. Therefore, the first learning structure should cover a migration strategies and guidelines.
- The employees should know the main part of the application that related with their daily business activity. Therefore, after the migration the employee should learn the system features based on priority that already described in Table V.
- The remote users such as client and consumer should learn how to use e-learning portal that will be deployed as an approach to make the new system could be used by the entire organization.

Figure 3 shows the learning plan that will be used as a framework for the learning content. Migration learning focuses to cover any actions that should be done to migrate from existing to new system. Essential learning phase covers the essential topic that have to be mastered to use the system. Enhancement learning phase is a continuous phase that will be updated regularly. The enhancement learning phase will cover additional topic that not covered by the essential learning phase. In this step, the implementation learning phase will be executed in three months.

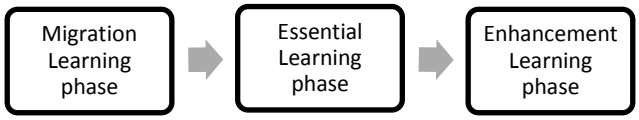



Figure 3. Learning Phases For Implementing ICT System

Each phase will have several contents based on the discussion within stakeholder and the 50 main users. Table VII describes the learning content in every phase.

TABLE VII. MILESTONE, PROBLEM, AND APPROACH PLANNING

| Learning Phase | Learning Topic |
|---|---|
| Migration Learning | • Migrating the address book of the existing system<br>• Archiving the old email and conversation<br>• Archiving and converting the document to prepare the online storage |
| Essential Learning | • Basic email configuring with Outlook Web Access (OWA)<br>• Downloading and managing email through Outlook client<br>• PIM management using OWA<br>• Instant messaging in OWA<br>• Office productivity using Office Web Apps<br>• Online storage and archiving with SkyDrive |
| Enhancement Learning | • Accessing an email in a limited connection<br>• Sharing and managing the online storage in SkyDrive<br>• Accessing and connecting other system through OWA<br>• Others topic will be discussed on demand |

After the creation of the learning plan, implementation team to create an ad-hoc team that focuses to develop the learning content. The team is composed three persons namely instruction design mentor and two talents. The instruction design mentor creates the tutorial scenarios. The tutorial itself works as a demo driven tutorial. After the scenario is created, the design mentor will propose the scenario to be recorded by the talents. The tutorial output is composed as three main outputs that are slide deck presentation, recorded video, and

demo script document. The outputs will be uploaded to the E-learning portal. The implementation of the E-learning content and the socialization of e-learning portal will be described in Section VI.

## V. LEARNING MODEL IMPLEMENTATION

The learning model implementation is aligned with the ICT implementation phase. The implementation model is done through two phases which are internal phase and external phase. Internal phase focuses in implementation activity in internal organization. It contains 100 employees that work permanently onsite in the organization. The external phase focuses implementation in external organization such as customer, alumni, community, and remote workers. It reaches 1500 persons that are separated geographically around Java, Sumatra, and Kalimantan.

Learning implementation in internal phase is done through blended learning. The blended learning contains an onsite workshop and self-paced online learning. Onsite workshop contains any material that related with essential learning that described in Table VII. On the other hand, the online learning is done by using organization online storage that stored in the new system. The online learning works as follow:

- The learning contents are uploaded into an online storage that parts of the new system. There are seven video lessons that uploaded and ready to view as self-paced e-learning.
- The learning contents link is distributed by the ICT supervisor in the organization. The link is distributed through a new email system.
- The learning content has a playlist as a step-by-step recommendation to learn the essential features of the system. Figure 4 shows the video playlist.

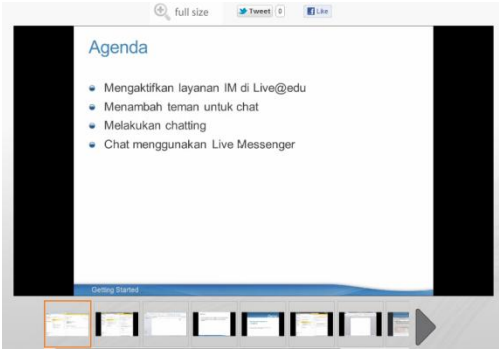



Figure 4. Video Playlist For Self-Paced E-Learning

The internal phase implementation model enhances the usages of the system by 27% for the first month. After that, the increment of the system usage is between 20-25% and it is reached to use by the entire employees in the fourth months.

The external phase implementation is started two months after the internal phase is started. The external phase focuses in three main activities which are:

- Building and configuring online E-learning that can be accessed through the web. Figure 5 shows the E-online learning that exposed to the external users.
- Collaborating and engaging the local community leader to learn and to acknowledge the others peers about online E-learning.
- The online E-learning contains the entire E-learning topic that is described in Table VII.

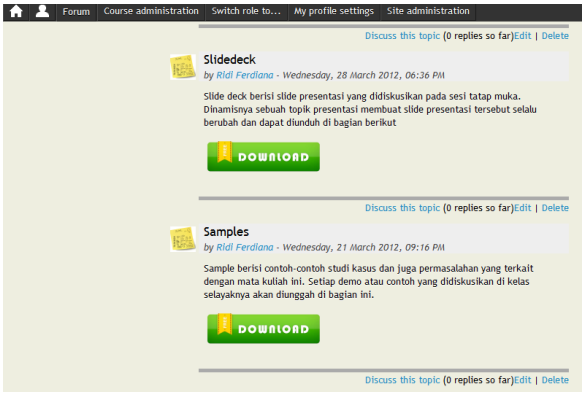



Figure 5. Online E-Learning For External Implementation Phase

The external phase implementation increases the usages of the system by 5% for the first week. The thousand numbers of users that separated in geographically make the implementation slower. Therefore, the further implementation and the usage of the new system are much depends with the local community leader. Therefore, the external phase implementation is still executed as a part of regular activity in the organization.

Table VIII shows the overall adoption that reach 27% average improvement using the proposed method. The percentage is calculated by counting the system usage in four months. The calculation compares between before and after adopting the ICT-IDM method. The percentage is calculated from the number of active user divides by the total user. The implementation phase in the table shows the total users based on external and internal phase for four months.

TABLE VIII. USAGE TABLES

| System Features | Initial Phase (Before Adopting The ITC-IDM) | Implementation phase (After Adopting The ICT-IDM) | Improvement Usage |
|---|---|---|---|
| Email | 13% | 76% | 63% |
| Calendar | 2% | 23% | 21% |
| Task | 0% | 21% | 21% |
| Online Storage | 5% | 38% | 33% |
| Public Address Book | 3% | 16% | 13% |
| Instant Messaging | 0% | 13% | 13% |
| Improvement average | | | 27% |

## VI. DISCUSSION AND FURTHER WORK

In this research, it is started with believe that a new ICT system that developed is not used optimally. Therefore, the research creates an initiative to do further ICT system

socialization by adopting e-learning initiative using IDM approaches. The research discovers several techniques and approaches to increase the new ICT system adoption such as follows.

- Closer Communication to the users and the stakeholder. This technique creates a clear view for the bottleneck and the real problem that exist in the ICT system implementation.
- Incremental adoption plan. This technique is done by doing implementation through three milestone phases. Each phase addresses the specific problem and approaches.
- Problem driven learning development. This technique is done by creating a learning development through several learning phases namely migration learning, essential learning, and enhancement learning. These learning phases use instruction design model
- Multimedia e-learning development. This technique is done by adopting IDM through multimedia e-learning content. Video, demo script and slide deck. The contents are uploaded into online storage to reach the users. The multimedia e-learning provides self-paced e-learning to decrease the skill gap between users by providing flexible way to learn the system.
- Iterative learning implementation. This technique is done by using two phases of the implementation which are external and internal phase. Both phases use iterative and continuous improvement model to increase the usage and the system adoption.

The techniques still has future work since the ICT-IDM only increase the system usage for just 27%. The hardest part is to socialize to the external organization. Therefore, it needs further improvement to eliminate geographical problem, huge numbers of user, and different skills of the users. The global model of the ICT-IDM adoption should be proposed for the further work of the research.

## ACKNOWLEDGMENT

We thank Partners in learning program that supports this research initiative. Microsoft Innovation Center UGM people and the ICT people at P4TK Mathematic organization that provided a helpful insight and approaches to do this research.

## REFERENCES

[1] Allen, Michael W. 2007. Designing Successful e-Learning: Forget What You Know about Instructional Design and Do Something Interesting, Michael Allen's e-Learning Library. Pfeiffer.

[2] Bernard R. Gifford and Noel D. Enyedy. 1999. Activity centered design: towards a theoretical framework for CSCL. In Proceedings of the 1999 conference on Computer support for collaborative learning (CSCL '99), USA.

[3] Brad Mehlenbacher, Krista Holstein, Brett Gordon, and Khalil Khammar. 2010. Reviewing the research on distance education and e-learning. In Proceedings of the 28th ACM International Conference on Design of Communication (SIGDOC '10). ACM, New York, USA.

[4] Carliner, Saul, and Patti Shank (eds). 2008. The e-Learning Handbook: Past Promises, Present Challenges. Pfeiffer.

[5] Carliner, Saul. 2002. Designing E-learning. ASTD. USA. 978-1-56286-332-6.

[6] Chao Boon Teo and Robert Kheng Leng Gay. 2006. A knowledge-driven model to personalize e-learning. J. Educ. Resour. Comput. 6, 1, Article 3 (March 2006).

[7] Dietmar P. F. Möller and Hamid Vakilzadian. 2010. Integrating modeling and simulation into an e-learning environment in engineering study programs. In Proceedings of the 2010 Conference on Grand Challenges in Modeling & Simulation (GCMS '10). Society for Modeling & Simulation International, Vista, CA, 90-97.

[8] Emma O'Brien and Timothy Hall. 2004. Training Needs Analysis: the first step in authoring e-learning content. In Proceedings of the 2004 ACM symposium on Applied computing (SAC '04). ACM, New York, NY, USA, 935-939.

[9] JE Stephenson, WB Morris, HG Tempest, DK Griffin, A Mileham, and AM Payne. 2007. The use of an e-learning constructivist solution in workplace learning. In Proceedings of the 14th European conference on Cognitive ergonomics: invent! explore! (ECCE '07).

[10] Jin-Sung Kim and Kwang-Jin Choi. 2009. Modeling rule for analysis and design of e-learning content. In Proceedings of the 2009 International Conference on Hybrid Information Technology (ICHIT '09). ACM, New York, NY, USA, 353-360.

[11] Rothwell, William J., and H.C. Kazanas. 2008. Mastering the Instructional Design Process: A Systematic Approach, Fourth Edition. Pfeiffer.

[12] S. S. (Thabo) Masemola and M. R. (Ruth) De Villiers. 2006. Towards a framework for usability testing of interactive e-learning applications in cognitive domains, illustrated by a case study. In Proceedings of the 2006 annual research conference of the Souuth African institute of computer scientists and information technologists on IT research in developing countries (SAICSIT '06), Judith Bishop and Derrick Kourie (Eds.). South African Institute for Computer Scientists and Information Technologists, , Republic of South Africa, 187-197.

[13] Satu Luojus and Olli Vilkki. 2008. Development of user-driven research methods as the starting point for living lab activities. In Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges (NordiCHI '08). ACM, New York, USA.

[14] Silvia Schiaffino, Analía Amandi, Isabela Gasparini, and Marcelo S. Pimenta. 2008. Personalization in e-learning: the adaptive system vs. the intelligent agent approaches. In Proceedings of the VIII Brazilian Symposium on Human Factors in Computing System.

[15] Stefano Ferretti, Silvia Mirri, Ludovico Antonio Muratori, Marco Roccetti, and Paola Salomoni. 2008. E-learning 2.0: you are We-LCoME!. In Proceedings of the 2008 international cross-disciplinary conference on Web accessibility (W4A) (W4A '08). ACM, New York, USA

[16] Tsui, Frank & Karam, Orlando. 2011. Essentials of software engineering, second edition. Jones and Bartlett Publishers.

[17] Vliet, Hans van. 2008. Software Engineering: Principles and Practice, Third Edition. John Wiley & Sons.

[18] Watkins, John & Mills, Simon. 2011. Testing it: an off-the-shelf software testing process, 2nd edition. Cambridge University Press.

[19] Yao Chin-Bang. 2009. Context-aware customization e-learning system with intelligent on-line examination mechanism. In Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human (ICIS '09). ACM, New York, USA.

## AUTHORS PROFILE

**Ridi Ferdiana**. got his doctoral degree at Universitas Gadjah Mada in 2011. He earned his master degree from the same university in 2006. In his professional area, he holds several professional certifications such as MCP, MCTS, MCPD, MCITP and MCT. In his daily research activities he really enjoys to learn about software engineering, business platform collaboration, and programming optimization. He can be reached at ridi@acm.org.

**Obert Hoseanto** is the Partners in Learning Manager of Microsoft Indonesia, based in Jakarta, Indonesia. His research interests are educational technology, e-learning and teacher education, and can be contacted at obert.hoseanto@microsoft.com.

# An Enhanced MPLS-TE For Transferring Multimedia packets

Abdellah Jamali
Dept. of Computer Science and Mathematics
ESTB, Hassan 1er University
Berrechid, Morocco

Najib Naja
Dept. of RIM
Institute of Posts and Telecommunications
Rabat, Morocco

Driss El Ouadghiri
Dept. of Computer Science and Mathematics
My Ismail University
Meknes, Morocco

***Abstract*— Multi-Protocol Label Switching is useful in managing multimedia traffic when some links are too congested; MPLS Traffic Engineering is a growing implementation in today's service provider networks. In This paper we propose an improvement of MPLS-TE called EMPLS-TE, it is based on a modification of operation of Forwarding Equivalence Class (FEC) in order to provide the quality of service to stream multimedia. The performance of the EMPLS-TE is evaluated by a simulation model under a variety of network conditions. We also compare its performance with that of unmodified MPLS-TE and MPLS. We demonstrate how a small change to the MPLS-TE protocol can lead to significantly improved performance results. We present a comparative analysis between MPLS, MPLS-TE and Enhanced MPLS-TE (EMPLS-TE). Our proposed EMPLS-TE has a performance advantageous for multimedia applications in their movement in a congested and dense environment. EMPLS-TE defines paths for network traffic based on certain quality of service. The simulation study is conducted in this paper; it is a means to illustrate the benefits of using this Enhanced MPLS-TE for multimedia applications.**

***Keywords- Multi-Protocol Label Switching (MPLS); Multi-Protocol Label Switching Traffic Engineering (MPLS-TE); Forwarding Equivalence Class (FEC); Quality Of Service (QoS); Simulation.***

## I. INTRODUCTION

The goal of Traffic Engineering (TE) is to provide QoS to multimedia packets by reservation of the resources and optimum resources utilization [9]. Multiprotocol Label Switching (MPLS) technology [2] allows traffic engineering and enhances the performance of the existing protocols over the traditional IPv4 network. The central idea of MPLS is to attach a short fixed-length label to packets at the ingress router of the MPLS domain. Packet forwarding then depends on the tagged label, not on longest address match, as in traditional IP forwarding. A router placed on the edge of the MPLS domain, named Label Edge Router (LER) that is associated to a label on the basis of a Forwarding Equivalence Class (FEC). In the MPLS network, internal routers that perform swapping and label-based packet forwarding are called Label Switching Routers (LSRs) [15].

MPLS TE also extends the MPLS routing capabilities with support for constraint-based routing. IGPs typically compute routing information using a single metric. Instead of that simple approach, constraint-based routing can take into account more detailed information about network constraints, and policy resources. MPLS TE extends current link-state protocols (IS-IS and OSPF) to distribute such information. There is another approach to provide QoS to multimedia traffic: DiffServ-aware Traffic Engineering (DS-TE) [7] [6], by using three signaling protocols in MPLS networks: Label Distribution Protocol (LDP) [5], Constraint based Routing LDP (CR-LDP) [3] and Resource Reservation Protocol-Traffic Engineering (RSVP-TE) [4].

In this paper we focus on our paper presented in [1] and MPLS-TE as a technology rather used by operators, then we make an improvement on MPLS-TE and propose EMPLS-TE (Enhanced MPLS-TE).

In order to provide a good service for transferring multimedia packets that requires a large flow in the MPLS-TE networks we make an improvement to the method of processing speed in the FEC in MPLS-TE.

Rest of paper is organized as below:

Section II defines QoS as services that provide some combination of high security, high reliability, low packet drop rate, low delay, and low jitter. The same section reviews the working of traditional IP, MPLS and MPLS-TE and their salient features. In section III, we will describe our proposed enhancement EMPLS-TE and its methods. In section IV, we will simulate the MPLS, MPLS-TE and EMPLS-TE, and then we compare it with original MPLS, and with original MPLS-TE. Section V, will conclude this paper. Routing, MPLS and MPLS-TE and their salient features. In section III, we will describe our proposed enhancement EMPLS-TE and its methods. In section IV, we will simulate the MPLS, MPLS-TE and EMPLS-TE, and then we compare it with original MPLS, and with original MPLS-TE. Section V will conclude this paper.

## II. OVERVIEW

### A. Internet Quality of Service (QoS)

Originally, the Internet was developed for transferring file and accessing remote machines. Therefore, the Internet was not expected to transfer multimedia data at large data rate. Today, many different types of applications in Internet demand more secure more reliable and faster services. Both non-real time and real-time applications require some kinds of

QoS, such as high reliability, bounded delay and jitter, and high security. Therefore, I would like to define QoS as services that provide some combination of high security, high reliability, low packet drop rate, low delay, and low jitter; in general ATM is an example of a network technology that provides good QoS.

Although ATM can be used to transmit both IP packets and ATM data, it is less suitable for best effort services IP packets mainly because ATM supports only a small part of IP services. The most common and major QoS problem in the backbone network is unevenly distributed traffic. MPLS-TE can distribute traffic evenly and optimize network utilization TE ensures that all available network resources are optimally used during times of failure or traffic routing, which is needed when congestion happens. Network congestion is not easily solved by IP because of its characteristics: connectionless and best effort service. As results, bursts of traffic appear unexpectedly, routers are easily congested, and packets are dropped. Therefore, the current Internet has poor reliability, unbounded delay and jitter, and varied throughput.

### *B. Traditional IP Routing*

The IP was created as a connectionless network layer protocol that makes no attempt to discriminate between various application types. IP uses routing protocols as traditional Interior Gateway Routing Protocol (IGRP) [10], Intermediate System-to-Intermediate System (IS-IS) [14], Open Shortest Path First (OSPF) [18] to build routing tables for active number the equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop.

Links in an area of network, and therefore transferring data between the source and the destination [16], the operation of these protocols depends on how to promote and distribute information on the state network that are broadcast regularly and depends also on how to update the routing tables of all routers located in the same autonomous system (AS).

Each router uses the information on the overall state of network to maintain an independently its own routing tables so that it can transfer data successfully using the shortest path or the link state as metric maintain before deciding to send data.

The major problem of some of these protocols is that they transfer the data on the paths with minimal hops, and since they do not use the paths with many hops that can lead the data to the destination, then this strategy produces quite congested links.

So, the traditional routing IP traffic is routed by the same types of paths (short), and therefore a fairly large amount of packets is lost.

To tackle the problem of low delay and packet loss during the delivery of multimedia applications, it is necessary to think of improvement methods to use more effectively the available network resources. MPLS and MPLS-TE (MPLS Traffic Engineering) are some process that provides this functionality.

### *C. MPLS*

Multiprotocol Label Switching (MPLS) can speed up the flow of network traffic and make it easier to manage. MPLS is flexible, fast, cost-efficient and allows for network segmentation and quality of service (QoS). MPLS also offers a better way of transporting latency-sensitive applications like voice and video. While MPLS technology has been around for several years, businesses are now taking advantage of service provider offerings and beginning their own corporate implementations.

MPLS can be considered a technology that has brought an oriented connection for IP protocol. Therefore, network services and applications can exploit all of the advantages of MPLS. In other words, MPLS is a connection oriented technology that uses a label swapping technique with IP network routing [12]. A label is a small, fixed index, which identifies a Forward Equivalence Class; a group of IP packets that are forwarded over the same path with the same packet treatments. With MPLS, the packet is faster than with use IP address because MPLS uses labels to quickly check the next hop that leads to the destination without going to the network layer to analyze the packets along the path.

MPLS consists of routers: Label Switching Routers (LSR) and Label Edge Routers (LER). These routers use labels to quickly send packets to the destination.

An LSR is a router that forwards both conventional IP packets and MPLS labelled packets. An LER is an LSR at the edge of the MPLS network to add and remove labels. An LER connects between the MPLS domain and the non-MPLS domain such as IP network.

A flow of packets coming from a non-MPLS domain is first assigned a label at an incoming LER and its forward along the path as an old label is replaced with a new label at LSRs on the path. Therefore, a label is used to reach the next node.

Although the exchange of label is required on the path, and the search of the network layer is not required at LSRs routers due to transmission of the link layer with labels. In routers LERS the labels are completely removed and the packets are transmitted directly to other networks. MPLS label switched paths are an essential element in delivering end-to-end QoS. Without them, it is not possible to control the path of packet flows from requested packet treatments.

The assignment of labels to packets is based on the concept of forwarding equivalence class (FEC). According to this concept, packets which belong to the same FEC are assigned the same label at an ingress node to an MPLS domain. A FEC consists of packets entering a network through the same ingress node and exiting the network through the same ingress node. A FEC also consists of packets requiring similar QoS or packet treatment across the MPLS domain. The path traversed by a FEC is called a Label Switched Path (LSP). A signal protocol such as LDP (Load Distribution Protocol) or RSVP (Resource reservation Protocol) [17] is used to establish and release LSPs [13].

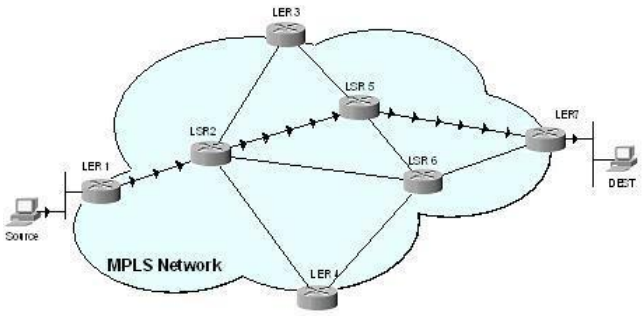



Figure 1. MPLS Network

### D. Traffic Engineering within MPLS

MPLS Traffic Engineering [11] is an obligation for network operators to provide a fairly reliable infrastructure and provides quality performance. Traffic Engineering provides efficient routing of traffic in the network to the use of network resources. This allows operators the ability to better exploit bandwidth resources on the network [8].

As a result of the unprecedented growth in demand for network resources and the competitiveness amongst providers, Traffic Engineering has become the primary application for MPLS.

MPLS Traffic Engineering responds to the ineffectiveness of some routing protocols in terms of datagram processing in the case of congestion. It allows a wider distribution flow of traffic across all available resources. Load balancing for TE in IP network requires an ability to control traffic flow precisely. In the traditional metric-based control, an administrator can change only link metrics, and the changes of some link metrics may affect the overall traffic flow. To manage the performance of a network, it is necessary to have explicit control over the paths that traffic flows traverse so that traffic flows can be arranged to maximize resource commitments and utilization of the networks [13].

The connection-oriented nature of MPLS allows ISPs to implement TE in their networks and achieve a variety of goals, including bandwidth assurance, different routing, load balancing, path redundancy, and other services that lead to QoS [9].

MPLS networks can use native TE mechanisms to minimize network congestion and improve network performance. TE modifies routing patterns to provide efficient mapping of traffic streams to network resources. This efficient mapping can reduce the occurrence of congestion and improves service quality in terms of the latency, jitter, and loss that packets experience. Historically, IP networks relied on the optimization of underlying network infrastructure or Interior Gateway Protocol (IGP) tuning for TE. Instead, MPLS extends existing IP protocols and makes use of MPLS forwarding capabilities to provide native TE. In addition, MPLS TE can reduce the impact of network failures and increase service availability. RFC 2702 discusses the requirements for TE in MPLS networks.

MPLS TE brings explicit routing capabilities to MPLS networks. An originating label switching route (LSR) can set up a TE label switched path (LSP) to a terminating LSR through an explicitly defined path containing a list of intermediate LSRs. IP uses destination-based routing and does not provide a general and scalable method for explicitly routing traffic. In contrast, MPLS networks can support destination-based and explicit routing simultaneously. MPLS TE uses extensions to RSVP and the MPLS forwarding paradigm to provide explicit routing. These enhancements provide a level of routing control that makes MPLS suitable for TE.

### E. Problem Context And Enhacem Ent Of Mpls-Te

We propose an improvement for the FEC group treatment in order to consider the throughput as an important parameter for multimedia applications that allows it to select the best paths in its routing.

MPLS-TE determines LSP as a sequence of labels in the packet to construct a path and to convey through these paths established by the protocol for distributing labels. The problem of MPLS-TE is how to select the FEC groups that satisfy some parameters of quality of service and in particular the throughput which can be considered as important parameter for some types of applications. The choice of FEC group is also according to several parameters (source address, destination address, QoS parameters). To solve this problem, we group all packets for multimedia applications in a specific FEC, with a high throughput, and LSP that consists of a sequence of labels for multimedia packets is associated with this FEC by the LDP protocol which provides this information to routers LSR on the throughput which we have chosen for packets multimedia in a specified FEC. Hence these multimedia packets take the paths that correspond to throughput as an important quality of service parameter. After the improvement in the FEC group associated with multimedia packets which are labeled, these packets with the principle of MPLS-TE are switched toward the MPLS-TE network by using number of label and the LSP paths. The LSR routers of MPLS-TE network switches the FEC labels that we improved to LER routers, taking into account the throughput that we have set for these multimedia packets.

### F. Simulation And Analysis Of The Solution

To analyze the proposed solution and the effectiveness of our suggested enhancement in MPLS-TE, we use an event-driven network simulator targeted at networking research. The software version used in this paper is ns-2.34 with MPLS Network Simulator (MNS 2.0).

MPLS-TE and EMPL-TE as discussed in the previous sections have several desirable capabilities. However in this paper, the simulation was chosen to demonstrate the ability of EMPLS-TE in providing Traffic Engineering. To demonstrate this capability, the simulations were setup using a normal MPLS, and a normal MPLS with Traffic engineering implemented (MPLS-TE). The results from these simulations are used for the comparison between the three approaches and evaluate our proposed scheme. Both simulations are based on the common topology.

### G. Simulation environment

The network consists of 90 nodes (in backbone, sources and destination). All links were set up as duplex with 15 ms delay and using Drop Tail Queuing, which serve packets on a First Come First Serve (FCFS) basis. The simulation time is 200s and the links have a capacity of 1.5 Mbps and the transmitted flux in the network is multimedia.

TABLE I. SIMULATION PARAMETRS

| *Simulation Parameter* | *Value* |
|---|---|
| Simulator | NS-2.34 |
| Simulation Time | 200s |
| Node Max. IFQ Length | 50 |
| Data Packet Size | 512 bytes |
| Traffic type | CBR(UDP) |
| Packet rate | 4pkt/sec |

### H. Performance Metrics

The following metrics are used in varying scenarios to evaluate different protocols:

Packet delivery ratio - This is defined as the ratio of the number of data packets received by the destinations to those sent by the CBR sources.

End-to-end delay of data packets - This is defined as the delay between the time at which the data packet was originated at the source and the time it reaches the destination. Data packets that get lost en route are not considered. Delays due to route discovery, queuing and retransmissions are included in the delay metric.

The metrics are measured against various mobility scenarios and with varying number of data connections.

### I. Comparison between MPLS, MPLS-TE and EMPLS-TE

In this subsection, we present a comparative analysis of the performance metrics of the MPLS, MPLS-TE and our approach EMPLS-TE.

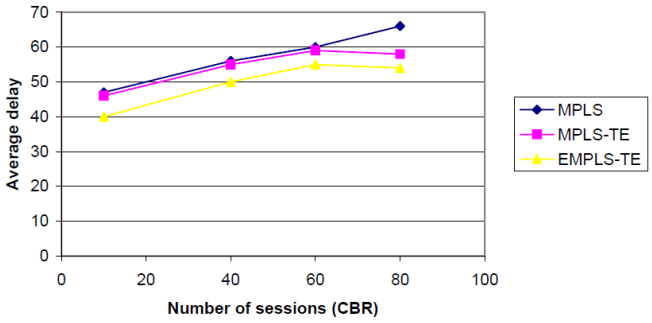



Figure 2. *Average delay Vs number of sessions CBR*

Packet Delivery Ratio:

Figure 3 gives the packet delivery ratio when the number of sessions (CBR) varies. With number of sessions from 60 to 80 both EMPLS-TE and MPLS-TE has almost same packet delivery ratio but as with number of sessions from 10 to 60 the packet delivery fraction of EMPLS-TE is better. The ratio decreases rapidly in case of MPLS whereas MPLS-TE maintains the same ratio. Thus with the increase in number of sessions EMPLS-TE gives more packet delivery fraction thereby outperforming MPLS and MPLS-TE.

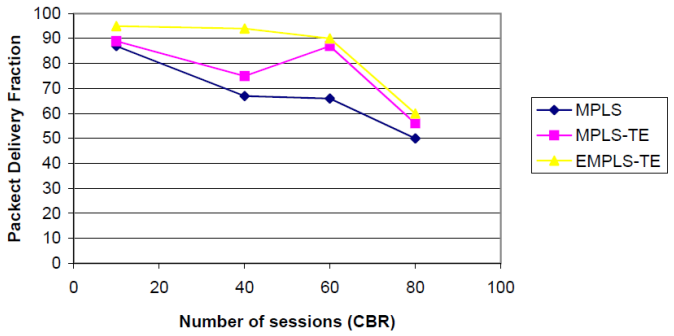



Figure 3. *Packet delivery fraction Vs number of sessions CBR*

Packets loss :

EMPLS-TE has less packets loss than both MPLS and MPLS-TE under almost all possible values of number of sessions. The difference is magnified under high number of sessions (40 and 60). The primary reason is that in MPLS-TE architecture, we group all packets for multimedia applications in a specific FEC, with a high throughput as compared to that in MPLS. MPLS-TE performs considerably better than both MPLS and MPLS-TE, because MPLS and MPLS-TE focus on LSP routes with the fewest hops, while MPLS-TE tends to choose the least congested route with a specific FEC. Also, when utilizing promiscuous listening MPLS-TE has to spend time processing any control packet it receives, even if it is not the intended recipient. For the time of simulation, the packets loss increases with an increase in the number of sessions.

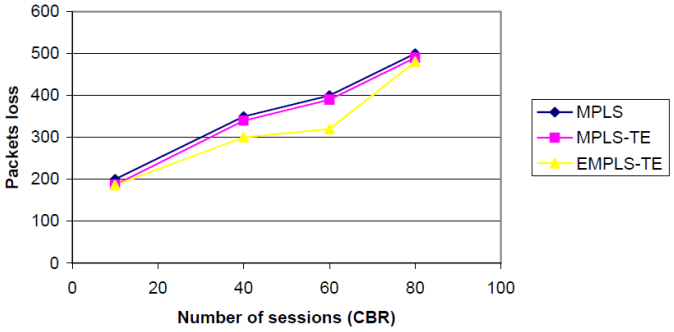



Figure 4. Packets Vs Number of sessions

Throughput with UDP :

From Figure 5, it is clear that at 10 second, MPLS-TE gives better throughput than EMPLS-TE. As the simulation times increases to 50, both MPLS and MPLS-TE have almost the same throughput but as the simulation times increases beyond 200 EMPLS-TE outperforms MPLS-TE and MPLS (as the throughput of all MPLS, MPLS-TE and EMPLS-TE increase with simulation times).

The throughput of EMLPS-TE is similar to MPLS and MPLS-TE between 10s and 50s. The architecture suffers a

little at fewer simulation times. At low simulation time, the throughput does not exceed 0, 4 Mbps in MPLS, MPLS-TE and EMPLS-TE due to packet collisions. This is because the number of collisions increases in EMPLS-TE due to the additional pending data packets sent by the intermediate routes during route discovery. The throughput increases quickly with increase in simulation times from 100s. Our EMPL-TE solution is very efficient at 200s. The obtained results show that EMPL-TE is an architecture designed for long periods.

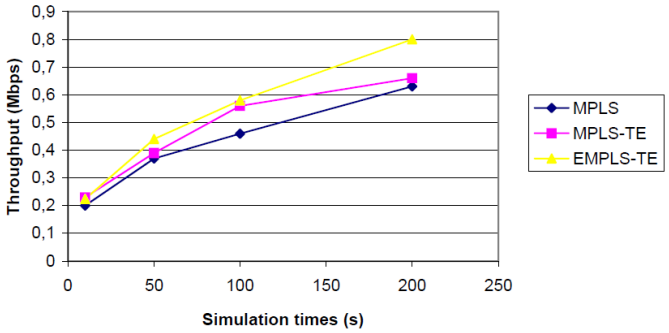



Figure 5. Throughput Vs Simulation times

Throughput with TCP :

In the first two source nodes send the multimedia stream to their destination through the MPLS network, MPLS-TE network and EMPLS-TE network. And we calculate the TCP throughput and UDP throughput at two destinations (see Figure. 5 and Figure. 6), we note that between 0 second and 10 seconds, our approach EMPLS-TE is more efficient, we obtain an important value of packet delivery ratio, but with author approach (MPLS-TE), the result is not efficient between 0s and 10s.

This improvement of the packet delivery ratio is due to enhanced throughput with FEC that we changed in MPLS-TE, and as result it performs the transmission of multimedia stream.

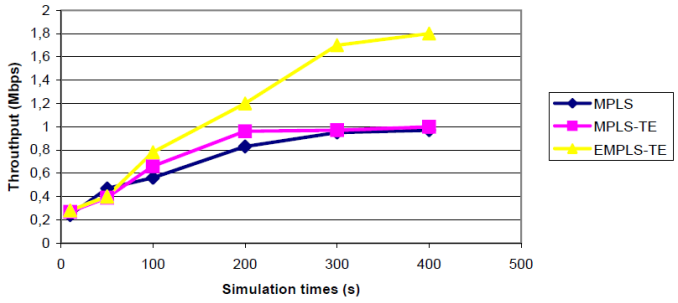



Figure 6. Throughput Vs Simulation times

## III. Conclusion

Through simulation results and analysis, it was clear that MPLS-TE does not provide a reliable service and improved packet delivery ratio as an important performance metric to ensure the arrival of received packets for sensitive applications as multimedia packets.

## ACKNOWLEDGMENT

This work was supported in part by the networks laboratory at Institute of the Post and Telecommunications-Rabat. We would like to thank for his help and the anonymous referees for their valuable comments.

## REFERENCES

[1] A. Jamali, N. Naja, D. El Oudghiri, and R. Benaini, “Improving quality of service (QoS) in multi-protocol label switching module,” in MMS’2009 IEEE Xplore.

[2] R. Callon, E. Rosen, and A. Viswanathan, “Multiprotocol label switching architecture,” In IETF RFC3031, Janvier, 2001.

[3] B. Jamoussi, and al. “Constraint based isp setup using idp,” In IETF RFC3212, Junuary 2002 ggdd

[4] D. Awduche and al. “Rsvp-te : extensions to rsvp for isp tunnels,” In IETF RFC3209, December 2001.

[5] W. Lai, “Requirements for support of differentiated services-aware mpls traffic engineering,” In IETF RFC3564, July 2003.

[6] A. Lund, C. Reingold, N. Rexford, J. Feldmann, and A. Greenberg, “Traffic engineering for ip networks,” IEEE Network Magazine, vol. 14, pp. 11–19, April 2000.

[7] C. L. Hedrick, “An introduction to igrp,” In cisco white paper, August 1991.

[8] H. Wang, J. Liebeherr, J. Wang, and S. Patek, “Traffic engineering with aimd in mpls networks,” In LNCS, May 2002.

[9] A. Toguyni, A. Rahmani, K. Lee, and A. Noce, “Comparison of multipath algorithms for load balancing in a mpls network,” ICOIN’05, Published by Springer as Lecture Notes in Computer Science (LNCS), pp. 463-470, February 2005.

[10] L. Kyeongja, ”Global model for quality of service in fai networks : integration of diffserv and traffic engineering based on mpls,” In Thesis, order number 0025. university of science and technology , Central School .

[11] A. Martey, “Integrated is-is routing protocol concepts,” In Cisco Press, May 2002.

[12] K. A. M. Lutfullah, M. Z. Hassan, M. R. Amin, M. Arifur Rahman, and A. H. Kabir, “ Performance analysis and the study of the be- havior of mpls protocols,” In International Conference on Computer and Communication Engineering, pp. 13–15, Kuala Lumpur, Malaysia, May 2008.

[13] P. Lewis James, “Cisco tcp/ip routing professional reference,”revised and expanded. 1998.

[14] S. Berson, S. Herzog, S. Jamin, R. Braden, and L. Zhang, “Resource reservation protocol (rsvp),” In RFC2205, Septembre 1997.

[15] D. Williams R. Guerin, and A. Orda, “Qos routing mechanisms and ospf extensions,” In of the Global Internet Miniconference, 1997.

## AUTHORS PROFILE

**Jamali Abdellah** is a researsh and professor at Hassan 1er University, Settat, Morroco, since Ocyober 2011. He was born in Ouarzazate, Morocco. He received a thesis in Computer Science from the University of Hassan II, Mohammedia, Morocco. He is a founding member, in 2007, of a research group e-NGN (e-Next Generation Networks) for Africa and Middle East. His research interests include the computer networks, IPv6 Networks, Quality of Service in MPLS and QoS in Ad Hoc and Networks performance analysis.

**Najib NAJA** was born in Maohammadia, Morroco. In 1989, He received his engineering degree from TELECOM Bretagne, France, option : Computer Science and Networks. In 1994, he received the PhD thesis in Computer Science from the University of Rennes I, France. In 1997, he received a thesis in telecommunications from the University of Mohammed V, Morocco. His research interests include the Information

system, Ad Hoc Network, signal processing, information technology, QoS in Ad Hoc, Networks performance analysis. He is Professor in The National Institute of Posts and Telecommunications (INPT) in Rabat, Morocco since 1994.

**Driss El Ouadghiri** is a research and an associate professor at Science Faculty, Moulay Ismail University, Meknes, Morocco, since September 1994. He was born in Ouarzazate, Morocco. He got his "License" in applied mathematics and his "Doctorat de Spécialité de Troisième Cycle" in computer networks, respectively, in 1992 and 1997 from Mohamed V

University, Rabat, Morocco. In 2000 he got his PhD in performance evaluation in wide area networks from Moulay Ismail University, Meknes, Morocco. He is a founding member, in 2007, of a research group e-NGN (e-Next Generation Networks) for Africa and Middle East. His research interests focus on performance evaluation in networks (modelling and simulation), DiffServ architecture (mechanisms based active queue management) and IPv6 networks.

# A New Algorithm for Data Compression Optimization

I Made Agus Dwi Suarjaya
Information Technology Department
Udayana University
Bali, Indonesia

***Abstract*— People tend to store a lot of files inside theirs storage. When the storage nears it limit, they then try to reduce those files size to minimum by using data compression software. In this paper we propose a new algorithm for data compression, called j-bit encoding (JBE). This algorithm will manipulates each bit of data inside file to minimize the size without losing any data after decoding which is classified to lossless compression. This basic algorithm is intended to be combining with other data compression algorithms to optimize the compression ratio. The performance of this algorithm is measured by comparing combination of different data compression algorithms.**

***Keywords- algorithms; data compression; j-bit encoding; JBE; lossless.***

## I. INTRODUCTION

Data compression is a way to reduce storage cost by eliminating redundancies that happen in most files. There are two types of compression, lossy and lossless. Lossy compression reduced file size by eliminating some unneeded data that won't be recognize by human after decoding, this often used by video and audio compression. Lossless compression on the other hand, manipulates each bit of data inside file to minimize the size without losing any data after decoding. This is important because if file lost even a single bit after decoding, that mean the file is corrupted.

Data compression can also be used for in-network processing technique in order to save energy because it reduces the amount of data in order to reduce data transmitted and/or decreases transfer time because the size of data is reduced [1].

There are some well-known data compression algorithms. In this paper we will take a look on various data compression algorithms that can be use in combination with our proposed algorithms. Those algorithms can be classified into transformation and compression algorithms. Transformation algorithm does not compress data but rearrange or change data to optimize input for the next sequence of transformation or compression algorithm.

Most compression methods are physical and logical. They are physical because look only at the bits in the input stream and ignore the meaning of the contents in the input. Such a method translates one bit stream into another, shorter, one. The only way to understand and decode of the output stream is by knowing how it was encoded. They are logical because look only at individual contents in the source stream and replace common contents with short codes. Logical compression method is useful and effective (achieve best compression ratio) on certain types of data [2].

## II. RELATED ALGORITHMS

### A. Run-length encoding

Run-length encoding (RLE) is one of basic technique for data compression. The idea behind this approach is this: If a data item *d* occurs *n* consecutive times in the input stream, replace the n occurrences with the single pair *nd* [2].

RLE is mainly used to compress runs of the same byte [3]. This approach is useful when repetition often occurs inside data. That is why RLE is one good choice to compress a bitmap image especially the low bit one, example 8 bit bitmap image.

### B. Burrows-wheeler transform

Burrows-wheeler transform (BWT) works in block mode while others mostly work in streaming mode. This algorithm classified into transformation algorithm because the main idea is to rearrange (by adding and sorting) and concentrate symbols. These concentrated symbols then can be used as input for another algorithm to achieve good compression ratios.

Since the BWT operates on data in memory, you may encounter files too big to process in one fell swoop. In these cases, the file must be split up and processed a block at a time [3]. To speed up the sorting process, it is possible to do parallel sorting or using larger block of input if more memory available.

### C. Move to front transform

Move to front transform (MTF) is another basic technique for data compression. MTF is a transformation algorithm which does not compress data but can help to reduce redundancy sometimes [5]. The main idea is to move to front the symbols that mostly occur, so those symbols will have smaller output number.

This technique is intended to be used as optimization for other algorithm likes Burrows-wheeler transform.

### D. Arithmetic coding

Arithmetic coding (ARI) is using statistical method to compress data. The method starts with a certain interval, it reads the input file symbol by symbol, and uses the probability of each symbol to narrow the interval. Specifying a narrower interval requires more bits, so the number constructed by the algorithm grows continuously. To achieve compression, the algorithm is designed such that a high-probability symbol

narrows the interval less than a low-probability symbol, with the result that high-probability symbols contribute fewer bits to the output [2].

Arithmetic coding, is entropy coder widely used, the only problem is its speed, but compression tends to be better than Huffman (other statistical method algorithm) can achieve [6]. This technique is useful for final sequence of data compression combination algorithm and gives the most for compression ratio.

## III. PROPOSED ALGORITHM

J-bit encoding (JBE) works by manipulate bits of data to reduce the size and optimize input for other algorithm. The main idea of this algorithm is to split the input data into two data where the first data will contain original nonzero byte and the second data will contain bit value explaining position of nonzero and zero bytes. Both data then can be compress separately with other data compression algorithm to achieve maximum compression ratio. Step-by-step of the compression process can be describe as below:

1. Read input per byte, can be all types of file.
2. Determine read byte as nonzero or zero byte.
3. Write nonzero byte into data I and write bit ‘1’ into temporary byte data, or only write bit ‘0’ into temporary byte data for zero input byte.
4. Repeat step 1-3 until temporary byte data filled with 8 bits of data.
5. If temporary byte data filled with 8 bit then write the byte value of temporary byte data into data II.
6. Clear temporary byte data.
7. Repeat step 1-6 until end of file is reach.
8. Write combined output data
    a) Write original input length.
    b) Write data I.
    c) Write data II.
9. If followed by another compression algorithm, data I and data II can be compress separately before combined (optional).

Figure 1 shows visual step-by-step compression process for J-bit encoding. Inserted original input length into the beginning of the output will be used as information for data I and data II size. As for step-by-step of the decompression process can be describe below:

1. Read original input length.
2. If was compressed separately, decompress data I and data II (optional).
3. Read data II per bit.
4. Determine whether read bit is '0' or '1'.
5. Write to output, if read bit is '1' then read and write data I to output, if read bit is '0' then write zero byte to output.
6. Repeat step 2-5 until original input length is reach.

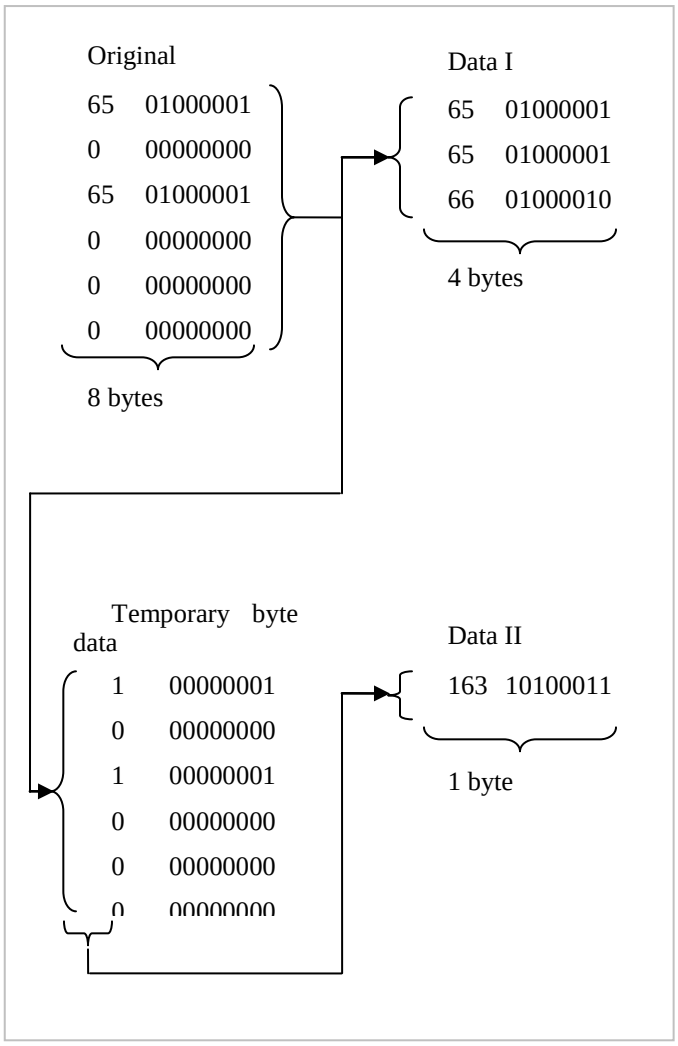



Figure 1. J-bit Encoding process

## IV. COMBINATION COMPARISON

Five combinations of data compression algorithm are used to find out which combination with the best compression ratio. The combinations are:

1. RLE+ARI.
2. BWT+MTF+ARI.
3. BWT+RLE+ARI.
4. RLE+BWT+MTF+RLE+ARI (as used in [3]).
5. RLE+BWT+MTF+**JBE**+ARI.

Those combinations are tested with 5 types of files. Each type consists of 50 samples. Each sample has different size to show real file system condition. All samples are uncompressed, this include raw bitmap images and raw audio without lossy

compression. Average compression ratio for each type of file is used. Samples for the experiment are show in table 1.

TABLE I. SAMPLES FOR COMBINATION INPUT

| No | Name | Qty | Type | Spec. |
|---|---|---|---|---|
| 1 | Image | 50 | Bitmap Image | Raw 8 bit |
| 2 | Image | 50 | Bitmap Image | Raw 24 bit |
| 3 | Text | 50 | Text Document | |
| 4 | Binary | 50 | Executable, library | |
| 5 | Audio | 50 | Wave Audio | Raw |

## V. RESULT

Figure 2 shows that 8-bit bitmap images are compressed with good compression ratio by algorithms that combined with J-bit encoding.

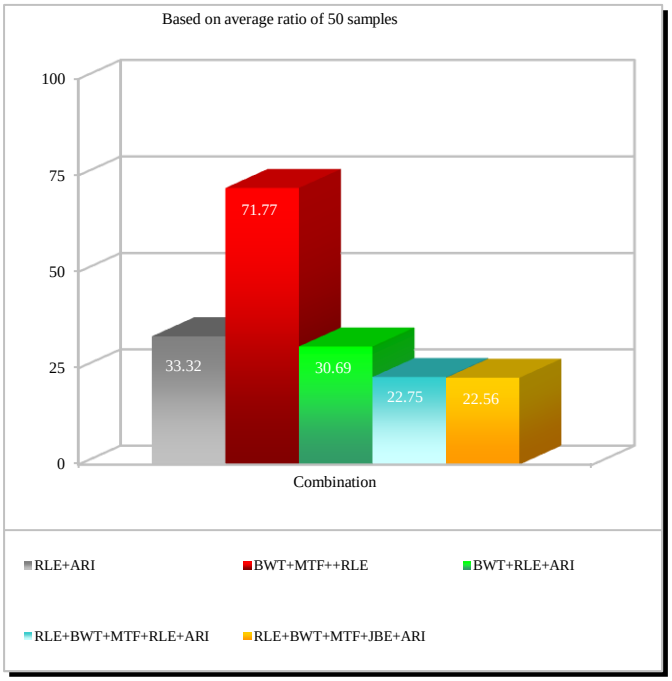



Figure 2. Ratio comparison for 8-bit bitmap image

Figure 3 shows that 24-bit bitmap images are compressed with better compression ratio by algorithms that combined with J-bit encoding. A 24 bit bitmap image has more complex data than 8 bit since it is store more color. Lossy compression for image would be more appropriate for 24 bit bitmap image to achieve best compression ratio, even thought that will decrease quality of the original image.

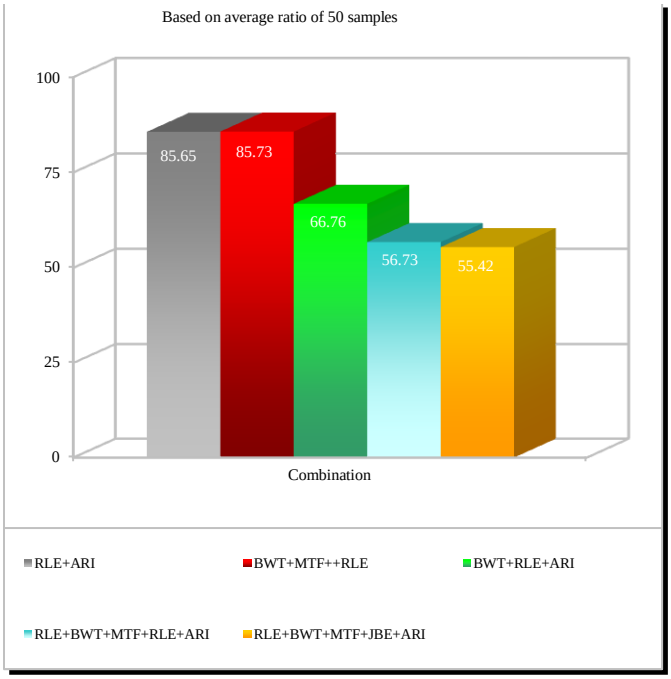



Figure 3. Ratio comparison for 24-bit bitmap image

Figure 4 shows that text files are compressed with better compression ratio by algorithms that combined with J-bit encoding.

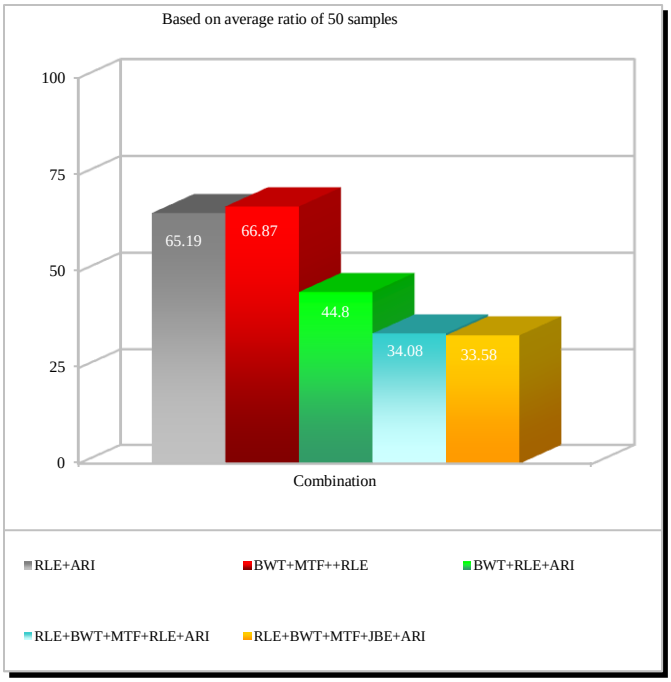



Figure 4. Ratio comparison for text

Figure 5 show that binary files are compressed with better compression ratio by algorithms that combined with J-bit encoding.

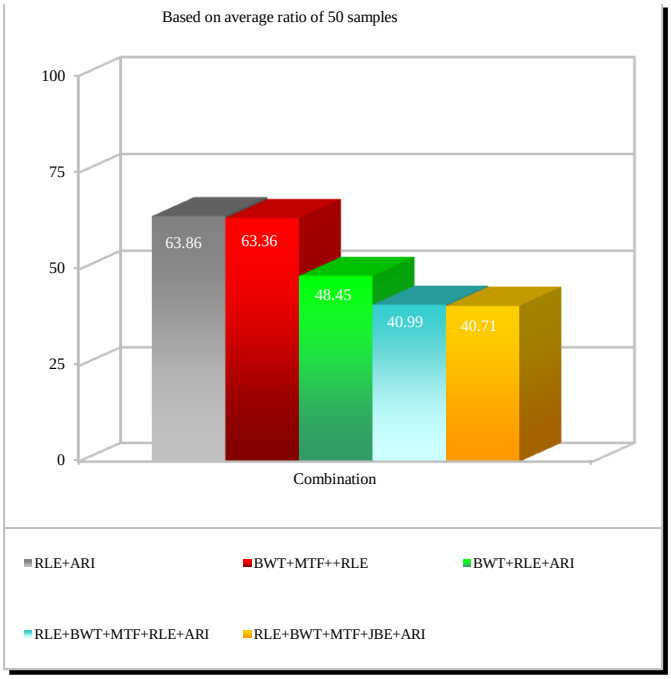



Figure 5. Ratio comparison for binary

Figure 6 shows that wave audio files are compressed with better compression ratio by algorithms that combined with J-bit encoding.

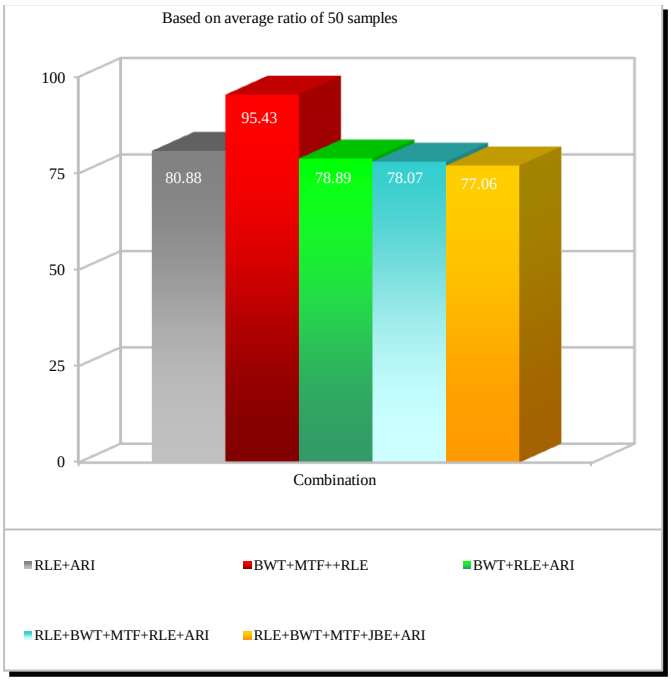



Figure 6. Ratio comparison for wave

## VI. CONCLUSION

This paper proposes and confirms a data compression algorithm that can be used to optimize other algorithm. An experiment by using 5 types of files with 50 different sizes for each type was conducted, 5 combination algorithms has been tested and compared. This algorithm gives better compression ratio when inserted between move to front transform (MTF) and arithmetic coding (ARI).

Because some files consist of hybrid contents (text, audio, video, binary in one file just like document file), the ability to recognize contents regardless the file type, split it then compresses it separately with appropriate algorithm to the contents is potential for further research in the future to achieve better compression ratio.

## REFERENCES

[1] Capo-chichi, E. P., Guyennet, H. and Friedt, J. K-RLE a New Data Compression Algorithm for Wireless Sensor Network. In Proceedings of the 2009 Third International Conference on Sensor Technologies and Applications.

[2] Salomon, D. 2004. Data Compression the Complete References Third Edition. Springer-Verlag New York, Inc.

[3] Nelson, M. 1996. Data compression with Burrows-Wheeler Transform. Dr. Dobb's Journal.

[4] Campos, A. S. E. Run Length Encoding. Available: http://www.arturocampos.com/ac_rle.html (last accessed July 2012).

[5] Campos, A. S. E. Move to Front. Available: http://www.arturocampos.com/ac_mtf.html (last accessed July 2012).

[6] Campos, A. S. E. Basic arithmetic coding. Available: http://www.arturocampos.com/ac_arithmetic.html (last accessed July 2012).

### AUTHORS PROFILE

**I Made Agus Dwi Suarjaya** received his Bachelors degree in Computer System and Information Science in 2007 from Udayana University and Masters degree in Information Technology in 2009 from Gadjah Mada University. He served as a full-time lecturer at Faculty of Engineering, Information Technology Department in Udayana University. His research interest include software engineering, networking, security, computing, artificial intelligent, operating system and multimedia.

# Monte Carlo Based Non-Linear Mixture Model of Earth Observation Satellite Imagery Pixel Data

Kohei Arai [1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*— Monte Carlo based non-linear mixel (mixed pixel) model of visible to near infrared radiometer of earth observation satellite imagery is proposed. Through comparative studies with actual real earth observation satellite imagery data between conventional linear mixel model and the proposed non-linear mixel model, it is found that the proposed mixel model represents the pixels in concern much precisely rather than the conventional linear mixel model.**

***Keywords- remote sensing satellite; visible to near infrared radiometer; mixed pixel: mixel; Monte Carlo simulation model.***

## I. INTRODUCTION

The pixels in earth observed images which are acquired with Visible to Near Infrared: VNIR sensors onboard remote sensing satellites are, essentially mixed pixels (mixels) which consists of several ground cover materials [1]. Some mixel model is required for analysis such as un-mixing of the mixel in concern [2],[3]. Typical mixel is linear mixing model which is represented by linear combination of several ground cover materials with mixing ratio for each material [4]. It is not always true that the linear mixel model is appropriate [5]. Due to the influences from multiple reflections between the atmosphere and ground, multiple scattering in the atmosphere on the observed radiance from the ground surface, pixel mixture model is essentially non-linear rather than linear. These influence is interpreted as adjacency effect [6].[7].

Method for representation of non-linear mixel model is not so easy. In particular, there is not sophisticated multi reflection model between ground materials. The representation method for non-linear mixel model is based on Monte Carlo Ray Tracing: MCRT model [8]. It is rather easy to designate surface slopes on the ground and multi reflection among trees for MCRT model. The proposed MCRT based non-linear mixel model is applied to real earth observation satellite imagery data of Advanced Spaceborn Thermal Emission and Reflection Radiometer / Visible and Near Infrared Radiometer: ASTER/VNIR onboard on Terra satellite. A comparison of radiance between the conventional linear mixel model and the proposed non-linear mixel model is conducted. As a result, validity of the proposed model is confirmed.

The following section describes the proposed non-linear mixel model based on MCRT followed by some experiments for validation of the proposed model. Then, finally, conclusions with some discussions are described.

## II. PROPOSED NON-LINEAR MIXEL MODEL

### A. Monte CarloRay Tracing Simulation

In order to show a validity of the proposed non-linear mixel model, MCRT simulation study and field experimental study is conducted. MCRT allows simulation of polarization characteristics of sea surface with designated parameters of the atmospheric conditions and sea surface and sea water conditions. Illustrative view of MCRT is shown in Fig.1.

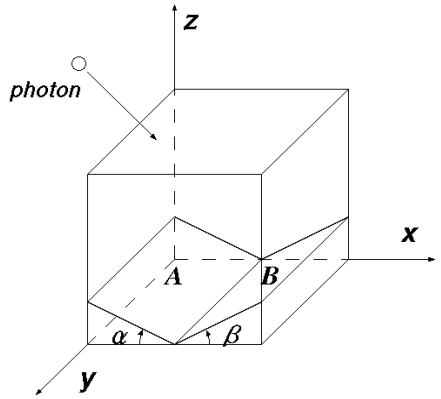
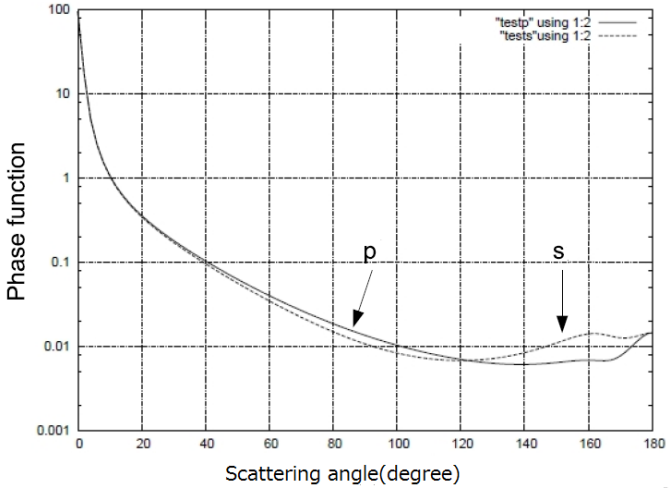


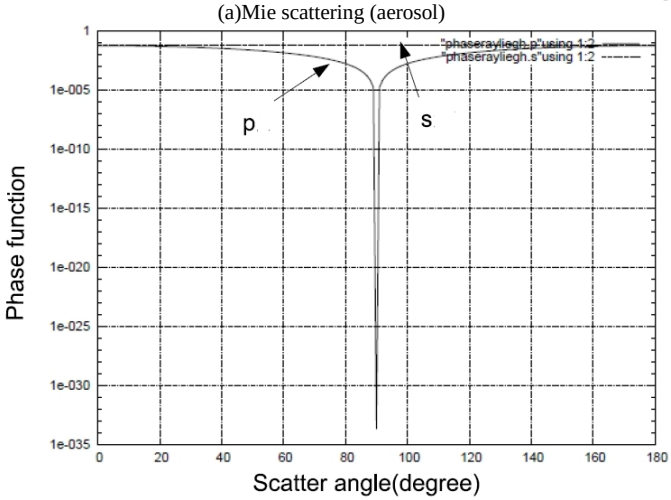

Figure 1 Illustrative view of MCRT for the atmosphere and sea water

Photon from the sun is input from the top of the atmosphere (the top of the simulation cell). Travel length of the photon is calculated with optical depth of the atmospheric molecule and that of aerosol. There are two components in the atmosphere; molecule and aerosol particles while three are also two components, water and particles; suspended solid and phytoplankton in the ocean. When the photon meets molecule or aerosol (the meeting probability with molecule and aerosol depends on their optical depth), then the photon scattered in accordance with scattering properties of molecule and aerosol. The scattering property is called as phase function[1]. In the visible to near infrared wavelength region, the scattering by molecule is followed by Rayleigh scattering law [10] while that by aerosol is followed by Mie scattering law [10]. Example of phase function of Mie scattering is shown in Fig.2 (a) while that of Rayleigh scattering is shown in Fig.2 (b).

[1] http://ejje.weblio.jp/content/phase+function




(a)Mie scattering (aerosol)




(b)Rayleigh scattering (Molecule)

Figure 2 Phase functions for Mie and Rayleigh scattering

In the atmosphere, there are absorption due to water vapor, ozone and aerosols together with scattering due to the atmospheric molecules, aerosols. Atmospheric Optical Depth: AOD (optical thickness) in total, Optical Depth: OD due to water vapor ($H_2O$), ozone ($O_3$), molecules (MOL), aerosols (AER), and real observed OD (OBS) are plotted in Fig.3 as an example.

For simplifying the calculations of the atmospheric influences, it is assumed that the atmosphere containing only molecules and aerosols. As shown in Fig.3, this assumption is not so bad. Thus the travel length of the photon at once, $L$ is expressed with equation (1).

$$L=L_0 \text{ RND}(i) \quad (1)$$

$$L_0=Z_{max}/\tau \quad (2)$$

where $Z_{max}$, $\tau$, RND($i$) are maximum length, altitude of the atmosphere, optical depth, and $i$-th random number, respectively. In this equation, $\tau$ is optical depth of molecule or aerosol. The photon meets molecule when the random number is greater than $\tau$. Meanwhile, if the random number is less than $\tau$, then the photon meats aerosol. The photon is scattered at the molecule or aerosol to the direction which is determined with the aforementioned phase function and with the rest of the travel length of the photon.

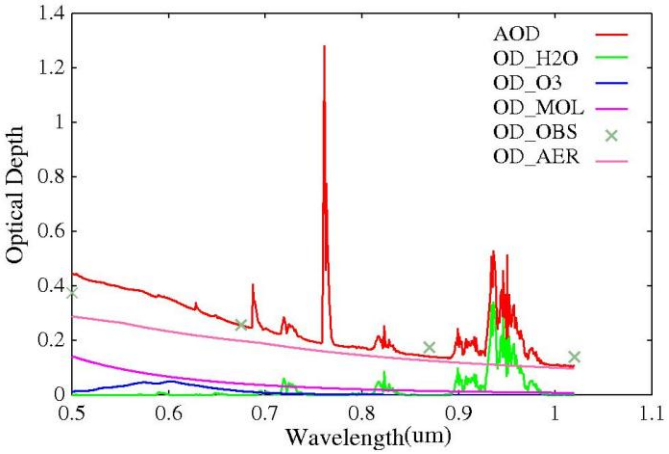



Figure 3 Example of observed atmospheric optical depth in total and the best fit curves of optical depth due to water vapor, ozone, molecules, and aerosols calculated with MODTRAN of atmospheric radiative transfer software code..

### B. *Ground Surface with Slopes*

When the photon reaches on the ground, the photon reflects at the ground surface to the direction which is determined by random number. Lambertian surface [11] is assumed. Therefore, reflectance is constant for all the directions. The reflected photon travels with the rest of travel length. Two adjacent slopes of Lambertian surfaces are assumed on the ground as shown in Fig.4. Slope angles for both are $\alpha$、$\beta$ while their reflectance are $\Gamma_A$ and $\Gamma_B$

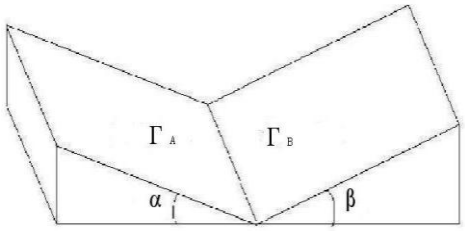



Figure 4 Two adjacent slopes of Lambertian surfaces which are assumed on the ground

### C. *Top of the Atmosphere: TOA Radiance Calculation*

If the photon reaches on the wall of the simulation cell, the photon disappears at the wall and it appears from the corresponding position on the opposite side wall. Then it travels with the rest of travel length. Eventually, the photons which are reached at the top of the atmosphere are gathered with the Instantaneous Field of View: IFOV of the Visible to Near Infrared Radiometer: VNIR onboard satellite. At sensor radiance, $I^+$ with direction and IFOV of $\mu$, $\mu_0$ can be calculated with equation (3)

$$I^+(\mu, \mu_0)=I\, N^+(\mu, \mu_0)/N_{total} \quad (3)$$

where $N^+$ is the number of photons which are gathered by VNIR, $N_{total}$ denotes the number of photons input to the simulation cell. Also $I$ denote extraterrestrial irradiance at the top of the atmosphere.

## III. EXPERIMENTS

### A. Validity of the Monte CarloRay Tracing Simulation

In order to confirm that the developed MCRT is valid, a comparative study is conducted between radiative transfer code of Gauss Seidel method and the MCRT derived TOA radiance. Because the Gauss Seidel method allows calculation of TOA radiance with flat surface of ground, 0.2 of reflectance of flat surface is assumed in the comparison. Also, 0.02 and 0.03 of optical depths are assumed for aerosol and molecule.

The size of simulation cell is determined as 50 km by 50 km by 50 km. Solar zenith angle is set at 30 degree while solar azimuth is set at 120 degree. 700,000 of photons are input to the simulation cell. TOA radiance derived from the Gauss Seidel method is 0.565 ($mW/m^2/sr/\mu m$) while that from the MCRT is 0.579 ($mW/m^2/sr/\mu m$) at the 500nm of wavelength. For both cases, IFOV of the VNIR radiometer is assumed to be $2\pi$, all of the photons output from the top of the atmosphere are counted. Therefore, the developed MCRT seems valid enough.

### B. TOA Radiance for the Different Combination of Optical Depths of Aerosol and Molecule and for the Ground with the Different Slopes

TOA radiance at 500 nm of wavelength for the different combination of optical depths of aerosol and molecule which ranges from 0.01 to 0.04 and for the ground with the different slopes, 0 and 20 degree are calculated.

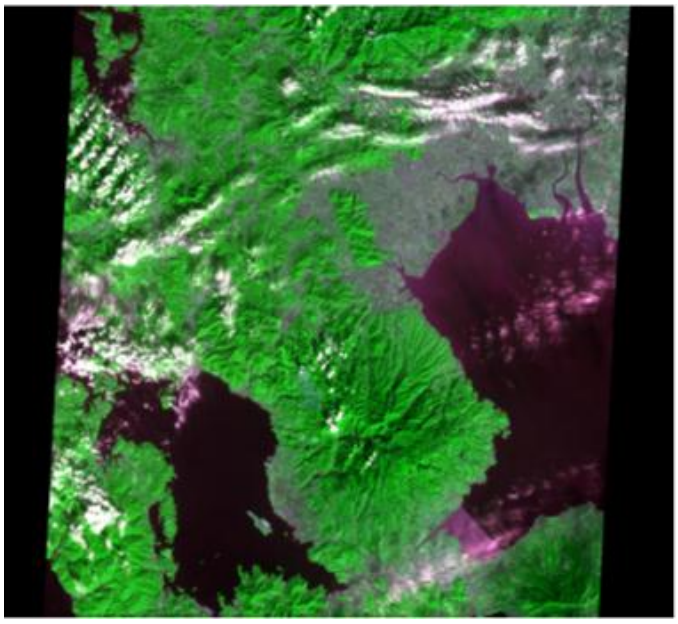
Again, IFOV of the VNIR radiometer is assumed to be $2\pi$, all of the photons output from the top of the atmosphere are counted. The reflectance for both slopes are same as 0.5. The results are shown in Table 1. In the table, $\tau_{aer}$ , $\tau_{mol}$ are optical depths of aerosol and molecule, respectively.

TABLE I. TOP OF THE ATMOSPHERE: TOA RADIANCE FOR THE COMBINATIONS OF ATMOSPHERIC CONDITIONS

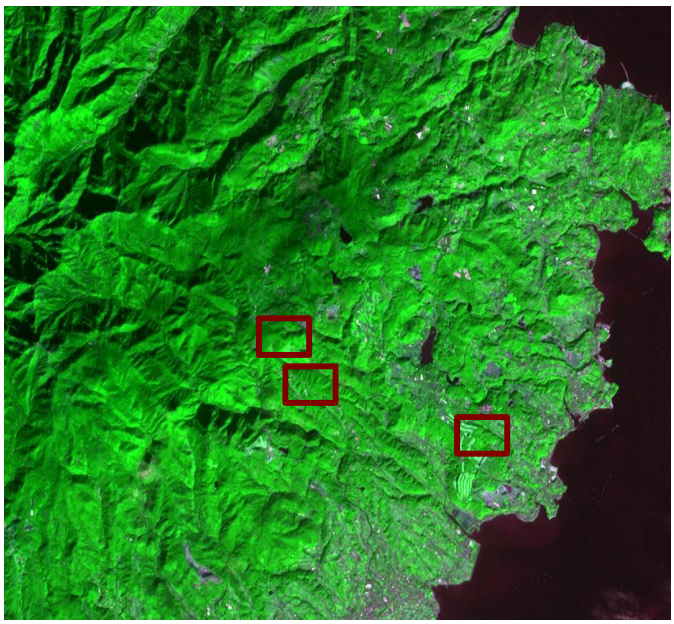
| | TOA radiance ($mW/m^2/sr/\mu m$) | | | |
|---|---|---|---|---|
| $\tau_{aer} \setminus \tau_{mol}$ | 0.01 | 0.02 | 0.03 | 0.04 |
| 0.01 | 3.51 | 0.317 | 3.99 | 6.51 |
| 0.02 | 5.65 | 1.91 | 1.09 | 3.04 |
| 0.03 | 5.7 | 3.08 | 0.622 | 10.7 |
| 0.04 | 3.29 | 3.85 | 3.97 | 7.45 |

### C. Validity of the Proposed Non-Linear Mixel Model with Real VNIR Data

The proposed non-linear mixel model based on MCRT is validated with real earth observation satellite imagery data of ASTER/VNIR onboard Terra satellite which is acquired at 11:09 Japanese Standard Time: JST on December 15 2004. IFOV of ASTER/VNIR is 15m with 60km of swath width. Whole scene of ASTER/VNIR is shown in Fig.5 (a) while Fig.5 (b) shows a portion of the scene.



(a)Whole scene of ASTER/VNIR image



(b)A portion of the scene

Figure 5 ASTER/VNIR image used for experiment

Three test sites, Area #1, 2, 3 are extracted from the scene. Attribute information of these sites are listed in Table 2.

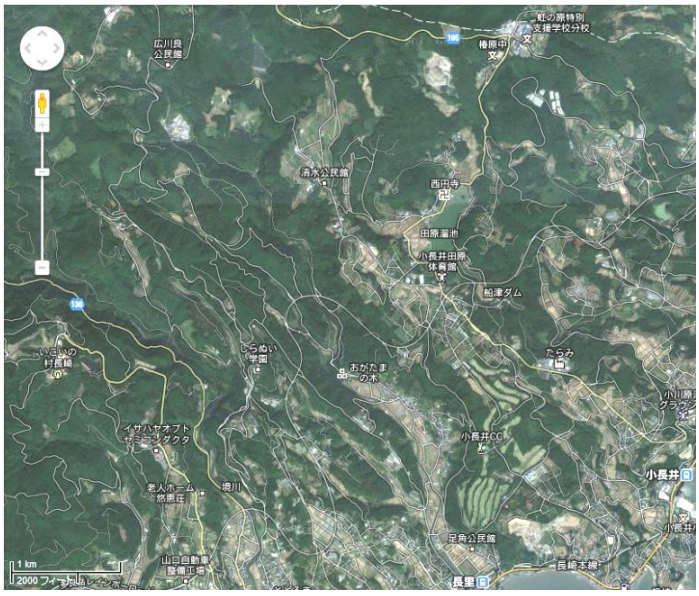
Fig.6 (a) shows three test sites on ASTER/VNIR image while Fig.6 (b) shows three test sites on Google map. Other than these, topographic map of three test sites which is corresponding to the Google map is shown in Fig.6 (c) while the extracted portion of each test site on ASTER/VNIR image is shown in Fig.6 (d), (e) and (f), respectively. These digital elevation models for three test sites are taken into account in the MCRT simulations.

Also, solar zenith angle of 58 degree and solar azimuth angle of 17 degree are taken into account in the simulations. From the atmospheric optical depth measurement data with sun photometer, optical depth of total atmosphere is calculated.

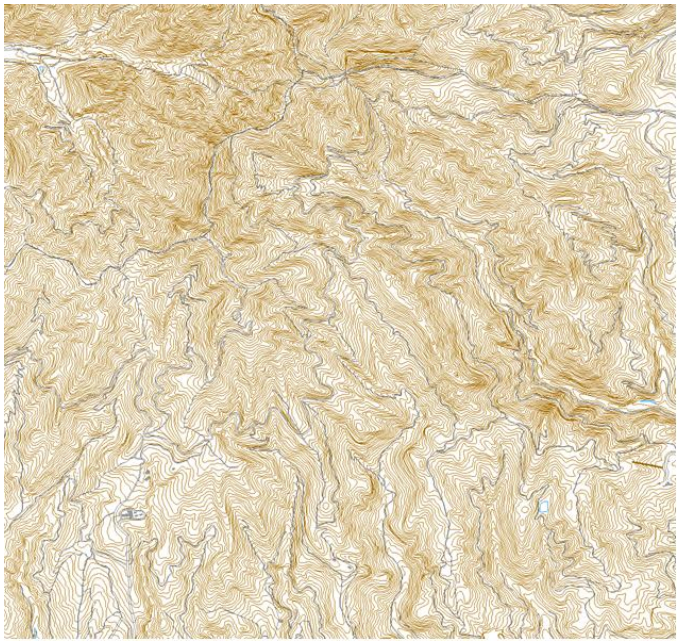
TABLE II. ATTRIBUTIONS FOR THE TEST SITE WITH SLOPES

| | Area #1 | Area #2 | Area #3 |
|---|---|---|---|
| Area Name | Korai-cho, Ochiai-gawa | Korai-cho, Ochiai-gawa | Konagai Golf Club |
| Latitude | 32°57'30" | 32°56'33" | 32°56'13" |
| Longitude | 130°7'19" | 130°7'25" | 130°10'21" |
| Slope A(°) | 24 | 30 | 20 |
| Slope B(°) | 28 | 26 | 0 |
| $\Gamma_A$ | 0.14 | 0.2 | 0.14 |
| Material | Deciduous | Bare Soil | Deciduous |
| $\Gamma_B$ | 0.08 | 0.08 | 0.12 |
| Material | Coniferous | Coniferous | Paddy |
| OD-Aerosol | 0.35 | 0.35 | 0.35 |
| OD-Molecule | 0.14 | 0.14 | 0.14 |


(a)Three test sites on ASTER/VNIR image



(b)Three test sites on Google Map



(c) Topographic map of corresponding area of three test sites on Google map

Coniferous (above), Deciduous (bottom)

(d)Area#1 (Korai-cho, Ochiai-gawa, Nagasaki, Japan)

Coniferous (above), Bare Soil (bottom)

(e)Area #2 (Korai-cho, Ochiai-gawa, Nagasaki, Japan)

Deciduous (above), Paddy field (bottom)

(f)Area #3 (Konagai Country Club, Nagasaki, Japan)

Figure 6 Three test site, Area #1, 2, 3.

Furthermore, molecule optical depth $\tau_R$ is calculated with equation (4) as a function of atmospheric pressure P which is measured on the ground.

$$\tau_R(\lambda) = \frac{P}{P_0} \cdot 0.00864\lambda^{-(3.916+0.074\lambda+\frac{0.05}{\lambda})} \quad (4)$$

where $P_0$ denotes standard atmospheric pressure on the ground (1013 hPa) while $\lambda$ denotes wavelength. Then aerosol optical depth is calculated from total atmospheric optical depth by subtracting molecule optical depth.

Comparative study is conducted between ASTER/VNIR derived radiance of Band 2 (Green band) and the radiance which derived from the conventional linear mixel model and the proposed non-linear mixel model.

Table 3 shows the calculated radiance in unit of $\mu W/m^2/sr/\mu m$ and the radiance difference between ASTER/VNIR and the estimated with the conventional and the proposed mixel models.

TABLE III. COMPARISON OF RADIANCE BETWEEN REAL ASTER/VNIR AND THE CONVENTIONAL LINEAR MIXEL MODEL AS WELL AS THE PROPOSED NON-LINEAR MIXEL MODEL DERIVED RADIANCE

| | Area #1 | Area #2 | Area #3 |
|---|---|---|---|
| ASTER/VNIR | 14.1 | 15.5 | 16 |
| Linear | 12.9 | 13.7 | 14.6 |
| Non-Linear | 13.7 | 14.3 | 15 |
| VNIR-Linear | 1.2 | 1.8 | 1.4 |
| VNIR-Non-Linear | 0.4 | 1.2 | 1 |

It is found that the estimated radiance with the proposed non-linear mixel model is much closer rather than that with the conventional linear mixel model.

## IV. CONCLUSION

Monte Carlo based non-linear mixel (mixed pixel) model of visible to near infrared radiometer of earth observation satellite imagery is proposed. Through comparative studies between ASTER/VNIR derived radiance and the conventional linear mixel model derived radiance as well as the proposed non-linear mixel model derived radiance, it is found that the estimated radiance with the proposed non-linear mixel model is much closer to ASTER/VNIR derived radiance (around 6%) rather than that with the conventional linear mixel model.

One of the disadvantages of the proposed non-linear mixel model based on MCRT is time consumable computations. Acceleration is highly required.

## ACKNOWLEDGMENT

The author would like to thank Dr. Yasunori Terayama and Mr. Kohei Imaoka of Saga University for their effort to simulation study and experiments.

## REFERENCES

[1] Masao Matsumoto, Hiroki Fujiku, Kiyoshi Tsuchiya, Kohei Arai, Category decomposition in the maximum likelihood classification, Journal of Japan Society of Phtogrammetro and Remote Sensing, 30, 2, 25-34, 1991.

[2] Masao Moriyama, Yasunori Terayama, Kohei Arai, Clafficication method based on the mixing ratio by means of category decomposition, Journal of Remote Sensing Society of Japan, 13, 3, 23-32, 1993.

[3] Kohei Arai and H.Chen, Unmixing method for hyperspectral data based on subspace method with learning process, Techninical Notes of the Science and Engineering Faculty of Saga University,, 35, 1, 41-46, 2006.

[4] Kohei Arai and Y.Terayama, Label Relaxation Using a Linear Mixture Model, International Journal of Remote Sensing, 13, 16, 3217-3227, 1992.

[5] Kohei Arai, Yasunori Terayama, Yoko Ueda, Masao Moriyama, Cloud coverage ratio estimations within a pixel by means of category decomposition, Journal of Japan Society of Phtogrammetro and Remote Sensing, 31, 5, 4-10, 1992.

[6] Kohei Arai, Non-linear mixture model of mixed pixels in remote sensing satellite images based on Monte Carlo simulation, Advances in Space Research, 41, 11, 1715-1723, 2008.

[7] Kohei Arai, Kakei Chen, Category decomposition of hyper spectral data analysis based on sub-space method with learning processes, Journal of Japan Society of Phtogrammetro and Remote Sensing, 45, 5, 23-31, 2006.

[8] Kohei Arai, Adjacency effect of layered clouds estimated with Monte-Carlo simulation, Advances in Space Research, Vol.29, No.19, 1807-1812, 2002.

[9] Ramachandran, Justice, Abrams(Edt.),Kohei Arai et al., Land Remote Sensing and Global Environmental Changes, Part-II, Sec.5: ASTER VNIR and SWIR Radiometric Calibration and Atmospheric Correction, 83-116, Springer 2010.

[10] Kohei Arai, Lecture Note for Remote Sensing, Morikita Publishing Inc., (Scattering), 2004.

[11] Kohei Arai, Fundamental Theory for Remote Sensing, Gakujutsu-Tosho Publishing Co., Ltd.,(Lambertian), 2001.

## AUTHORS PROFILE

**Kohei Arai,** He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission "A" of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.

# A Modified Feistel Cipher Involving Substitution, shifting of rows, mixing of columns, XOR operation with a Key and Shuffling

V.U.K Sastry
Dean R&D, Department of Computer Science and Engineering, Sreenidhi Institute of Science & Tech.
Hyderabad, India.

K. Anup Kumar
Associate Professor, Department of Computer Science and Engineering, SNIST,
Hyderabad, India

***Abstract*— In this paper, we have developed a modification to the Feistel cipher by taking the plaintext in the form of a pair of matrices and introducing a set of functions namely, substitute, shifting of rows, mixing of columns and XOR operation with a key. Further we have supplemented this process by using another function called shuffling at the end of each round of the iteration process. In this analysis, the cryptanalysis clearly indicates that the strength of the cipher is quite significant and this is achieved by the introduction of the aforementioned functions.**

***Keywords- encryption; decryption; cryptanalysis; avalanche effect; XOR operation.***

## I. INTRODUCTION

The study of the Feistel cipher has been a fascinating fundamental area in the development of block ciphers in cryptography. In the recent years, we have offered several modifications [1-4] to the classical Feistel cipher by taking the plaintext in the form of a pair of matrices. In all these investigations, we have made use of the multiplication with a single key matrix or the multiplication with a pair of key matrices as a fundamental tool in the development of the cipher. This is associated with the mod operation. Further, we have introduced some operations such as mixing, permutation, blending or shuffling in order to achieve confusion and diffusion, so that, the strength of the cipher becomes significant.

In the present investigation, our objective is to study a modification of the Feistel cipher, wherein we use the fundamental operations such as substitution, shifting of rows, mixing of columns, XOR operation and Shuffling. It may be noted here that the operations, substitution, shifting of rows and mixing of columns are very well utilized in Advanced Encryption Standard (AES) [5]. Our interest here is to develop a strong block cipher which exceeds, in strength, almost all the other ciphers available in the literature.

In what follows we present the plan of the paper. In section 2, we deal with the development of the cipher and present the flowcharts and algorithms required in this analysis. In section 3, we mention an illustration of the cipher and describe the avalanche effect. We study the cryptanalysis in section 4. Finally, in section 5, we discuss the computations and draw conclusions.

## II. DEVELOPMENT OF THE CIPHER

Consider a plaintext P containing $2m^2$ characters. On using the EBCIDIC code, the characters occurring in the plaintext can be represented in terms of decimal numbers wherein each number lies in [0 - 255]. Then, these numbers can be written in the form of a pair of square matrices $P_0$ and $Q_0$, wherein each one is of size m.

Let us consider a key matrix K, where K is a square matrix whose size is m.

The flowcharts depicting the encryption and the decryption are given below.

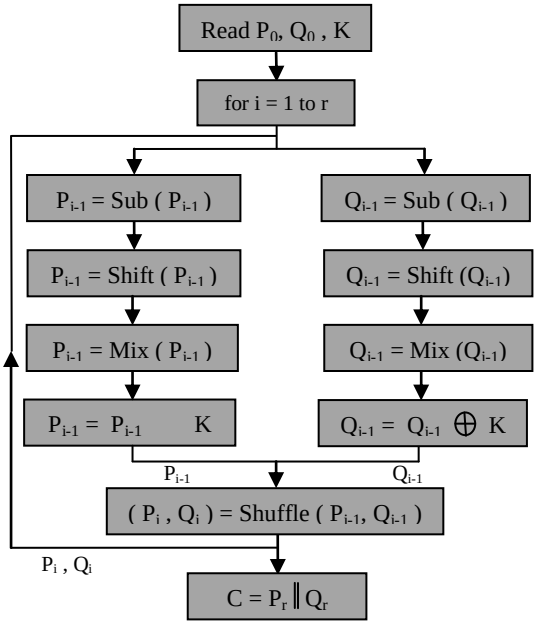



Fig 1. The process of Encryption

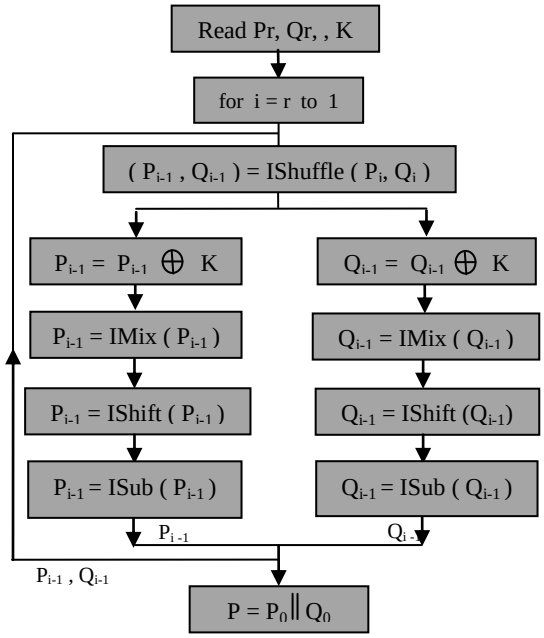



Fig 2. The process of Decryption

Now we write the algorithms for the processes encryption and decryption as given below.

*A. **Algorithm for Encryption***

1. Read P, K
2. $P_0$ = Left half of P.
3. $Q_0$ = Right half of P.
4. for i = 1 to r

begin

$P_{i-1}$ = Sub ($P_{i-1}$)

$P_{i-1}$ = Shift ($P_{i-1}$)

$P_{i-1}$ = Mix ($P_{i-1}$)

$P_{i-1}$ = $P_{i-1} \oplus_1$ K

$Q_{i-1}$ = Sub ($Q_{i-1}$)

$Q_{i-1}$ = Shift ($Q_{i-1}$)

$Q_{i-1}$ = Mix ($Q_{i-1}$)

$Q_{i-1}$ = $Q_{i-1} \oplus_1$ K

($P_i$ , $Q_i$) = Shuffle ( $P_{i-1}$, $Q_{i-1}$ )

end

5. C = $P_r \| Q_r$ /* ‖ represents concatenation */
6. Write(C)

*B. **Algorithm for Decryption***

1. Read C, K
2. $P_r$ = Left half of C.
3. $Q_r$ = Right half of C.
4. for i = r to 1

begin

($P_{i-1}$ , $Q_{i-1}$) = IShuffle ( $P_i$, $Q_i$ )

$P_{i-1}$ = $P_{i-1} \oplus_1$ K

$P_{i-1}$ = IMix ($P_{i-1}$)

$P_{i-1}$ = IShift ($P_{i-1}$)

$P_{i-1}$ = ISub ($P_{i-1}$)

$Q_{i-1}$ = $Q_{i-1} \oplus_1$ K

$Q_{i-1}$ = IMix ($Q_{i-1}$)

$Q_{i-1}$ = IShift ($Q_{i-1}$)

$Q_{i-1}$ = ISub ($Q_{i-1}$)

end

5. P = $P_0 \| Q_0$ /* ‖ represents concatenation */
6. Write (P)

Let us now explain the basic ideas underlying in the functions Sub ( ), Shift ( ), Mix ( ), used for substitution, shifting of rows, mixing of columns respectively.

Firstly, Let us focus our attention on the substitution process involved in the function Sub ( ).

Consider the EBCIDIC code which can be written in the form a matrix given by

E (i, j) = 16*(i-1) + (j–1), i = 1 to 16 and j = 1 to 16 (2.1)

All these numbers can be placed in the form of a table.

Let us arrange these numbers, which are lying in the interval [0-255], in a random manner.

We represent these numbers in the hexadecimal notation. All these numbers can be written in the form of a table given below (table 2).

In the encryption process, when we come across a number lying in [0-255], we will replace it by the corresponding number in the substitution table. For example, if we come across the number 70, in the process of encryption, this will be converted into hexadecimal number as 46. Then, 70 will be replaced by the number which is occurring in the $4^{th}$ row, $6^{th}$ column of the substitution table, i.e by 5A ( = 90 in decimal notation). This is the process of substitution. Keeping the EBCIDIC code matrix and the substitution table in view, we form the inverse substitution table which is given in Table 2.

The inverse substitution table will be utilized while carrying out the decryption process and it is denoted by function ISub ( ).

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 63 | 7C | 77 | 7B | F2 | 6B | 6F | C5 | 30 | 01 | 67 | 2B | FE | D7 | AB | 76 |
| 1 | CA | 82 | C9 | 7D | FA | 59 | 47 | F0 | AD | D4 | A2 | AF | 9C | A4 | 72 | C0 |
| 2 | B7 | FD | 93 | 26 | 36 | 3F | F7 | CC | 34 | A5 | E5 | F1 | 71 | D8 | 31 | 15 |
| 3 | 04 | C7 | 23 | C3 | 18 | 96 | 05 | 9A | 07 | 12 | 80 | E2 | EB | 27 | B2 | 75 |
| 4 | 09 | 83 | 2C | 1A | 1B | 6E | 5A | A0 | 52 | 3B | D6 | B3 | 29 | E3 | 2F | 84 |
| 5 | 53 | D1 | 00 | ED | 20 | FC | B1 | 5B | 6A | CB | BE | 39 | 4A | 4C | 58 | CF |
| 6 | D0 | EF | AA | FB | 43 | 4D | 33 | 85 | 45 | F9 | 02 | 7F | 50 | 3C | 9F | A8 |
| 7 | 51 | A3 | 40 | AF | 92 | 9D | 38 | F5 | BC | B6 | DA | 21 | 10 | FF | F3 | D2 |
| 8 | CD | 0C | 13 | EC | 5F | 97 | 44 | 17 | C4 | A7 | 7E | 3D | 64 | 5D | 19 | 73 |
| 9 | 60 | 81 | 4F | DC | 22 | 2A | 90 | 88 | 46 | EE | B8 | 14 | DE | 5E | 0B | DB |
| A | E0 | 32 | 3A | 0A | 49 | 06 | 24 | 5C | C2 | D3 | AC | 62 | 91 | 95 | E4 | 79 |
| B | E7 | C8 | 37 | 6D | 8D | D5 | 4E | A9 | 6C | 56 | F4 | 6A | 65 | 7A | AE | 08 |
| C | BA | 78 | 25 | 2E | 1C | A6 | B4 | C6 | E8 | DD | 74 | 1F | 4B | BD | 8B | 8A |
| D | 70 | 3E | B5 | 66 | 48 | 03 | F6 | 0E | 61 | 35 | 57 | B9 | 86 | C1 | 1D | 9E |
| E | E1 | F8 | 98 | 11 | 69 | D9 | 8E | 94 | 9B | 1E | 87 | E9 | CE | 55 | 28 | DF |
| F | 8C | A1 | 89 | 0D | BF | E6 | 42 | 68 | 41 | 99 | 2D | 0F | B0 | 54 | BB | 16 |

Table 1. Substitution Table

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 52 | 09 | 6A | D5 | 30 | 36 | A5 | 38 | BF | 40 | A3 | 9E | 81 | F3 | D7 | FB |
| 1 | 7C | E3 | 39 | 82 | 9B | 2F | FF | 87 | 34 | 8E | 43 | 44 | C4 | DE | F9 | CB |
| 2 | 54 | 7B | 94 | 32 | A6 | C2 | 23 | 3D | EE | 4C | 95 | 0B | 42 | FA | C3 | 4E |
| 3 | 08 | 2E | A1 | 66 | 28 | D9 | 24 | B2 | 76 | 5B | A2 | 49 | 6D | 8B | D1 | 25 |
| 4 | 72 | F8 | F6 | 64 | 86 | 68 | 98 | 16 | D4 | A4 | 5C | CC | 5D | 65 | B6 | 92 |
| 5 | 6C | 70 | 48 | 50 | FD | ED | B9 | DA | 5E | 15 | 46 | 57 | A7 | 8D | 9D | 84 |
| 6 | 90 | D8 | AB | 00 | 8C | BC | D3 | 0A | F7 | E4 | 58 | 05 | B8 | B3 | 45 | 06 |
| 7 | D0 | 2C | 1E | 8F | CA | 3F | 0F | 02 | C1 | AF | BD | 03 | 01 | 13 | 8A | 6B |
| 8 | 3A | 91 | 11 | 41 | 4F | 67 | DC | EA | 97 | F2 | CF | CE | F0 | B4 | E6 | 73 |
| 9 | 96 | AC | 74 | 22 | E7 | AD | 35 | 85 | E2 | F9 | 37 | E8 | 1C | 75 | DF | 6E |
| A | 47 | F1 | 1A | 71 | 1D | 29 | E5 | 89 | 6F | B7 | 62 | 0E | AA | 18 | BE | 1B |
| B | FC | 56 | 3E | 4B | C6 | D2 | 79 | 20 | 9A | DB | C0 | FE | 78 | CD | 5A | F4 |
| C | 1F | DD | A8 | 33 | 88 | 07 | C7 | 31 | B1 | 12 | 10 | 59 | 27 | 80 | EC | 5F |
| D | 60 | 51 | F7 | A9 | 19 | B5 | 4A | 0D | 2D | E5 | 7A | 9F | 93 | C9 | 9C | EF |
| E | A0 | E0 | 3B | 4D | AE | 2A | F5 | B0 | C8 | EB | BB | 3C | 83 | 53 | 99 | 61 |
| F | 17 | 2B | 04 | 7E | BA | 77 | D6 | 26 | E1 | 69 | 14 | 63 | 55 | 21 | 0C | 7D |

Table 2. Inverse Substitution Table

Now let us see the process of shifting involved in the function Shift ( ), during the encryption process we come across plaintext $P_{i-1}$ and $Q_{i-1}$ in the process of iteration. As $P_{i-1}$ is a square matrix of size m, it can be written in the form

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots\cdots\cdots & p_{1m} \\ p_{21} & p_{22} & p_{23} & \cdots\cdots\cdots & p_{2m} \\ p_{31} & p_{32} & p_{33} & \cdots\cdots\cdots & p_{3m} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ p_{m1} & p_{m2} & p_{m3} & \cdots\cdots\cdots & p_{mm} \end{bmatrix} \quad (2.4)$$

On converting each decimal number in (2.4) into its binary form, we get

$$\begin{bmatrix} p_{111}p_{112}\ldots\ldots p_{118} & p_{121}p_{122}\ldots\ldots p_{128} & \cdots\cdots & p_{1m1}p_{1m2}\ldots\ldots p_{1m8} \\ p_{111}p_{112}\ldots\ldots p_{118} & p_{121}p_{122}\ldots\ldots p_{128} & \cdots\cdots & p_{1m1}p_{1m2}\ldots\ldots p_{1m8} \\ p_{111}p_{112}\ldots\ldots p_{118} & p_{121}p_{122}\ldots\ldots p_{128} & \cdots\cdots & p_{1m1}p_{1m2}\ldots\ldots p_{1m8} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ p_{111}p_{112}\ldots\ldots p_{118} & p_{121}p_{122}\ldots\ldots p_{128} & \cdots\cdots & p_{1m1}p_{1m2}\ldots\ldots p_{1m8} \end{bmatrix} \quad (2.5)$$

Here each row contains 8m binary bits. In the process of shifting, we offer a right shift of 4 bits in the first row, 12 bits in the second row, 20 bits in the third row and in general,

4 + 8 * (i-1) bits right shift in the $i^{th}$ row.

This process is carried out till we exhaust all the rows. It may be noted here that IShift ( ) denotes the reverse process of Shift ( ).

In this, the binary bits are obviously given a left shift in an appropriate manner.

To have a clear insight into the mixing process denoted by the function Mix ( ), let us consider again the matrix $P_{i-1}$, which is represented in the form (2.5).

Let us restrict our attention only to a plaintext matrix, wherein, m=4. This can be written in the form given below

$$\begin{bmatrix} p_{111}\,p_{112}\ldots\ldots p_{118} & p_{121}\,p_{122}\ldots\ldots p_{128} \cdots\cdots & p_{141}\,p_{142}\ldots p_{148} \\ p_{211}\,p_{212}\ldots\ldots p_{218} & p_{221}\,p_{222}\ldots\ldots p_{228} \cdots\cdots & p_{241}\,p_{242}\ldots p_{248} \\ p_{311}\,p_{312}\ldots\ldots p_{318} & p_{321}\,p_{322}\ldots\ldots p_{328} \cdots\cdots & p_{341}\,p_{342}\ldots p_{348} \\ p_{411}\,p_{412}\ldots\ldots p_{418} & p_{421}\,p_{422}\ldots\ldots p_{428} \cdots\cdots & p_{441}\,p_{442}\ldots p_{448} \end{bmatrix} \quad (2.6)$$

This has 4 rows and 32 columns. On concatenating the binary bits of the $1^{st}$ column and the $17^{th}$ column we get a string of binary bits, which can be converted into a decimal number. This can be considered as new $p_{11}$.

On considering the binary bits of the $2^{nd}$ column and the $18^{th}$ column and concatenating them, we get another decimal number which will be called as $p_{12}$.

On adopting the same process till we exhaust all the columns taken in pairs, we get the decimal numbers which correspond to the other elements of the matrix written in the row wise order. Thus we have, the new plaintext matrix, which is obtained after the completion of mixing. Imix ( ) is the reverse process of Mix ( ).

For a detailed discussion of the function shuffle ( ), wherein we are shuffling the columns of two matrices, we refer to [4].

## III. Illustration Of The Cipher

Consider the plaintext given below

My dear young lady! We both are well qualified. You have done your B.Tech and I have completed my M.S, where is the problem! We can fly anywhere. Why your father and mother are not accepting our marriage. We both belong to the same cast, we both are farmers. What is the objection of your father and your mother, are they having any thinking regarding my financial status? We are having as much landed property as your father is having. My father and your father both are well trained seasonal politicians. I wonder why your father is not accepting and why your mother is not accepting. Our marriage must happen soon. Yours loving Mr.X (3.1)

Let us focus our attention on the first 32 characters of the plaintext. This is given by

My dear young lady! We both are

On using EBCIDIC code, we get the plaintext matrix P in the form

$$P = \begin{bmatrix} 77 & 121 & 32 & 100 & 101 & 97 & 114 & 32 \\ 121 & 111 & 117 & 110 & 103 & 32 & 108 & 97 \\ 100 & 121 & 33 & 32 & 87 & 101 & 32 & 98 \\ 111 & 116 & 104 & 32 & 97 & 114 & 101 & 32 \end{bmatrix} \quad (3.2)$$

This can be written in the form of a pair of matrices given by

$$P_0 = \begin{bmatrix} 77 & 121 & 32 & 100 \\ 121 & 111 & 117 & 110 \\ 100 & 121 & 33 & 32 \\ 111 & 116 & 104 & 32 \end{bmatrix} \quad (3.3)$$

and

$$Q_0 = \begin{bmatrix} 101 & 97 & 114 & 32 \\ 103 & 32 & 108 & 97 \\ 87 & 101 & 32 & 98 \\ 97 & 114 & 101 & 32 \end{bmatrix} \quad (3.4)$$

Let us take the key matrix K in the form

$$K = \begin{bmatrix} 45 & 128 & 192 & 53 \\ 133 & 200 & 150 & 16 \\ 100 & 150 & 33 & 120 \\ 13 & 189 & 164 & 55 \end{bmatrix} \quad (3.5)$$

On applying the encryption algorithm, given in section 2, we get the ciphertext C in the form

$$C = \begin{bmatrix} 51 & 145 & 164 & 146 & 108 & 237 & 147 & 173 \\ 155 & 18 & 82 & 72 & 85 & 155 & 19 & 71 \\ 182 & 102 & 90 & 237 & 150 & 142 & 218 & 60 \\ 11 & 150 & 219 & 226 & 237 & 177 & 36 & 100 \end{bmatrix} \quad (3.6)$$

On using the decryption algorithm on (3.6), we get back the original plaintext P given by (3.2).

Let us now study the avalanche effect which throws some light on the strength of the cipher.

On changing the first row, first column element of $P_0$, from 77 to 76, we get a 1 bit change in the plaintext. On applying the encryption algorithm on the modified plaintext, keeping up the key as it is, we get the ciphertext C in the form

$$C = \begin{bmatrix} 218 & 88 & 129 & 219 & 201 & 58 & 54 & 101 \\ 157 & 209 & 7 & 186 & 109 & 153 & 44 & 75 \\ 219 & 120 & 243 & 158 & 95 & 55 & 38 & 117 \\ 43 & 233 & 147 & 229 & 81 & 38 & 133 & 187 \end{bmatrix} . \quad (3.7)$$

On comparing (3.6) and (3.7), after converting them into their binary form, we notice that they differ by 128 bits out of 256 bits. This indicates that the cipher is quite good from the view point of its strength.

Let us now consider a one bit change in the key. This is achieved by changing first row, first column element of the key K, given by (3.5), from 45 to 44.

Now on using the modified key and applying the encryption algorithm, keeping the plaintext as it is, we get the cipher text C in the form

$$C = \begin{bmatrix} 79 & 149 & 68 & 154 & 22 & 239 & 105 & 98 \\ 232 & 131 & 221 & 63 & 57 & 229 & 243 & 114 \\ 103 & 82 & 190 & 152 & 14 & 222 & 73 & 209 \\ 179 & 44 & 237 & 153 & 44 & 75 & 219 & 120 \end{bmatrix} \quad (3.8)$$

Now on comparing (3.6) and (3.8), after converting both into their binary form, we find that these two ciphertexts differ by 134 bits out of 256 bits.

This also shows that, the strength of the cipher is expected to be significant.

## IV. CRYPTANALYSIS

In cryptography, determination of the strength of the cipher is a very important aspect. In the literature of cryptography, it is well known that the cryptanalysis can be carried out by the following approaches.

1. Ciphertext only attack ( Brute force attack )
2. Known plaintext attack
3. Chosen plaintext attack
4. Chosen ciphertext attack

As William Stallings [6] has pointed out that every cipher must be designed so that it withstands the first two attacks at least.

Let us now consider the brute force attack.

Here the key is containing $m^2$ decimal numbers. Thus the size of the key space

$$= 2^{8m^2}.$$

Let us suppose that, the time required for the computation of the cipher with one value of the key is $10^{-7}$ seconds. Then the time required for processing the cipher with all the possible values of the key in the key space is

$$\frac{(2)^{8m^2} \times 10^{-7}}{365 \times 24 \times 60 \times 60} = \frac{10^{(2.4)\,m^2 - 7}}{365 \times 24 \times 60 \times 60} = 3.12 \times 10^{(2.4)m^2 - 15} \text{ years}$$

This time is very large when m is greater than or equal to 3.

In our example as we have taken m=4, the attack on this cipher, by the brute force approach, is totally ruled out.

Let us now investigate the known plaintext attack. In this case, we know as many plaintext and ciphertext pairs as we require, making an attempt for breaking the cipher. In the light of the above information, we have as many pairs of P and C as we require.

If we take r=1, that is, if we confine our attention to a single round of the iteration process, then we have the relations connecting C and P as follows:

$$P_0 = \text{Sub}(P_0) \quad (4.1)$$
$$P_0 = \text{Shift}(P_0) \quad (4.2)$$
$$P_0 = \text{Mix}(P_0) \quad (4.3)$$
$$P_0 = P_0 \oplus K \quad (4.4)$$
$$Q_0 = \text{Sub}(Q_0) \quad (4.5)$$
$$Q_0 = \text{Shift}(Q_0) \quad (4.6)$$
$$Q_0 = \text{Mix}(Q_0) \quad (4.7)$$
$$Q_0 = Q_0 \oplus K \quad (4.8)$$
$$(P_1, Q_1) = \text{Shuffle}(P_0, Q_0) \quad (4.9)$$
$$C = P_1 \parallel Q_1 \quad (4.10)$$

In the known plaintext attack, we know $P_0$ and $Q_0$ corresponding to the initial stage. We also know the C obtained at the end.

As C is known to us, we can determine $P_1$ and $Q_1$ from (4.10)

On using the IShuffle ( ), on (4.9), we get the current $P_0$ and $Q_0$ which are occurring on the left hand side of (4.4) and (4.8). On using initial the $P_0$ and the Sub ( ), we get $P_0$ on the left hand side of (4.1). After that, on using shift ( ) on the available $P_0$, we get $P_0$ occurring on the Left hand side of (4.2). Then on using the function Mix( ) on the current $P_0$, we have the $P_0$ occurring on the left side of (4.3). Thus, we can readily determine the key K from (4.4). Hence this cipher can be broken by the known plaintext attack if we confine only to one step in the iteration process.

Let us now study the cipher when r = 2. Then the equations governing the cipher are (4.1) to (4.10) and the following

$$P_1 = \text{Sub}(P_1) \quad (4.11)$$
$$P_1 = \text{Shift}(P_1) \quad (4.12)$$
$$P_1 = \text{Mix}(P_1) \quad (4.13)$$
$$P_1 = P_1 \oplus K \quad (4.14)$$
$$Q_1 = \text{Sub}(Q_1) \quad (4.15)$$
$$Q_1 = \text{Shift}(Q_1) \quad (4.16)$$
$$Q_1 = \text{Mix}(Q_1) \quad (4.17)$$
$$Q_1 = Q_1 \oplus K \quad (4.18)$$
$$(P_2, Q_2) = \text{Shuffle}(P_1, Q_1) \quad (4.19)$$
$$C = P_2 \parallel Q_2 \quad (4.20)$$

In the known plaintext attack, we know C, obtained at the end of the iteration process, and the corresponding $P_0$ and $Q_0$, which are available at the very beginning of the iteration process.

As we know C, we can determine $P_2$ and $Q_2$ from (4.20). On using IShuffle on (4.19), we get $P_1$ and $Q_1$ which are occurring on the left side of (4.14) and (4.18). We cannot determine K as we do not know the $P_1$ and $Q_1$ occurring in the right hand side of (4.14) and (4.18). Here, we notice that, though $P_0$ and $Q_0$ are known to us, we cannot determine the $P_1$ and $Q_1$ which are occurring on the right hand side of (4.14) and (4.18), by starting at the beginning as the key K is occurring in (4.4) and (4.8). In the light of these facts, this cipher cannot be broken by the known plaintext attack, when we have confined to r=2. This shows that it is impossible to break the cipher by the known plaintext attack when we carry out all the sixteen rounds in the iteration.

Intuitively choosing a plaintext or ciphertext and determining the key or a function of the key is a formidable task in the case of this cipher.

From the above discussion we conclude that this cipher is not breakable by all the possible attacks that are available in cryptography.

## V. COMPUTATIONS AND CONCLUSIONS

In this investigation, we have offered a through modification in the Feistel cipher by taking the plaintext in the form of a pair of matrices, and by applying several procedures, namely, substitution, shifting, mixing, XORing with the key and shuffle operation. Each one of these procedures modifies the plaintext in a through manner and creates confusion and diffusion in the development of the cipher. The iteration process, which is the basic one in this cipher, supports all the above procedures in a strong way.

Here it may be noted that the substitution table generated in a random manner by using the numbers [0-255] is to be sent to the receiver by the sender.

The programs for encryption and decryption are written in C language.

The plaintext given in (3.1) is divided into 20 blocks, wherein each block is containing 32 characters. We have appended in the last block by adding 13 blank characters, so that it becomes a complete block. On applying the encryption algorithm given in section 2 we get the cipher text corresponding to the entire plaintext (excluding the first block for which the cipher text is already given in (3.6) ), in the form

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 212 | 111 | 166 | 213 | 179 | 183 | 219 | 102 | 51 | 84 | 223 | 38 | 165 | 45 | 198 | 253 |
| 244 | 153 | 37 | 69 | 150 | 119 | 82 | 206 | 223 | 122 | 100 | 147 | 82 | 145 | 190 | 142 |
| 122 | 45 | 157 | 190 | 115 | 140 | 161 | 154 | 229 | 63 | 77 | 179 | 44 | 237 | 243 | 158 |
| 140 | 154 | 148 | 153 | 53 | 41 | 110 | 76 | 146 | 115 | 202 | 111 | 223 | 77 | 50 | 100 |
| 147 | 158 | 94 | 147 | 126 | 250 | 105 | 153 | 103 | 121 | 34 | 63 | 71 | 62 | 155 | 102 |
| 51 | 93 | 211 | 211 | 35 | 125 | 54 | 173 | 157 | 186 | 100 | 149 | 22 | 94 | 115 | 140 |
| 161 | 154 | 229 | 63 | 77 | 179 | 44 | 237 | 243 | 158 | 140 | 154 | 148 | 153 | 53 | 41 |
| 110 | 55 | 38 | 73 | 81 | 237 | 201 | 146 | 84 | 89 | 103 | 121 | 34 | 63 | 71 | 62 |
| 155 | 102 | 51 | 93 | 211 | 211 | 14 | 113 | 148 | 51 | 92 | 228 | 201 | 42 | 61 | 185 |
| 79 | 211 | 108 | 203 | 59 | 124 | 231 | 142 | 242 | 68 | 126 | 142 | 140 | 154 | 148 | 153 |
| 53 | 41 | 110 | 76 | 146 | 115 | 218 | 100 | 201 | 39 | 60 | 189 | 38 | 253 | 244 | 211 |
| 50 | 206 | 242 | 68 | 126 | 142 | 125 | 54 | 204 | 102 | 187 | 167 | 166 | 70 | 250 | 109 |
| 91 | 59 | 124 | 231 | 25 | 67 | 53 | 202 | 126 | 155 | 102 | 89 | 219 | 231 | 61 | 25 |
| 53 | 41 | 14 | 113 | 148 | 51 | 92 | 228 | 201 | 42 | 61 | 185 | 79 | 211 | 108 | 203 |
| 59 | 124 | 231 | 142 | 242 | 68 | 126 | 142 | 140 | 154 | 148 | 153 | 53 | 41 | 110 | 76 |
| 146 | 115 | 218 | 100 | 201 | 39 | 60 | 189 | 38 | 253 | 244 | 211 | 50 | 206 | 242 | 68 |
| 126 | 142 | 125 | 54 | 204 | 102 | 187 | 167 | 166 | 91 | 81 | 190 | 155 | 86 | 206 | 223 |
| 217 | 140 | 219 | 103 | 172 | 102 | 143 | 209 | 207 | 108 | 198 | 109 | 70 | 250 | 109 | 91 |
| 59 | 125 | 182 | 99 | 53 | 77 | 242 | 106 | 82 | 220 | 111 | 223 | 73 | 146 | 84 | 89 |
| 103 | 117 | 44 | 237 | 247 | 166 | 73 | 53 | 41 | 27 | 232 | 231 | 162 | 217 | 219 | 231 |
| 56 | 202 | 25 | 174 | 83 | 244 | 219 | 50 | 206 | 223 | 57 | 232 | 201 | 169 | 73 | 147 |
| 82 | 150 | 228 | 201 | 39 | 60 | 166 | 253 | 244 | 211 | 38 | 73 | 57 | 229 | 233 | 55 |
| 239 | 166 | 153 | 150 | 119 | 146 | 35 | 244 | 115 | 233 | 182 | 99 | 53 | 221 | 61 | 50 |
| 55 | 211 | 106 | 217 | 219 | 166 | 73 | 81 | 101 | 231 | 56 | 202 | 25 | 174 | 83 | 244 |
| 219 | 50 | 206 | 223 | 57 | 232 | 201 | 169 | 73 | 147 | 82 | 150 | 227 | 114 | 100 | 149 |
| 30 | 220 | 153 | 37 | 69 | 150 | 119 | 146 | 37 | 221 | 61 | 48 | 231 | 25 | 67 | 53 |
| 206 | 76 | 146 | 163 | 219 | 148 | 253 | 54 | 204 | 179 | 183 | 206 | 120 | 239 | 36 | 71 |
| 232 | 232 | 201 | 169 | 73 | 147 | 82 | 150 | 228 | 201 | 39 | 61 | 166 | 76 | 146 | 115 |
| 203 | 210 | 111 | 223 | 77 | 51 | 44 | 239 | 36 | 71 | 232 | 231 | 211 | 108 | 198 | 107 |
| 186 | 122 | 100 | 111 | 166 | 213 | 179 | 183 | 206 | 113 | 148 | 51 | 92 | 167 | 233 | 182 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 101 | 157 | 190 | 115 | 209 | 147 | 82 | 144 | 231 | 25 | 67 | 53 | 206 | 76 | 146 | 163 |
| 219 | 148 | 253 | 54 | 204 | 179 | 183 | 206 | 120 | 239 | 36 | 71 | 232 | 232 | 201 | 169 |
| 73 | 147 | 82 | 150 | 228 | 201 | 39 | 61 | 166 | 76 | 146 | 115 | 203 | 210 | 111 | 223 |
| 77 | 51 | 44 | 239 | 36 | 86 | 233 | 233 | 150 | 212 | 111 | 166 | 213 | 179 | 183 | 206 |
| 203 | 206 | 113 | 148 | 51 | 92 | 167 | 233 | 182 | 101 | 157 | 190 | 115 | 209 | 147 | 82 |
| 147 | 38 | 165 | 45 | 198 | 228 | 201 | 42 | 61 | 185 | 50 | 74 | 139 | 44 | 239 | 36 |
| 218 | 182 | 118 | 250 | 198 | 104 | 253 | 26 | 93 | 211 | 211 | 14 | 113 | 148 | 51 | 92 |
| 228 | 201 | 42 | 61 | 185 | 79 | 211 | 108 | 203 | 59 | 124 | 231 | 142 | 242 | 68 | 126 |

The cryptanalysis, carried out in this investigation, clearly shows that this cipher is a strong one. This has become a very good cipher as we have taken the length of the plaintext as large as possible (2048 bits), and supported the encryption process with a good number of functions so that the plaintext undergoes a through transformation ( in each round of the iteration process) before it becomes the ciphertext. In this analysis, the substitution table generated in the random manner plays a very important role.

## REFERENCES

[1] V.U.K Sastry and K. Anup Kumar, " A Modified Feistel Cipher involving a key as a multiplicant on both the sides of the Plaintext matrix and supplemented with Mixing Permutation and XOR Operation", International Journal of Computer Technology and Applications ISSN: 2229-6093. Vol. 3, No.1, pp. 23-31, 2012.

[2] V.U.K Sastry and K. Anup Kumar, "A Modified Feistel Cipher Involving a Key as a Multiplicant on Both the Sides of the Plaintext Matrix and Supplemented with Mixing, Permutation, and Modular Arithmetic Addition", International Journal of Computer Technology and Applications ISSN: 2229-6093. Vol. 3, No.1, pp. 32-39, 2012.

[3] V.U.K Sastry and K. Anup Kumar, "A Modified Feistel Cipher Involving a Pair of Key Matrices, Supplemented with XOR Operation, and Blending of the Plaintext in each Round of the Iteration Process", International Journal of Computer Science and Information Technologies ISSN: 0975-9646. Vol. 3, No.1, pp. 3133-3141, 2012.

[4] V.U.K Sastry and K. Anup Kumar, "A Modified Feistel Cipher involving a pair of key matrices, Supplemented with Modular Arithmetic Addition and Shuffling of the plaintext in each round of the iteration process", International Journal of Computer Science and Information Technologies ISSN: 0975-9646. Vol. 3, No.1, pp. 3119-3128, 2012.

[5] Daemen J, and Rijmen V, "Rijndael, the Advanced Encryption Standard (AES)", Dr. Dobbs Journal, Vol. 26(3), pp. 137 -139, Mar 2001.

[6] William Stallings, Cryptography and Network Security, Principles and Practice, Third Edition, Pearson, 2003.

## AUTHORS PROFILE

**Dr. V. U. K. Sastry** is presently working as Professor in the Dept. of Computer Science and Engineering (CSE), Director (SCSI), Dean (R & D), SreeNidhi Institute of Science and Technology (SNIST), Hyderabad, India. He was Formerly Professor in IIT, Kharagpur, India and
Worked in IIT, Kharagpur during 963 – 1998. He guided 12 PhDs, and published more than 40 research papers in various international journals. His research interests are Network Security & Cryptography, Image Processing, Data Mining and Genetic Algorithms.

**Mr. K. Anup Kumar** is presently working as an Associate Professor in the Department of Computer Science and Engineering, SNIST, Hyderabad India. He obtained his B.Tech (CSE) degree from JNTU Hyderabad and his M.Tech (CSE) from Osmania University, Hyderabad. He is now pursuing his PhD from JNTU, Hyderabad, India, under the supervision of Dr. V.U.K. Sastry in the area of Information Security and Cryptography. He has 10 years of teaching experience and his interest in research area includes Cryptography,Steganography and Parallel Processing Systems.

# Automatic Association of Strahler's Order and Attributes with the Drainage System

Mohan P. Pradhan
Department of CSE,
SMIT
Sikkim India

M. K. Ghose
Department of CSE,
SMIT
Sikkim India

Yash R. Kharka
Department of CSE,
SMIT
Sikkim India

***Abstract*— A typical drainage pattern is an arrangement of river segment in a drainage basin and has several contributing identifiable features such as leaf segments, intermediate segments and bifurcations. In studies related to morphological assessment of drainage pattern for estimating channel capacity, length, bifurcation ratio and contribution of segments to the main stream, association of order with the identified segment and creation of attribute repository plays a pivotal role. Strahler's (1952) proposed an ordering technique that categories the identified stream segments into different classes based on their significance and contribution to the drainage pattern. This work aims at implementation of procedures that efficiently associates order with the identified segments and creates a repository that stores the attributes and estimates of different segments automatically. Implementation of such techniques not only reduces both time and effort as compared to that of manual procedures, it also improves the confidence and reliability of the results.**

***Keywords- Stream; digitization; Strahler's order.***

## I. INTRODUCTION

A drainage pattern pertaining to a terrain is a mesh of interconnected streams. This mesh may be of different types such as Dendritic, Trellied or Lattice, Radial or Concentric, Parallel, Rectangular, Deranged, Centripetal and Violent. The formation of mesh depends upon the morphological aspect of the terrain the drainage system is subjected to such as slope, varied resistance of rocks and its geological and geomorphological past[1]. In studies related to drainage system and its effects on the terrain demands classification of the system into identifiable classes or identifiable orders.

There are different systems for ordering drainage pattern designed by Horton (1945), Strahler (1952), Scheideggar (1965) and Shreve (1967) for associating order with stream segment in drainage segment.[1]

The order of a stream segment depends upon the order of its tributaries. The point where tributaries meet is called as a junction. In studies related to drainage system streams are also referred to as links.

Links are typically of two types namely internal and external. Links are classified based on whether they do or do not have tributaries. Link that stretches from source to a junction is referred to as external links whereas link that stretches from on junction to another is referred to as internal links. Each of these identified streams have their own order, length, channel capacity and bifurcation ratio.

In order to associate order with the stream segment in a drainage pattern either traditional tedious manual technique can be used or a process can be designed based on certain criteria or knowledge of any ordering techniques.

Traditional techniques for associating order with segment involve manual digitization, manual association of order and attributes etc. This demands greater effort, time and cost investment. Design of automated procedure for the same would greatly reduce investments; in addition it enhances quality and reliability of the results.

This proposed work automatically extracts segments from a drainage pattern, associates order and also estimates various attributes related to the same. This work relies on the concept of Strahler's ordering technique for performing qualitative and quantitative assessment of the drainage system. In order to minimize the time and space complexity of identification and storing attributes, two important procedures are used.

On analyzing the drainage pattern it was observed that the contributing tributaries always converges inwards to form a segment of higher order and these converging streams are always located towards the bound. So, in order to identify these streams traditional row column traversal technique proves ineffective and would consume more time.

To minimize the time required for determining segments an efficient and effective spiral navigation technique is used [2]. In order to store the indentified segments of variable length an efficient 2D data structure (jazzed array) is used to store the detail that optimally utilizes memory space. The proposed schema is shown in figure 1 below.

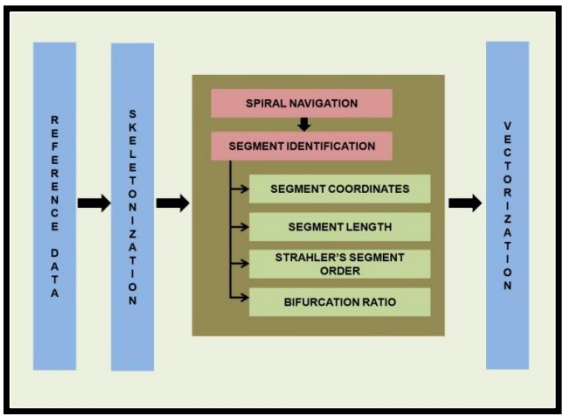



Figure 1. Schema for stream order & attribute association.

## II. STREAM ORDERING

In Strahler's (1952) system [3]

- if a stream has no contributing tributaries, then order of the stream is 1
- else if the stream has more than on tributaries, and their orders are i and j then
  - if i==j then the order of the resulting stream will be i+1 or j+1
  - else if i<j then the order of the resulting stream will be j
  - else if i>j then the order of the resulting stream will be i

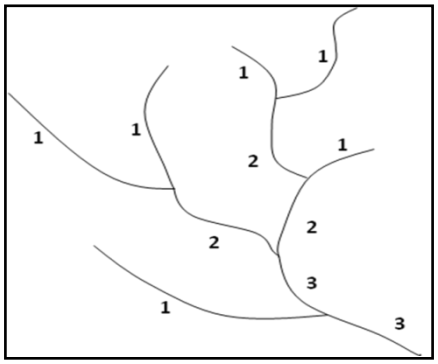



Figure 2. Strahler's Ordering Scheme

Two streams with same order *i* unite to give a stream of order *i+1* and if the streams of different order unite the new stream retains the order of the highest order stream.

## III. RELATED WORK

Stream ordering has wide range of application in hydrological studies for e.g. run-off modeling, fisheries etc. Association of stream ordering is tedious and time consuming task, thus method for determining stream order automatically for various drainage network topologies plays a crucial role in various studies related to drainage pattern.

Andy Ward et al [4] in their study related to stream classification has expressed that Stream order plays significant role in determining the expected ecological function of drainage system. Stream order is extensively used in River Continuum Concept RCC which is used for classifying, describing flowing water and classification of sections of waters based on presence of indicator organisms and Flood Pulse Concept (FPC), a concept that describes the interaction between water and land or hydrological conditions takes into basis the orientation of the stream in a drainage pattern for analysis.

Dawson et al [5] have developed an automated computer based extraction procedure on GIS system which helps in classification of rivers in Britain.

Many researchers have used terrain analysis of a DEM for extracting drainage pattern and associating attributes with each stream using Strahler's ordering technique. P. Venkatachalam et al. [5] have suggested a procedure for delineating drainage and watersheds from DEM of a terrain. David et al. [7] have used DEM as source for automatic delineation of flow path, sub-watershed and flow networks for hydrographic modeling. Storey and Wadhwa [8] have also estimated the length of each stream using terrain analysis of DEM based on LiDAR survey to produce stream channel maps. Ejstrud [9] have also constructed digital dataset of streams, lakes and wetlands using DEMS.

Alper Sen et al. [10] in their study related to drainage pattern used k-means clustering method for grouping rivers into different categories for creating reduced scale map for a drainage system taking into consideration various river attributes.

In spite of wide hydrological application, very few works has been carried for automatic digitization and association of order.

## IV. METHODOLOGY USED

A typical stream segment is represented by (Start, End, Length, Order and Bifurcation Ratio). Start represents starting coordinates of a stream segment. End represents ending coordinates of a stream segment. Length represents Length of the stream. Order represents the order of the segment. Bifurcation ratio is the number of streams of order u divided by number of streams of order (u-1).

### *A. Data Traversal Procedure*

In order to decrease the total amount of time required for determining the peripheral stream this work uses an efficient traversal scheme based on spiral navigation [1] rather than the traditional row column approach for navigation.

The peripheral streams that converge to form a main stream are often oriented around the main stream, so to identify these contributing stream spiral traversal proves efficient. The traversal process in case of spiral traversal may end at any step between 0 to $n^2$ against $n^2$ in case of traditional row column approach.

This traversal scheme exhaust the values in the data set in spiral manner either in a clockwise direction or an anti-clock wise direction, in order to ensure that the procedure terminates at a single point or coordinate the dataset has to be odd order so, in case if the dimension is not odd then row or column of non-significant values are added.

*B. Procedure for Determination of segment initiation point*

The implemented process takes in skeletonized binary image (0 (non-significant) or 1(significant)) of the drainage system for performing the classification process.

While traversing the data set in a spiral manner two possible types of values might be encountered either 0 or 1.

- On encountering 0 (insignificant value) the spiral traversal continues.
- On encountering 1 (significant value) the spiral traversal temporarily stop and either of the two actions are performed.
    - If the encountered 1 is in previously visited coordinate, then spiral traversal continues.
    - If the encountered 1 is an unvisited coordinate, then the spiral navigation is temporarily put to halt and segment navigation procedure is initiated.
- On completing segment navigation procedure the spiral traversal process resumes.

The process is repeated until all external links are exhausted.

*C. Segment Navigation Procedure*

This procedure aims at determining various segments, their starting coordinate, intermediate coordinates and terminating coordinate. It also determines length of the identified segment and associates order for an identified stream.

- *Identification of leaf segments*

On encountering a significant unvisited value during spiral navigation, the significant value is traversed until a junction is not encountered. On encountering a junction the segment navigation process stops and the navigation status is saved in a data structure.

The traversal status includes information such as starting coordinate, intermediate coordinates and terminating coordinates along with the length of the segment.

In order to prevent misinterpretation of traversed coordinates a status variable is maintained of equal dimensions as the actual data. As and when a coordinate value is visited the status of that coordinate is changed to visited (1 for visited and 0 for not visited).

The dataset is traversed until and unless all the leaf segments are determined. The traversed leaf segments have variable length, so the process demanded an efficient data management scheme that utilizes the memory efficiently, so in this work a jazzed data structure is implemented in order to store the segment information.

The individual rows in the jazzed data structure represent a segment and store the coordinate points for the segment. Additional information such as segment coordinate count is also maintained for determining length. Every leaf stream encountered is assigned a Strahler's order 1.

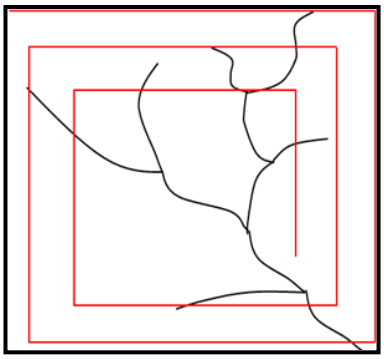


Figure 3. Spiral Navigation Scheme

- *Identification of non leaf segments*

Upon completing identification of leaf segments, the process for identifying non leaf segment is initiated. For determining the non leaf segment, spiral traversal is not used rather in order to reduce the complexity, the initially created jazzed structure is taken as input. The entries in the jazzed structure are scanned for determining junction. Junction is determined by determining entries that terminate at same coordinate. All such junctions are identified. For each identified junction segment navigation is done in order to identify second order streams and their status is saved in the data structure.

In order to determine $i^{th}$ order streams, all entries related to $(i-1)^{th}$ order streams in the jazzed structure are taken into consideration.

This process terminates when there are no junctions left to be extended.

- *Procedure for Strahler's ordering*

If in case

i. the terminating coordinates in the jazzed structure are same and their orders are same (say *i*) then the order of the segment extending from their intersection will be (*i+1*)

ii. the terminating coordinates in the jazzed structure are same and their orders are different then the order of the segment will be the highest of the orders

This process is repeated for all orders until and unless all the streams are not exhausted.

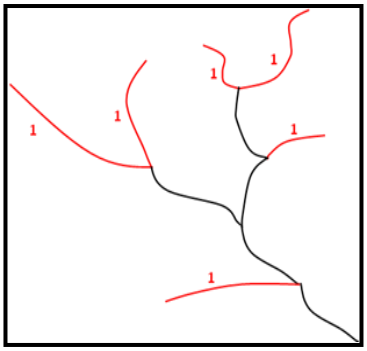



Figure 4. Identification of 1st Order Streams using spiral traversal

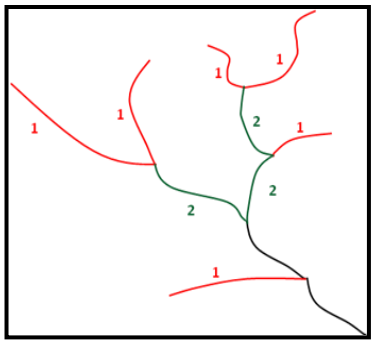



Figure 5. Identification of 2nd Order Stream using junction extension

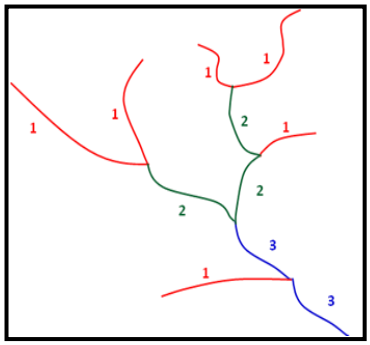



Figure 6. Identification of 3nd Order Stream using junction extension

## V. Results and Discussion

The skeletonized images of the drainage pattern were taken into consideration for evaluation the performance of the techniques.

The skeletonized image was processed in order to determine the number of segments, coordinate points, length, and order and bifurcation ratio. The input and the output image along with the various attributes are represented below in table 1.

## VI. Conclusion

This works aim at developing an automatic procedure for digitizing and vectoring drainage pattern. Implementation of such process tremendously reduces the amount of effort and time required in order to digitize and classify segments in a drainage pattern. In addition to classification it also associates attributes with the identified segments of drainage pattern.

## References

[1] Garde R J, River Morphology, 2nd edition, New Age International Publishers Ltd.-New Delhi, pp 14-15.

[2] Mohan P. Pradhan, M.K. Ghose, "Automatic Association of Stream Order for Vector Hydrograph Using Spiral Traversal Technique", IOSR Journal of Computer Enginering, Volume 1, Issue 5, pp 09-12, 2012.

[3] Strahlers A. N., "Quantitative Analysis of Watershed Geomorphology", American Geophysical Union Transactions, Volume 38, Pages 912-920, 1957.

[4] Andy Ward, Jessica L. D'Ambrosio, Dan Mecklenburg, "Stream Classification", Environmental hydrology, CRC Press.

[5] F.H. Dawson, D.D. Hornby, J. Hilton, "A method for the automated extraction of environmental variables to help the classification of rivers in Britain", Special Issue: Sustainable River Basin Management in the UK: Needs and Opportunities, Volume 12, Issue 4, pages 391–403, July/August 2002.

[6] P. Venkatachalam et al.," Automatic delineation of watersheds for Hydrological Applications", Proc. of 22nd Asian Conference on Remote Sensing, 2011.

[7] David Tarboton, "Terrain Analysis Using Digital Elevation Models in Hydrology", 23rd ESRI International Users Conference, 2003, San Diego, California.

[8] Storey, R., Wadhwa, S. , "An Assessment of the Lengths of Permanent, Intermittent and Ephemeral Streams in the Auckland Region", Prepared by NIWA for Auckland Regional Council. Auckland Regional Council Technical Report 2009/028, 2009.

[9] Bo Ejstrud, "Reconstructing drainage networks", [available online: web.sdu.dk/ejstrud/forskning/GIS/Arbejdspapir_Streams_Lakes.pdf ]

[10] Alper Sen, Turkay Gokgoz, "Clustering Approaches for Hydrographic Generalization", GIS Ostrava 2012 - Surface Models for Geosciences, 2012.

## Authors Profile

**Mohan P Pradhan**, He received B. Tech, M. Tech degrees in 2006 and 2009 respectively from Sikkim Manipal University, Sikkim, India. He is currently persuing his Ph. D for Sikkim Manipal University in the field of automatic digitization for GIS applications. He is currently working as Assistant Professor in department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Sikkim Manipal University since August 2006. His research interest includes RS and GIS, Algorithms, Image Processing and Software Engineering.

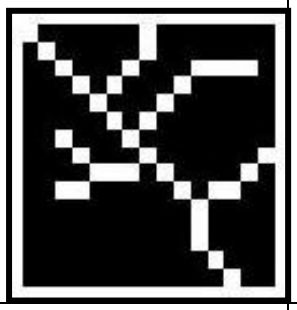
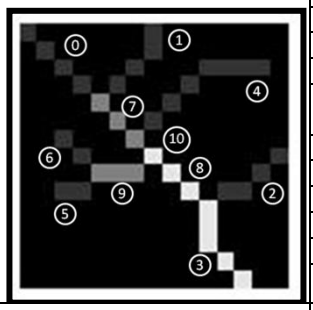
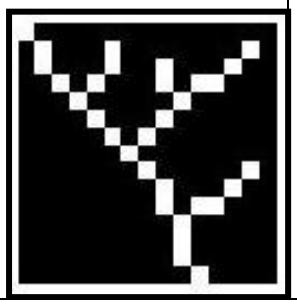
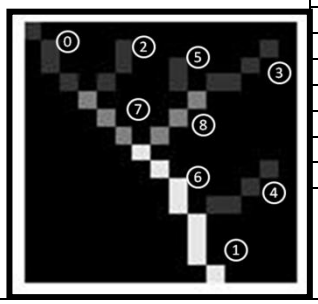
**Professor (Dr.) M. K. Ghose** is the Dean (R & D), SMIT and Professor & Head of the Department of Computer Science & Engineering at Sikkim Manipal Institute of Technology, Mazitar, Sikkim, India since June, 2006. Prior to this, Dr. Ghose worked in the internationally reputed R & D organization ISRO – during 1981 to 1994 at Vikram Sarabhai Space Centre, ISRO, Trivandrum in the areas of Mission Simulation and Quality & Reliability Analysis of ISRO Launch vehicles and Satellite systems and during 1995 to 2006 at Regional Remote Sensing Service Centre, ISRO, IIT Campus, Kharagpur in the areas of RS & GIS techniques for the natural resources management. He was also associated with Regional Engg. College (NIT), Silchar (1979 – 1981) as Teaching Asst. and Assam Central University, Silchar as COE and HOD of Computer Science Department (1997-2000).

His areas of research interest are Data Mining, Simulation & Modelling, Network, Sensor Network, Information Security, Optimization & Genetic Algorithm, Digital Image processing, Remote Sensing & GIS and Software Engineering.

**Yash R Kharka**, he received his B. Tech in Computer Science from Sikkim Manipal University, Sikkim, India, 2011. He is currently working as Assistant Systems Engineer at Tata Consultancy Services.

His research interest includes Algorithms, Networking and Image Processing.

TABLE I. DIGITIZATION OF DRAINAGE PATTERN USING STRAHLERS ORDERING TECHNIQUE AND ATTRIBUTE ASSOCIATION

| Input | Output | Segment | Points | Count | Order | Bi. R |
|---|---|---|---|---|---|---|
|  |  | 0 | (0,0)(1,1)(2,2)(3,3)(4,4) | 5 | 1 | 2.000 |
| | | 1 | (0,7) (1,7) (2,6)(3,5) (4,4) | 5 | 1 | 2.000 |
| | | 2 | (7,14)(8,13)(9,12)(9,11) (10,10) | 5 | 1 | 2.000 |
| | | 3 | (14,12)(13,11)(12,10) (11,10)(10,10) | 5 | 3 | - |
| | | 4 | (2,13)(2,12)(2,11)(2,10) (3,9) (4,8)(5,7)(6,6) | 8 | 1 | 2.000 |
| | | 5 | (9,2) (9,3) (8,4) | 3 | 1 | 2.000 |
| | | 6 | (6,2) (7,3) (8,4) | 3 | 1 | 2.000 |
| | | 7 | (4,4) (5,5) (6,6) | 3 | 2 | 1.500 |
| | | 8 | (10,10) (9,9) (8,8) (7,7) | 4 | 3 | - |
| | | 9 | (8,4) (8,5) (8,6) (7,7) | 4 | 2 | 1.500 |
| | | 10 | (6,6) (7,7) | 2 | 2 | 1.500 |
|  |  | 0 | (0,0)(1,1)(2,1)(3,2)(4,3) | 5 | 1 | 2.500 |
| | | 1 | (14,10)(13,9)(12,9)(11,9) | 4 | 3 | - |
| | | 2 | (1,5)(2,5)(3,4)(4,3) | 4 | 1 | 2.500 |
| | | 3 | (1,13)(2,12)(3,11)(3,10)(4,9) | 5 | 1 | 2.500 |
| | | 4 | (8,13)(9,12)(10,11)(10,10)(11,9) | 5 | 1 | 2.500 |
| | | 5 | (2,8)(3,8)(4,9) | 3 | 1 | 2.500 |
| | | 6 | (11,9)(10,8)(9,8)(8,7)(7,6) | 5 | 3 | - |
| | | 7 | (4,3)(5,4)(6,5)(7,6) | 4 | 2 | 1.000 |
| | | 8 | (4,9)(5,8)(6,7)(7,6) | 4 | 2 | 1.000 |

# Performance model to predict overall defect density

Dr. J. Venkatesh
Associate Professor, School of Management Studies
Anna University: Chennai 600025, Regional Office Coimbatore
Jothipuram Post, Coimbatore – 641 047, Tamilnadu, India.

Mr. Priyesh Cherurveettil
Ph.D Part Time Research Scholar, School of Management Studies, Anna University: Chennai 600025
Regional Office: Coimbatore, Jothipuram Post, Coimbatore – 641 047, Tamilnadu, India.

Mrs. Thenmozhi. S
Assistant Professor
Department of Computer Applications
Gnanamani College of Technology,
AK Sumuthiram, Namakkal District. Tamilnadu, India.

Dr. Balasubramanie. P
Professor
Department of Computer Science & Engineering
Kongu Engineering College
Perundurai, Erode - 638 052.
Tamilnadu, India

***Abstract*— Management by metrics is the expectation from the IT service providers to stay as a differentiator. Given a project, the associated parameters and dynamics, the behaviour and outcome need to be predicted. There is lot of focus on the end state and in minimizing defect leakage as much as possible. In most of the cases, the actions taken are re-active. It is too late in the life cycle. Root cause analysis and corrective actions can be implemented only to the benefit of the next project. The focus has to shift left, towards the execution phase than waiting for lessons to be learnt post the implementation. How do we pro-actively predict defect metrics and have a preventive action plan in place. This paper illustrates the process performance model to predict overall defect density based on data from projects in an organization.**

***Keywords- process; performance; defect density; metrics.***

## I. INTRODUCTION TO OVERALL DEFECT DENSITY

Number of defects leaked into production is a key metric that IT service providers will track month on month and intend to show a downward trend to their clients. Number of defects as a measure alone might not make sense, its relationship with size or effort is important.

For example, if we have 40 defects leaked in January, 30 in February and 5 in March, it doesn't mean that there is a downward trend. 30 defects could be for a project effort of 1000 hrs whereas 5 defects could be for a project size of 50 hrs. Hence, the metric that need to be closely tracked is overall defect density, number of defects leaked against the project effort.

## II. PROCESS PERFORMANCE MODELS IN ORGANIZATIONS

Process Performance Models (PPM) is probabilistic, statistical and simulative in nature. It can predict interim and final outcome, it is a proactive measure of tracking the end goal instead of a reactive one. It can model the variation of factor and help us understand the predicted range or the variation of its outcomes. Mid-course correction can be made to achieve desired outcome. Interestingly, PPMs enable "What-if" analysis for project planning, dynamic re-planning and problem resolution during project execution. We can run "what if" exercises holding one or more values constant. We can see the effect of tradeoffs between schedule, effort, defects, staff and functionality.

CMMI Dev 1.2 predominantly focuses on development and enhancement type of projects. CMMI for Services focuses on production support or maintenance type of projects. At an organization level, the respective Software Engineering Process Group (SEPG) develops few standard process performance models. These models are developed based on the data gathered from different types of projects within the organization. The organization focuses on three to four key models; typically they are around defect density, productivity, and schedule variance. The important step is to agree to standard definitions for the entities and their measured attributes. When we use terms like defect, productivity, size, and even project, different teams will tend to interpret in their own context. Hence it is important to have a common definition.

Managers should have a good understanding of process performance models and should use it to pro-actively manage the customer needs. Project Managers will check the availability of organizational developed process performance models with respect to the project objectives and quality and process performance objectives. During planning phase, managers are expected to implement process models to manage the objectives. Project manager has to define the project objective and consider the organization objective if there is no client objective defined. The specification limit of the objective is derived from process performance baseline arrived at the organization level. Project Manager is expected to provide the values for the controllable factors in process model prediction section. The prediction intervals are automatically generated based on the values provided

## III. VARIABLES ASSOCIATED IN PREDICTION MODEL

Based on brain storming session with the project team in the organization the different parameters that influence overall defect density were looked at. The team shortlisted following factors to start with, domain experience, technical experience, defects identified during design and coding phase, overall review efficiency and usage of tools. Operational definitions for these parameters were baseline and data was collected from projects in a particular account against these parameters. Linear regression was performed against the data to find out which are the key variables that influences the overall defect density.

After many trial and error methods the below three variables were established as the x factors.

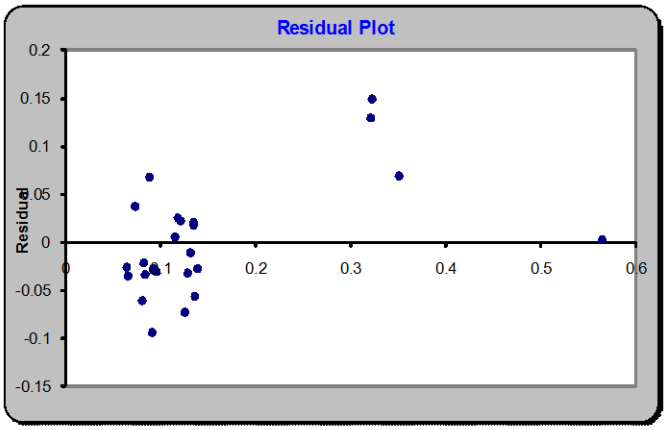
1. Y - Overall defect density – No of defects identified in the entire life cycle of the project against total effort for the project
2. X1 - Technical experience – Average technical experience of the team, in person months
3. X2 – DDD - Design Defect Density - Defects attributed to design identified during design review against effort spent for design
4. X3 – CDD – Coding Defect Density - Defects attributed to coding, identified during code review against effort spent for coding.

## IV. DEFECT DENSITY – REGRESSION EQUATION

The project data collated for the x and y factors are as shown in the Table 3.1. Data points from 25 projects in an organization were collected and considered for analysis. Projects factored in were similar in nature.

| Y | X1 | X2 | X3 |
|---|---|---|---|
| Overall Defect Density | Technical Experience (in months) | Design Defect Density DDD | Coding Defect Density CDD |
| 0.092 | 36 | 1.813 | 0.231 |
| 0.093 | 31 | 0.158 | 0.052 |
| 0.095 | 35 | 0.258 | 0.140 |
| 0.114 | 35 | 0.044 | 0.083 |
| 0.120 | 47 | 0.425 | 0.192 |
| 0.119 | 42 | 0.458 | 0.069 |
| 0.128 | 33 | 0.510 | 0.250 |
| 0.134 | 40 | 0.520 | 0.125 |
| 0.126 | 34 | 1.525 | 0.310 |
| 0.139 | 22 | 0.650 | 0.055 |
| 0.093 | 30 | 0.058 | 0.057 |
| 0.064 | 38 | 0.143 | 0.022 |
| 0.066 | 44 | 0.035 | 0.213 |
| 0.074 | 54 | 0.079 | 0.051 |
| 0.080 | 41 | 0.390 | 0.310 |
| 0.083 | 44 | 0.090 | 0.290 |
| 0.131 | 33 | 0.540 | 0.150 |
| 0.134 | 39 | 0.510 | 0.125 |
| 0.136 | 36 | 1.625 | 0.300 |
| 0.322 | 12 | 0.026 | 0.004 |
| 0.565 | 20 | 0.558 | 2.125 |
| 0.350 | 50 | 5.000 | 0.500 |
| 0.089 | 59 | 0.014 | 0.071 |
| 0.083 | 58 | 0.435 | 0.424 |
| 0.320 | 38 | 0.500 | 0.500 |

Table 3.1 – Project data values



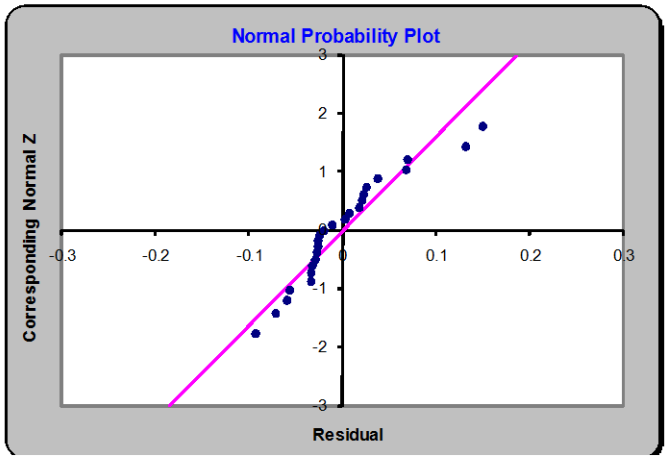




Figure 3.1 – Residual Plot

Mirror pattern is not found in Figure 3.1, Residual Plot and hence no heteroscedasticity is found. The normal probability plot is approximately linear. This would indicate that the normality assumption for the errors has not been violated.

Looking at the p value, since it is 0.0003 which is < 0.05, null hypothesis is not valid, which means the variables selected have an impact to overall defect density

| Intercept | Technical Experience (in months) | Design Defect Density DDD | Coding Defect Density CDD |
|---|---|---|---|
| 0.21375 | -0.0035 | 0.0295 | 0.1892 |

Table 3.2 – Regression Equation

As shown in Table 3.2, technical experience has a negative influence on overall defect density. As the team's technical experience increases the overall defect density is reduced. The influence of Design Defect Density and Coding Defect Density is positive.

This means that when the values of Design Defect Density and Coding Defect Density are low the overall defect density will be low and vice versa.

## V. DEFECT DENSITY – COMPONENTS OF PREDICTION MODEL

Based on project data analyzed it is evident that overall defect density is critically influenced by design sub process, code sub process and technical experience. Organization Metrics group would help with the baseline data for these metrics. For code sub process and CDD metric, based on the technology (Java, .Net) and review type (manual, tool, FxCop) the baseline values can be tabled. Organization Metrics group will share the baseline values for these combinations. Baseline values will include lower specification limit (LSL), goal and upper specification limit (USL).

The same can be gathered for Design Defect Density as well. Project team needs to choose the process that they would be following for coding or design sub process. Based on the composition of sub process, project goal for DDD and CDD would be calculated. It is also important for the project team to justify why they have gone with a particular sub process and the rationale. Table 4.1 gives the sub process performance baseline for Coding Defect Density and Design Defect Density. The values are represented by A1, A2, A3 and so on. Organization Metrics team would have the actual baseline values for LSL, Goal and USL for these identified metrics. Based on the current project context, the type of technology and review type will selected as shown in Table 4.2, Selected Sub process performance baseline.

| **Sub process** | **Metric** | **Technology** | **LSL** | **Goal** | **USL** |
|---|---|---|---|---|---|
| Code Review | CDD | .Net | A1 | A2 | A3 |
| Code Review | CDD | Java | C1 | C2 | C3 |
| Code Review | CDD | J2ee | D1 | D2 | D3 |
| Design Review | DDD | .Net | E1 | E2 | E3 |
| Design Review | DDD | Java | F1 | F2 | F3 |

Table 4.1 - Sub process performance baseline

| **Sub process** | **Metric** | **Technology** | **Goal** | **Comments** |
|---|---|---|---|---|
| Code Review | CDD | .Net | A2 | |
| Design Review | DDD | .Net | E2 | |

Table 4.2 – Selected Sub process performance baseline

## VI. DEFECT DENSITY MODEL – PRACTICAL USAGE

One of the current releases in design phase was considered for the practical usage of this model. The below steps will illustrate the prediction model.

1. X factors baseline data was used as input. Technical experience goal is 28 months with LSL of 14 months and USL of 48 months
2. Sub process performance baseline data was reviewed and based on the current project context the below selection was made. As shown in Table 5.1 the sub process code and design review were selected. Based on the project usage, .net technology was selected. The goal, upper specification limit and lower specification limit are chosen from organization baseline report.

| **Sub process** | **Metric** | **Technology** | **LSL** | **Goal** | **USL** |
|---|---|---|---|---|---|
| Code Review | Coding Defect Density (defects/ person day) | .Net | 0.10 | **0.38** | 1.5 |
| Design Review | Design Defect Density (defects/ person day) | .Net | 0.15 | **0.22** | 0.45 |

Table 5.1 – Selected Sub process

3. Update the actual technical experience in the team and predict the overall defect density. The predicted value is based on Monte Carlo simulation.

| Average Technical Experience (in months) | DRDDE (Defects/ Personday) | CRDDE (Defects/ Personday) | Predicted Overall Defect Density |
|---|---|---|---|
| 24 | 0.22 | 0.38 | 0.21 |

Table 5.2 – Predicted Overall Defect Density

4. The data was compared against goal. The client goal for overall defect density is 0.15 whereas the predicted value is 0.21.
5. Perform what-if analysis and look at various combinations of the x factors and analyze the predicted overall defect density based on these factors. Based on the project experience choose the one which is close to reality. In this case the option of 24months technical experience, DDD 0.22 and CDD 0.38 was considered as the closest option.
6. List down the assumptions considered when the final decision is made on the values of x factors. Ensure all the relevant assumptions are documented. As need be, the assumptions need to be validated with the relevant stakeholders before the baseline process.
7. Understand the deviation and prepare preventive action plan

| Expected client overall defect density | Predicted overall defect density | Preventive Action | Responsibility |
|---|---|---|---|
| 0.15 | 0.21 | List down the top three preventive action items | Project Manager |

Table 5.3 – Deviation Analysis

8. Estimated effort in person days for the project is 1000 person days. Based on the predicted defect density and organizations standard effort distribution across phases, the defects that could be injected at each phase are predicted as show in Table 5.4

| Phases | Expected Injection | Actual Defects Captured | Remarks |
|---|---|---|---|
| Requirements | 12 | 10 | |
| Design | 38 | 28 | |
| Coding | 70 | | |
| Unit Testing | 62 | | |
| System Testing | 108 | | |

Table 5.4 – Predicted-Actual Defects phase wise

9. Based on the actual data collated, keep updating Table 5.4 to compare the expected and actual defects captured. Based on the actual value in each phases, the predicted value for next phases are accordingly impacted. If there any specific inputs or considerations on the actual values, those are highlighted in the remarks column.
10. Prepare the detailed defect prevention plan. Against each phase, list down the defect type, defect cause, root cause, preventive action planned, responsible person, target date and the status. Defect types could be incorrect functionality or missing functionality or incorrect user interface or missing user interface. Defect causes could be lack of knowledge, missing information or incidental. Root cause should be as detailed as possible to plan for preventive and corrective action. 5-Why analysis can be used to identify the root causes.

    Defect prevention plan is an on- going document that need to be tracked very closely. It is meant both for planning and tracking defect prevention activities. This plan has to be revisited after completion of each stage. If defects detected during the completed stage fall under different defect types and defect causes not identified for preventing at that stage, then these new types need to be included in the on-going phases.

## VII. CONCLUSION

IT organizations focus on customer satisfaction is the key for survival. Unfortunately the element of predictive behavior during planning phase is very minimal and subjective. While Capability Maturity Model recommends pro-active management using quantitative models, the practical implementation is very low. The context of the organization is important in building these models. It is also important to understand that the project managers need to be equipped with the right information, metrics baseline and subject matter expertise. Defect leakage is a standard concern in the industry. The practical case study demonstrated the influence of Coding Defect Density and Design Defect Density. The case study also helped us understand how the values need to be determined, the steps around what-if analysis, the defect prevention plan and the tracking mechanism. This illustration gives us the confidence that the predictive mechanism can be planned and executed well in an organization.

## REFERENCES

[1] CMMI® Product Development Team, CMMI for Development, Version 1.2, CMMI-DEV, V1.2. Carnegie Mellon University, Pittsburgh, PA., 2006.

[2] Stephen H. Kan, Metrics and Models in Software Quality Engineering 2nd ed., 2002

[3] Richard D. Stutzke, Estimating Software-Intensive Systems: Projects, Products, and Processes, Addison-Wesley, 2005

[4] Robert W. Stoddard, CMMI Process Performance Models and Reliability, Carnegie Mellon University, IEEE Reliability Society Annual Technology Report, 2007

[5] Umesh Kumar Mishra, K. Harihara sudhan, Shalu Gupta. Establishing Process Performance Baselines and Models for Statistical Control of Software Projects, Proceedings of ASCNT-2011, CDAC, Noida, India

[6] Tim Kasse, Practical Insight into CMMI, 2nd ed., Artech House, 2008

[7] Bill Self, Greg Roche and Nigel Hill, Customer Satisfaction Measurement for ISO 9000: 2000, 2002

[8] Michael Johnson and Anders Gustafsson, Improving Customer Satisfaction, Loyalty and Profit, 2000

[9] Margaret K. Kulpa and Kent A. Johnson, Interpreting the CMMI: A Process Improvement Approach, 2nd ed., 2008

[10] Tim Kasse, Practical Insight into CMMI, 2nd ed., 2008

[11] CMMI® Product Development Team, CMMI for Development, Version 1.2, CMMI-DEV, V1.2. by Carnegie Mellon University, Pittsburgh, PA, 2006

[12] Don Peppers and Martha Rogers, Managing Customer Relationships: A Strategic Framework, John Wiley & Sons, 2004

[13] Forrest W. Breyfogle III, Implementing Six Sigma: Smarter Solutions Using Statistical Methods, John Wiley & Sons, 2nd ed., 2003

# Spontaneous-braking and lane-changing effect on traffic congestion using cellular automata model applied to the two-lane traffic

Kohei Arai [1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

Steven Ray Sentinuwo [2]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*— In the real traffic situations, vehicle would make a braking as the response to avoid collision with another vehicle or avoid some obstacle like potholes, snow, or pedestrian that crosses the road unexpectedly. However, in some cases the spontaneous-braking may occur even though there are no obstacles in front of the vehicle. In some country, the reckless driving behaviors such as sudden-stop by public-buses, motorcycle which changing lane too quickly, or tailgating make the probability of braking getting increase. The new aspect of this paper is the simulation of braking behavior of the driver and presents the new Cellular Automata model for describing this characteristic. Moreover, this paper also examines the impact of lane-changing maneuvers to reduce the number of traffic congestion that caused by spontaneous-braking behavior of the vehicles.**

***Keywords- spontaneous-braking; traffic congestion; cellular automata; two-lane trafficcomponent.***

## I. INTRODUCTION

The study of traffic flow has received a lot of attention for the past couple of decades. The simulations of traffic congestion become the most important aspect in the field of traffic analysis and modeling. Traffic congestion can be defined as the saturation condition of road network that occurs as increased traffic volume or interruption on the road, and is characterized by slower speed, longer trip times, and increased vehicular queuing. The investigated situations in the real traffic condition are those of traffic congestion caused by some main reason, such as insufficient road capacity, incidents, work zones (e.g., road maintenance or constructions near the road that requires space), weather events (e.g., in the case of rain or snow) which can hampers visibility therefore a driver have to slowdown its vehicle to compensate, or emergencies situations (e.g., hurricanes or severe snowstorms). However, in this paper, we concern to investigate the effect of individual braking behavior of the driver towards traffic congestion.

In more detail, this paper interests to describe and reproduce the characteristic of spontaneous-braking probability and its effects to the traffic behavior. In the real traffic situations, vehicle would make a braking as the response to avoid collision with another vehicle or avoid some obstacle like potholes, snow, or pedestrian that crosses the road unexpectedly. However, in some cases the spontaneous-braking may occur even if there are no obstacles in front of the vehicle. In some country, the reckless driving behaviors such as sudden-stop by public-buses, motorcycle which changing lane too quickly, or tailgating make the probability of braking getting increase.

One of the famous microscopic models for the simulation of road traffic flow is Cellular Automata (CA) model. In comparison with another microscopic model, the CA model proposes an efficient and fast performance when used in computer simulation [18]. CA is a dynamic model developed to model and simulates complex dynamical system. The set of CA rules may illustrate complex evolution patterns, such as time and space evolution in a system. Those evolutions can be shown just by use simple rules of CA. Furthermore, the utilization of CA successfully explains the phenomenon of transportation. These so-called traffic cellular automata (TCA) are dynamical systems that are discrete in nature and powerful to capture all previously mentioned basic phenomena that occur in traffic flows [18]. The one dimensional cellular automata model for single lane freeway traffic introduced by Nagel and Schreckenberg (NaSch) [1] is simple and elegant that captures the transition from laminar flow to start-stop waves with increasing vehicle density. The space of CA is discrete and consists of a regular grid of cells, each one of which can be in one of finite number of possible states. The number and array of cells in the grid depends on the specific transportation behavior that is described. The simplicity of the NaSch model has prompted the use of it for studying many traffic situations.

This paper presents a new Cellular Automata model for describing the phenomena of spontaneous-braking behavior and lane-changing character in traffic flow. In this model, we investigate the effect of spontaneous-braking probability and lane-changing maneuver in two-lane highway with one-way traffic character. This proposed model extends the NaSch model that first introduced CA for traffic simulation. The set of rules in NaSch model are modified to better capture and describe the behavior of the driver while making spontaneous-braking and lane-changing maneuver in traffic flow. The base deceleration rule of NaSch model is applicable only to stationary vehicles, which is vehicles that are blocked by the

leading vehicle in the previous time step. This rule is not applicable to two conditions, in the condition of those vehicles which are stopped due to spontaneous-braking behavior, and in the two-lane highway that allows vehicle to make lane-changing maneuvers. Compared with the original NaSch model, this proposed model exhibits spontaneous-braking probabilities effect combined with acceleration, deceleration, and lane-changing maneuvers effects. Though it is well known that spontaneous-braking is extremely reducing the local speed of vehicles, the impact on the global system has not been studied.

This paper uses a two-lane highway character with a periodic boundary condition. The periodic boundary approach has been used to conserve the number of vehicles and the stability of the model. The goal of this paper is to analyze the phenomena of spontaneous-braking behavior in traffic flow then propose a new cellular automata model to describing this phenomena. Moreover, this paper also investigates the impact of lane-changing maneuvers towards traffic congestion that is caused by spontaneous-braking behavior.

This paper is organized as follows. Some studies relating with CA based traffic flow is quick reviewed in Section 2. Section 3 presents a short description of the theoretical aspect of traffic CA model. Section 4 explains about the proposed model. Section 5 contains simulation process and the results in the form of fundamental diagrams and space-time diagrams. Finally, Section 6 contains conclusion and a summary of findings.

## II. Related Research Works

The one dimensional cellular automata model for single lane freeway traffic introduced by Nagel and Schreckenberg (NaSch) [1] is a probabilistic CA model that captures the transition from laminar flow to start-stop waves with increasing vehicle density. NaSch model update the state of cells synchronously in discrete time steps. There is a finite set of local interaction rules. This set of rules manages the new state of a cell by taking into account the actual state of the cell and its neighbor cells. This local interaction allows capture micro-level dynamics and propagates it to macro-level behavior. This single-lane system consists of a one-dimensional grid of L sites with periodic boundary conditions. A site can either be occupied, or empty by one vehicle with integer velocity between zero and $v_{max}$. The velocity of each vehicle is equivalent to the number of sites that a vehicle advances in one update, if there is no obstacle ahead. Each of vehicles moves only in one direction. Refer to the Ricket et. al [6], they outlined the rules of single-lane model. The index $i$ denotes the number of vehicle, $x(i)$ is the position of vehicle $i$, $v(i)$ is the vehicle's current velocity, $v_d(i)$ is the maximum speed, $pred(i)$ is the number of preceding vehicle, $gap(i) = x(pred(i)) - x(i) - 1$ indicates the width of the gap to the predecessor. The rules are applied to all vehicles at the beginning of each time step by simultaneously, which mean using parallel update. Then the vehicles are advanced according to their new velocities [6].

The parallel update rules are the following:

- $v(i) \neq v_d(i) \Rightarrow v(i) := v(i) + 1$ (1)
- $v(i) > gap(i) \Rightarrow v(i) := gap(i)$ (2)
- $v(i) > 0 \Rightarrow rand < p_d(i) \Rightarrow v(i) := v(i) - 1$ (3)

The first rule represents the linear acceleration of each vehicle which is not at the maximum speed to accelerate its speed by one site (cell) until the vehicle has reached its maximum velocity $v_d$. Second rule ensures that vehicles having predecessors in their way slowdown in order not to run into them. In this rule, all vehicles are checked for their distance between the vehicle and its predecessor. If the distance is smaller than its speed then the speed is reduced to the number of empty cells between them to avoid the collision. Third rule consider the stochastic noise parameter.

The probability $p_d$ is the probability number of each car to reduce its speed by one unit (cell) per time step. This NaSch model encouraged another study toward traffic flow conditions [2]-[7]. Ricket, et al. [8] investigated a simple model for two-lane traffic. Their model introduced the lane changing behavior for two lanes traffic. It was found that the fundamental diagram for each lane is asymmetric but the maximum is shifted towards large values of vehicular density $\rho$ ($\rho_{max} > 1/2$ ). They proposed a symmetric rule set where the vehicle changes lanes if the following criteria are fulfilled:

- $v_{move} > gap_{same} \rightarrow v_{move} = \min(v_n + 1, v_{max})$
- $gap_{target} > gap_{same}$
- $gap_{back} \geq v_{max}$

The variable $gap_{same}$, $gap_{target,}$ and $gap_{back}$ denote the number of unoccupied cells between the vehicle and its predecessor on its current lane, and between the same vehicle and its two neighbor vehicles on the desired lane, respectively.

The advance analysis about lane-changing behavior has been done, which includes symmetric and asymmetric rules of lane-changing [9-14]. Symmetric rule can be considered as rules that threat both lanes equally, while asymmetric rule can be applied in special characters highway, like German highways simulation [15], where lane changes are dominated by right lane rather than left lane. Another studies focus on the effect of lane-changing behavior on a two-lane road in presence of slow vehicle and fast vehicle [13], [16-18]. While the NaSch model could reproduce some of basic phenomenon observed in real traffic situations such as the start-stop waves in congested traffic, but it has been observed that the base NaSch model lacks the ability to produce other more realistic traffic patterns [19].

In this paper, we consider two parameters in traffic behavior; those are the spontaneous-braking behavior and lane-changing maneuver that occurs in the real traffic situation. This proposed model using two-lanes traffic and also adopts the symmetric lane-changing rules.

## III. Traffic Cellar Automata Model

Cellular automaton (CA), at the basis of the model presented in this paper, is a discrete model studied in computability theory, mathematics, physics, complexity science, theoretical biology and microstructure modeling.

Currently, various fields have been using CA models to model the phenomena of their system, such as vehicular traffic flow, pedestrian behavior, escape and panic dynamic, collective behavior, and self-organization. CA model uses a simple approach for modeling and simulation of complex dynamical systems. The behavior of complex systems can be described by considering at the local interactions between their elementary parts. CA decomposes a complex phenomenon into a finite number of elementary processes.

The CA model consists of two components, a cellular space and a set of state. The state of a cell is completely determined by its nearest neighborhood cells. All neighborhood cells have the same size in the lattice. Each cell can either be empty, or is occupied by exactly one node. There is a set of local transition rule that is applied to each cell from one discrete time step to another (i.e., iteration of the system). This parallel updating from local simple interaction leads to the emergence of global complex behavior.

The Nagel-Schreckenberg (NaSch) model is one of the theoretical CA models for the simulation of freeway traffic [1]. This NaSch model known as the simple CA model for illustrate road traffic flow that can reproduce traffic congestion, like slow down car behavior in a high-density road condition. This model shows how traffic congestion can be thought of as an emergent or collective phenomenon due to interactions between cars on the road, when the density of cars is high and so cars are close to each on average. The NaSch model also known as *stochastic traffic cellular automaton* (STCA) because it included a stochastic term in one of its rules. Like in deterministic traffic CA models (e.g., CA-184 or DFI-TCA), this NaSch model contains a rule that reflect vehicle increasing speed and braking to avoid collision. However, the stochasticity term also introduced in the system by its additional rule. In one of its rules, at each time-step *t*, a random number $\xi(t) \in [0,1]$ is generated from a uniform distribution. This random number is then compared with a stochastic noise parameter $p \in [0,1]$. For it is based on this probability $p$ then a vehicle will slow down to $v(\mathrm{i}) - 1$ cells/time-step. According to Nagel and Schreckenberg, the randomization rule captures natural speed fluctuations due to human behavior or varying external conditions [20].

## IV. Proposed Method

This paper extends a probabilistic CA model that introduced by Nagel-Schrekenberg [1] for the description of single-lane highway traffic. While the original NaSch model uses a single lane that is represented by a one-dimensional array of L sites (cells), this paper considers two-lane highway with unidirectional traffic character in periodic boundaries condition. The two-lane model is needed to describe the more realistic traffic condition which has several types of vehicles with multiple desired velocities. In single-lane model, the vehicles with multiple desired velocities just resulting in the platooning effect with slow vehicle being followed by faster ones and the average velocity reduced to the free-flow velocity of the slowest vehicle [8].

The simulation model in this paper presents two additional elements. The first additional element is spontaneous-braking parameter. This element is needed to illustrate the probability of spontaneous-braking behavior of the vehicle that occur in the real traffic situation. The concept of spontaneous-braking probability is introduced for the description of the spontaneous reaction of the drivers while making a spontaneous-braking behavior. This reaction can be caused by several things e.g., as the response to avoid collision with another vehicles, the reckless driving behaviors such as sudden-stop by public-buses, motorcycle which changing lane too quickly, or tailgating. Those behaviors make the probability of braking getting increase.

In original NaSch model [1], there is no rule accommodate the spontaneous-braking behavior. NaSch model introduced a stochastic noise parameter $p \in [0,1]$ that can make a slowdown vehicle to $v(\mathrm{i}) - 1$ cells/time-step. However, in real traffic situations this rule is difficult to describe the nature of the braking, especially on spontaneous-braking behavior of the vehicle. In our opinion, the value of braking is a variable number and the spontaneous-braking represent the extreme value of a braking behavior. Thus, the slow-down rule of vehicle $v(\mathrm{i}) - 1$ cells/time-step cannot describe the characteristic of spontaneous-braking. This paper introduces a new additional rule to represent the behavior of spontaneous-braking by using a spontaneous-braking probability $P_b$: $v(i) \rightarrow v(i) - b_x$ . Here $bx$ denotes the characteristic of driver while make a braking. The value of $b_x$ is equal or less than the current speed $v(i)$. This rule takes into account the dynamic characteristic of the driver while make a braking of its car. Already mentioned before, a two-lane unidirectional highway model with periodic boundary system is used in this computational model. Refer to the discrete NaSch model, a one-dimensional chain of L cells of length 7.5 m represents each lane. There are just two possibility states of each cell. Each cell can only be empty or containing by just one vehicle. The speed of each vehicle is integer value between $v = 0, 1, \ldots, v_{\max}$. In this model, all vehicles are considered as homogeneous then have the same maximum speed $v_{\max}$. In order to investigate the effect of spontaneous-braking behavior then the state of a road cell at the next time-step, from $t$ to $t + 1$ is dependent on the states of the direct frontal neighborhood cell of the vehicle and the core cell itself of the vehicle. The state of the road cells can be obtained by applying the following rules to all cells (vehicles) by parallel updated:

$$\text{Acceleration: } v(i) \rightarrow \min(v(i) + 1, v_{\max}) \quad (4)$$

$$\text{Deceleration: } v(i) \rightarrow \min(v(i), gap(i)) \quad (5)$$

$$\text{Spontaneous braking probability } p_b\text{: } v(i) \rightarrow v(i) - b_x \quad (6)$$

$$\text{Driving: } x(i) \rightarrow x(i) + v(i) \quad (7)$$

As this simulation model try to investigate the effect of spontaneous-braking behavior on traffic flow then this model deliberately eliminates the randomization rule of original NaSch ($v(\mathrm{i}) - 1$ cells/time-step). Here for the reason to avoid the speed reduction of vehicles caused by this rule that could influence our simulation results. The variable *gap(i)* indicates the distance between a vehicle *x(i)* and its predecessor

*x((i)+1). v*max represents the maximum speed of the vehicle.

The second additional element is lane-changing parameter. By using two-lane highway model and applying multiple

desired velocity types, then this paper also accommodates the lane-changing maneuvers of vehicles. In the real traffic situation, driver tends to make a lane-changing maneuver while encounter traffic congestion along its lane. This paper also intends to evaluate the impact of lane-changing maneuvers towards the traffic congestion that caused by spontaneous-braking behavior of the driver. In this model, the lane-changing maneuver is analogous as the movement of liquid. There is a different from the lane-changing model of Ricket et al. In this model, a vehicle would consider changing its lane only if the vehicles "see" another vehicle on its cell ahead and do so if possible. It means, as long as there is a cell free ahead on their lane then the vehicles would still remain on their lane. This lane-changing model will preserve the deceleration rule in our model that is showed in equation (5).

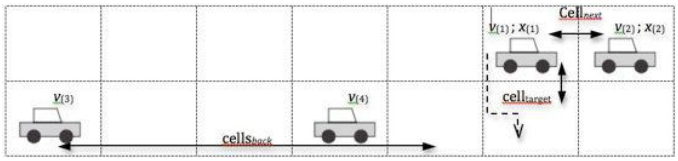



Figure 1. Schematic diagram of a lane-changing operation

The lane-changing rule is applied to vehicles to change from right lane to left lane and conversely. Vehicles are only move sideways and they do not advance. Fig. 1 shows the schematic diagram of lane-changing operation. A vehicle changes to the next lane if all of the following conditions are fulfilled:

- $\text{Cell}_{next} > 0$ (8)
- $\text{Cell}_{\text{target}} = 0$ (9)
- $x_{(cellsback)} + v_{(cellsback)}^{t+1} \neq cell_{t\ arget}$ (10)

$\text{Cell}_{\text{next}}$, $\text{Cell}_{\text{target}}$, and $\text{Cell}_{\text{back}}$ are the parameters that inform the state of one cell ahead, state of next cell, and state of cells behind on the other lane, respectively. If one cell is unoccupied or free-cell then its state is 0. In the real traffic situation, a driver also has to look back on the other lane and estimate the velocity of another cars-behind to avoid a collision. Equation (10) accommodates the driver behavior to estimate the velocity of vehicles before change the lane.

## V. SIMULATIONA DN RESULTS

The simulation starts with an initial configuration of N vehicles, with random distributions of positions on both lanes. This simulation use the same initial velocity for all vehicle $v_{\min} = 0$ and the maximum vehicle speed has been set to $v_{\max} = 5$ cell/time-step. Many simulations performed with different density ρ. The density ρ can be defined as number of cars N along the highway over number of cells on the highway L. During one simulation, the total number of cars on the highway cannot change. Vehicles go from left to right. If a vehicle arrives on the right boundary then it moves to the left boundary. Fig. 2 illustrates an environment, which exhibits a certain configuration.

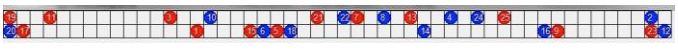


Figure 2. An environment with a certain configuration

This paper divides the analysis into two stages. The first stage investigates the effect of spontaneous-braking on the traffic flow. In this simulation stage, we analyze the traffic flow for the spontaneous-braking probability $b_p$ = 0; 0.3; and 0.7. The simulation was running 1000 time steps to let the system reaches its stable condition. The system automatically increase the vehicles density from minimum density ρ = 0 until maximum density ρ = 100 percent. Once the transient dies out, then the data extraction was started. The data was analyzed using fundamental diagrams, which plot the velocity of vehicle *vs* vehicle flow *vs* global density.

To show the system dynamics then the graph had written the last ten steps for each density before the end of simulation. Fig. 3 and Fig. 4 present the fundamental diagrams of this model. Fig. 3 shows the measurement of the average velocity $v(t)$ over all vehicles at each density. The red color, black color, and blue color of scatter graph present the average velocity in the condition with spontaneous-braking probability $P_b = 0$, $P_b = 0.3$, and $P_b = 0.7$, respectively.

One can be observed that in the traffic without spontaneous braking probability, the maximum velocity 5 unit of distance per unit of time could be achieved in the density ρ ≤ 0.12. When the probability of spontaneous-braking increased then the critical density point that maximum velocity can be achieved became lower than normal condition.

For the spontaneous-braking probability $P_b = 0.3$, the critical point of maximum velocity $v_{\max} = 5$ is around ρ = 0.04. While in the situation that spontaneous-braking probability $P_b = 0.7$, the vehicles were very difficult to reach their maximum speed $v_{\max} = 5$.

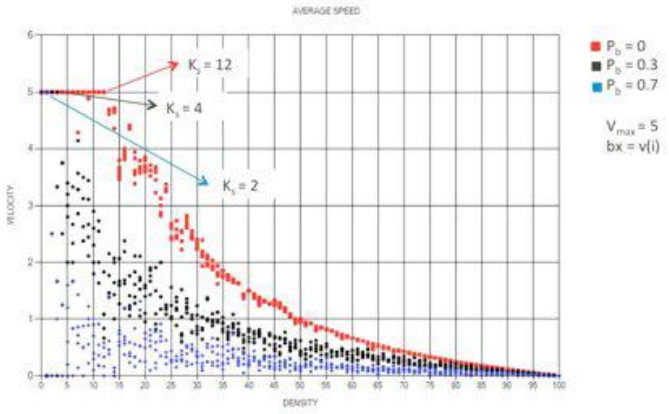



Figure 3. Average velocity (cell/time-step) *vs* density (cars/highway site)

In the phase after the critical density point of maximum velocity was reached, the vehicles reduced their velocity to synchronize with the gap between them and the vehicle ahead.

However, in the transition phase after the critical density point of maximum velocity, the vehicles still maintained their velocity. Regarding this average velocity graph, the traffic jam obviously appeared when the average velocity $v < 1$ cell/time.

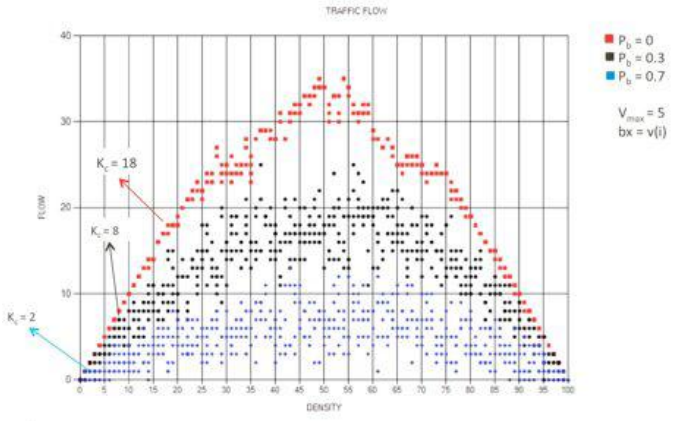



Figure 4. Traffic flow (cars/time step) *vs* density (cars/highway site)

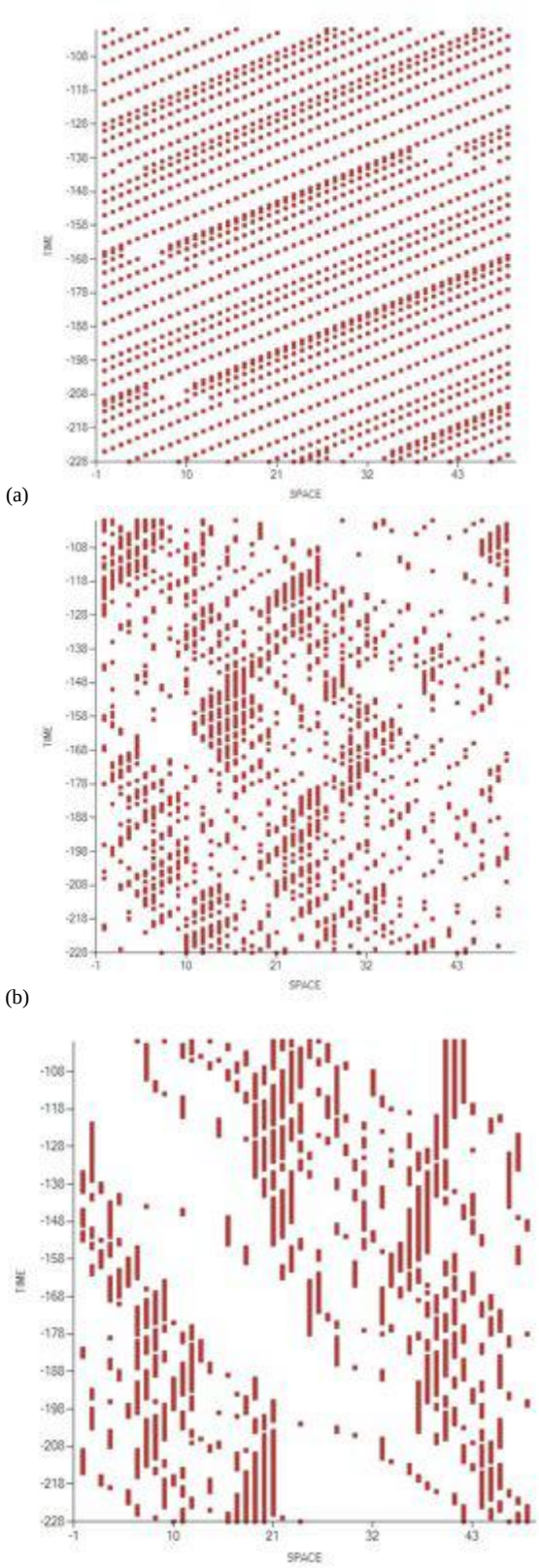
Fig. 4 illustrates the traffic flow over vehicles density for the spontaneous-braking probability $P_b = 0$, $P_b = 0.3$, and $P_b = 0.7$, respectively. The traffic flow indicates the number of moving vehicles per unit of time. While the density parameter means the number of vehicles per unit area of the highway. As can be seen from the graph, there is a reduction in traffic flow in the presence of spontaneous-braking parameter. We also consider the critical density $k_c$ that appeared in each traffic flow. Here, the critical density means a maximum density achievable under free flow. In the traffic flow with $P_b = 0$, the critical density $k_c$ situated at the density $\rho = 0.18$.

The critical density $k_c$ was getting lower when the spontaneous-braking parameter increased. Below the critical density $k_c$, all vehicles can make a movement. However, in the density after the critical density point, not every vehicle can move at each time step. This critical density point also indicates when the traffic congestion started to happen. To get an intuitive feel for the dynamics, we provide a set of space-time diagrams in Fig. 5, Fig. 6, and Fig. 7 for various density values.

The horizontal axis represents space and vertical axis represents the time. In order to get data to analyze, we simulate this model for density $\rho = 0.25$; 0.50; and 0.75 that represent light traffic, moderate traffic, and heavy traffic situations.

For density $\rho = 0.25$, it can be seen that the spontaneous-braking behavior has given a significant impact to produce traffic congestion (Fig. 5). The single vertical line which is shown in these time-space diagrams represents a stationary vehicle that is making a spontaneous-braking behavior. In the traffic with density value $\rho = 0.50$, there is a moderate impact of the spontaneous-braking behavior on the traffic congestion.

It can be seen that before the spontaneous-braking parameter was applied, the congestion already occurred on the traffic (Fig. 6). While in Fig. 7, the effect of spontaneous-braking on traffic congestion just a slightly impact is shown. That because in density value $\rho = 0.75$, the traffic congestion already appeared although in the condition without spontaneous-braking behavior.




(a)

(b)

(c)

Figure 5. Space-time diagram for density $\rho = 0.25$ and $P_b = 0$ (a), $P_b = 0.3$ (b), and $P_b = 0.7$ (c); without lane-changing maneuvers

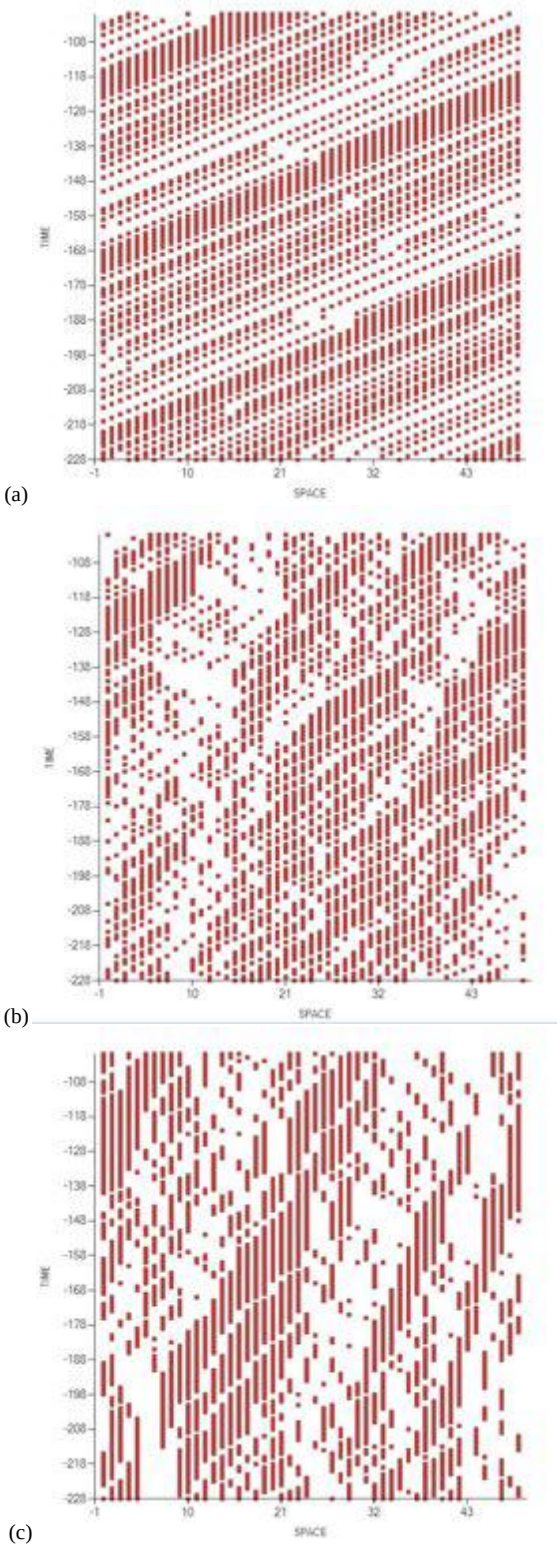



Figure 6. Space-time diagram for density ρ = 0.50 and $P_b$ = 0 (a), $P_b$ = 0.3 (b), and $P_b$ = 0.7 (c); without lane-changing maneuvers

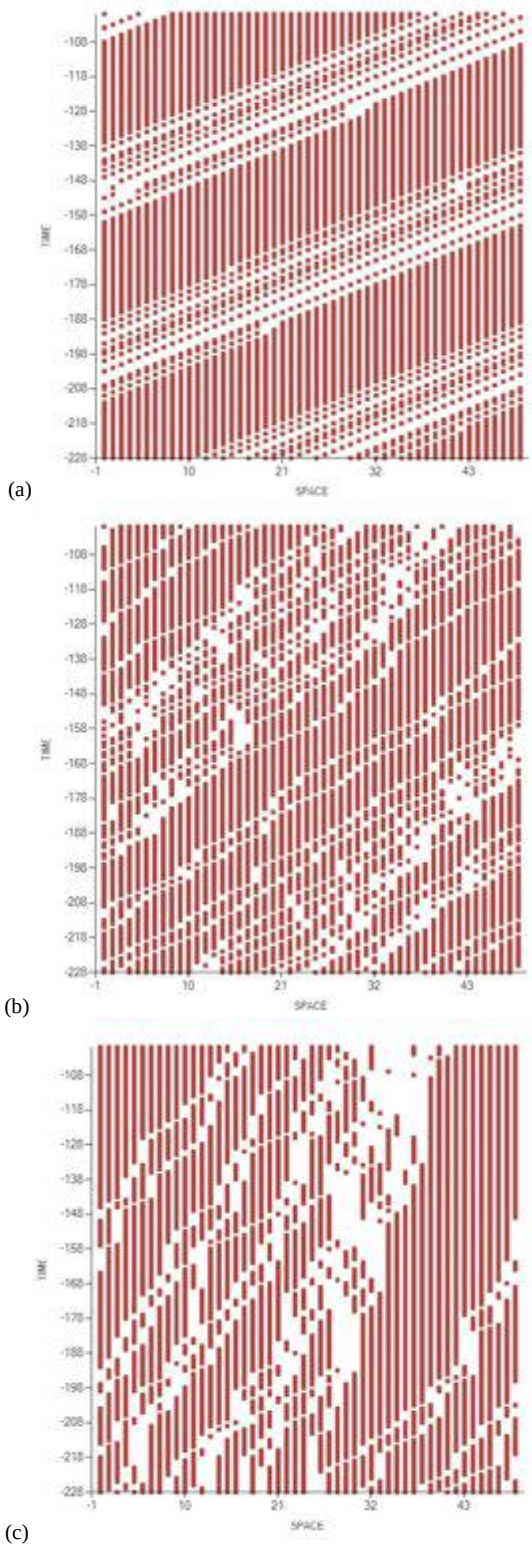
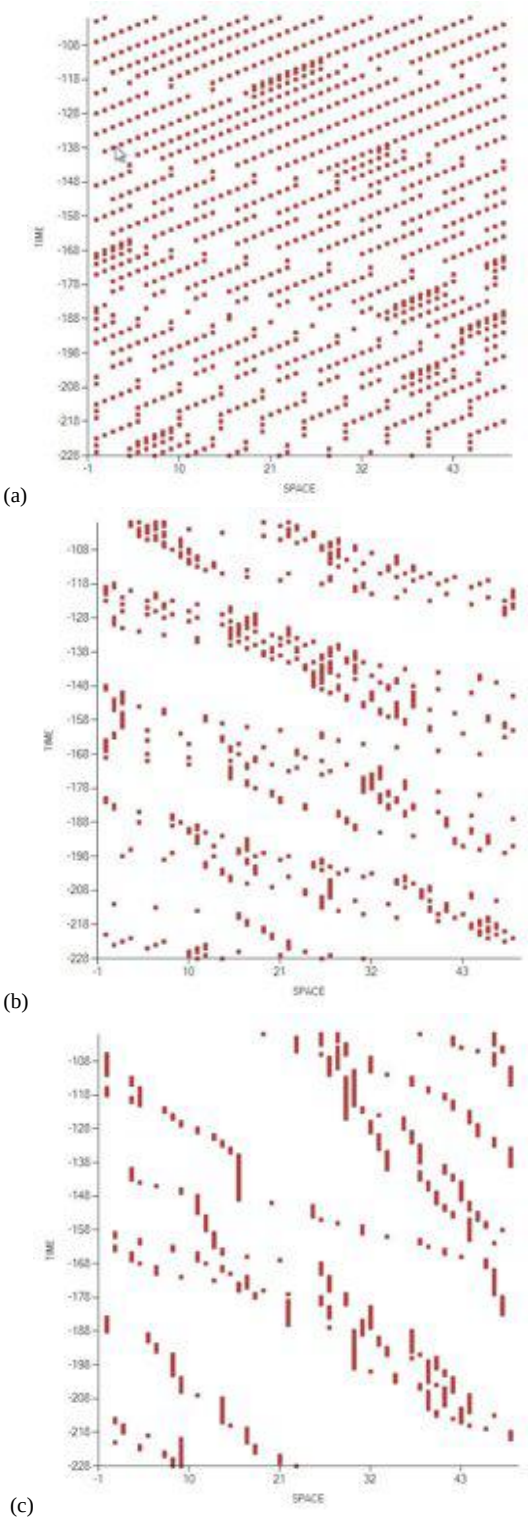



Figure 7. Space-time diagram for density ρ = 0.75 and $P_b$ = 0 (a), $P_b$ = 0.3 (b), and $P_b$ = 0.7 (c); without lane-changing maneuvers

The lane-changing effect on traffic congestion is discussed from here. As shown before that the spontaneous-braking behavior can contribute to the traffic congestion.

Therefore, in this section we evaluate the effect of lane-changing to reduce the congestion level. This lane-changing model was applying the equations (8), (9), and (10).

In this simulation, the vehicles can look back and estimate the situation along 5 cells behind on the other lane before make a lane-changing. We provide a set of space-time diagrams in Fig. 8, Fig. 9, and Fig. 10 for the density values ρ = 0.25; 0.50; and 0.75.

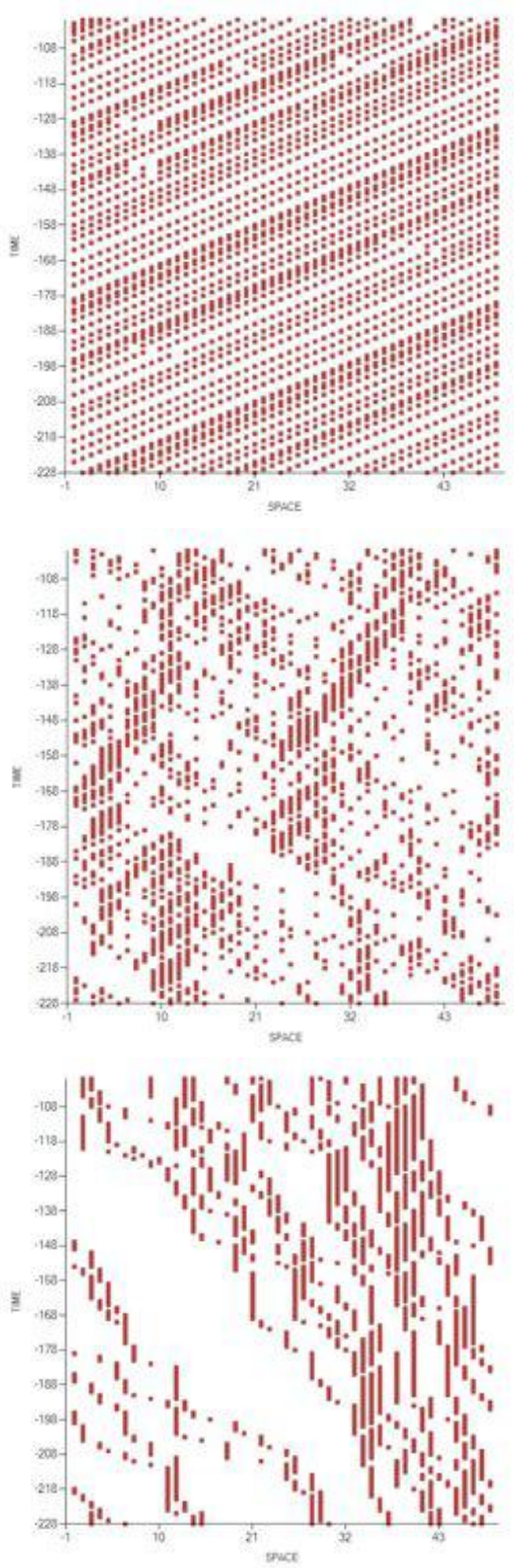



(a)

(b)

(c)

Figure 8. Space-time diagram for density ρ = 0.25 and $P_b$ = 0 (a), $P_b$ = 0.3 (b), and $P_b$ = 0.7 (c); with lane-changing maneuvers

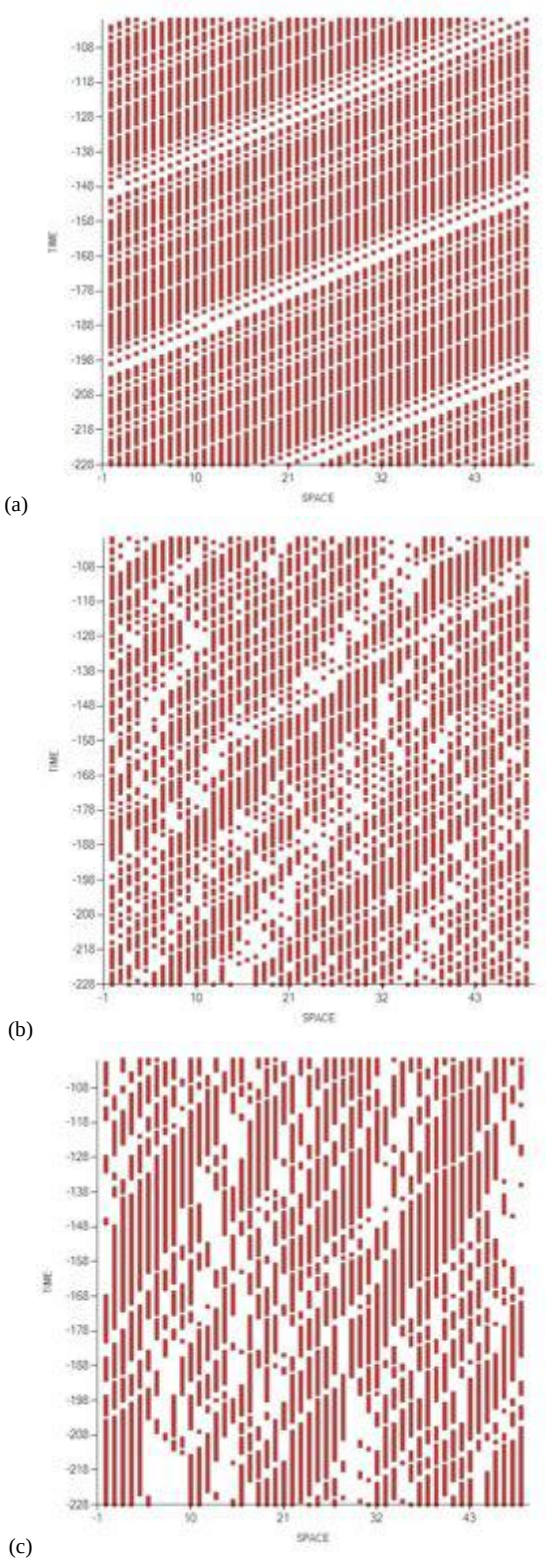



(a)

(b)

(c)

Figure 9. Space-time diagram for density ρ = 0.50 and $P_b$ = 0 (a), $P_b$ = 0.3 (b), and $P_b$ = 0.7 (c); with lane-changing maneuvers




(a) (b) (c)

Figure 10. Space-time diagram for density ρ = 0.75 and $P_b$ = 0 (a), $P_b$ = 0.3 (b), and $P_b$ = 0.7 (c); with lane-changing maneuvers.

The comparative graph shows that for the traffic density ρ < 0.75, the lane-changing maneuvers have given a good impact to reduce the congestion level. However, in all spontaneous-braking parameter value condition, the result shows that there is no significant impact that is contributed by lane-changing maneuver.

## VI. CONCLUSION

In this work, we simulate the braking behavior of the driver and present the new Cellular Automata model for describing this characteristic. The original NaSch model has been modified to accommodate the parameter of spontaneous-braking probability. This spontaneous-braking probability rule captures the natural of braking behavior due to human behavior. This simulation shows that the traffic congestion can be caused not only by the road capacity condition but also by driver behavior. Moreover, we also evaluate the effect of lane-changing to reduce the congestion that is caused by the parameter of spontaneous-braking probability.

## REFERENCES

[1] K. Nagel and M. Schreckenberg, "A cellular automaton model for freeway traffic," Journal of Physics I France, vol. 2, no. 12, pp.2221-2229, 1992.

[2] A. Schadschneider and M. Schreckenberg. "Cellular automaton models and traffic flow," Physics A, 1993.

[3] L. Villar and A. de Souza, "Cellular automata models for general traffic conditions on a line," Physica A, 1994.

[4] M. E. Lárraga, J. a. D. Río, and L. Alvarez-lcaza, "Cellular automata for one-lane traffic flow modeling," Transportation Research Part C: Emerging Technologies, vol. 13, no. 1, pp. 63-74, Feb. 2005.

[5] K. Nagel, "Particle hopping models and traffic flow theory," Physical review. E, vol. 53, no. 5, pp. 4655-4672, May 1996.

[6] K. Arai and Tri Harsono Agent and diligent driver behavior on the car-following part of the micro traffic flow in a situation of vehicles evacuation from Sidoarjo Prong roadway, International Journal of Computer Science and Network Security, 11, 1, 137-144, 2011.

[7] K. Arai, Tri Harsono, Ahmad Basuki, "Car-Following Parameters by Means of Cellular Automata in the Case of Evacuation," International Journal of Computer Science and Security (IJCSS), Vol (5), 2011.

[8] M. Rickert, K. Nagel, M. Schreckenberg, and A. Latour, "Two Lane Traffic Simulations using Cellular Automata," vol. 4367, no. 95, 1995.

[9] W. Knospe, L. Santen, A. Schadschneider, and M. Schrekenberg, "Disorder effects in cellular automata for two lane traffic," Physica A, vol. 265, no. 3-4, pp. 614–633, 1998.

[10] A. Awazu, "Dynamics of two equivalent lanes traffic flow model: selforganization of the slow lane and fast lane," Journal of Physical Society of Japan, vol. 64, no. 4, pp. 1071– 1074, 1998.

[11] E. G. Campri and G. Levi, "A cellular automata model for highway traffic," The European Physica Journal B, vol. 17, no. 1, pp. 159–166, 2000.

[12] L. Wang, B. H. Wang, and B. Hu, "Cellular automaton traffic flow model between the Fukui-Ishibashi and Nagel- Schreckenberg models," Physical Review E, vol. 63, no. 5, Article ID 056117, 5 pages, 2001.

[13] B. Jia, R. Jiang, Q. S. Wu, and M. B. Hu, "Honk effect in the two-lane cellular automaton model for traffic flow," Physica A, vol. 348, pp. 544–552, 2005.

[14] D. Chowdhury, L. Santen, and A. Schadschneider, "Statistical physics of vehicular traffic and some related systems," Physics Report, vol. 329, no. 4-6, pp. 199–329, 2000.

[15] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg, "A realistic two-lane traffic model for highway traffic," Journal of Physics A, vol. 35, no. 15, pp. 3369–3388, 2002.

[16] D. Chowdhury, L. Santen, and A. Schadschneider, "Statistical physics of vehicular traffic and some related systems," Physics Report, vol. 329, no. 4-6, pp. 199–329, 2000.

[17] R. J.Harris and R. B. Stinchcombe, "Ideal and disordered two- lane traffic models," Physica A, vol. 354, no. 1–4, pp. 582–596, 2005.

[18] X. G. Li, B. Jia, Z. Y. Gao, and R. Jiang, "A realistic two-lane cellular automata traffic model considering aggressive lane- changing behavior of fast vehicle," PhysicaA, vol. 367, pp. 479– 486, 2006.

[19] W. Knospe, L. Santen, A. Schadschneider, and M. Schreckenberg, "Empirical test for cellular automaton models of traffic flow," Phys. Rev. E, vol. 70, 2004.

[20] S. Maerivoet and B. D. Moor, "Transportation Planning and Traffic Flow Models," 05-155, Katholieke Universiteit Leuven, Department of Electrical Engineering ESAT-SCD (SISTA), July 2005.

## AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission A of ICSU/COSPAR since 2008. He wrote 30 books and published 307 journal papers

# Enhancing eHealth Information Systems for chronic diseases remote monitoring systems

Amir HAJJAM
Laboratoire Systèmes et Transports
Université de Technologie de Belfort Montbéliard, UTBM
Belfort, France

***Abstract*— Statistics and demographics for the aging population in Europe are compelling. The stakes are then in terms of disability and chronic diseases whose proportions will increase because of increased life expectancy. Heart failure (HF), a serious chronic disease, induces frequent re-hospitalizations, some of which can be prevented by up-stream actions. Managing HF is quite a complex process: long, often difficult and expensive. In France, nearly one million people suffer from HF and 120,000 new cases are diagnosed every year. Managing such patients, a telemedicine system tools associated with motivation and education can significantly reduce the number of hospital days that believes therefore that the patient is hospitalized for acute HF. The current development projects are fully in prevention, human security, and remote monitoring of people in their living day-to-day spaces, from the perspective of health and wellness. These projects encompass gathering, organizing, structuring and sharing medical information. They also have to take *into account* the main *aspects* of *interoperability*. A different approach has been used to capitalize on such information: data warehouse approach, mediation approach (or integration by views) or integration approach by link (or so-called mashup).**

**In this paper, we will focus on ontologies that take a central place in the Semantic Web: on one hand, they rely on modeling from conceptual representations of the areas concerned and, on the other hand, they allow programs to make inferences over them.**

***Keywords- Ontologies; Web Semantic; Remote Monitoring; Chronic Diseases.***

## I. INTRODUCTION

The pervasiveness of chronic diseases is highly growing with increased life expectancy. In most developed countries, those diseases are responsible for increasingly growing health spending. Today, there are more than 15 million patients suffering from such diseases in France as we do expect this number to grow over 20 million by 2020 [1]. Having those patients in specialized institutions (hospitals, nursing homes ...) is not only really desired but even not possible. A European study Catalan Remote Management Evaluation (CARME) [2] has shown that there was a 68% decrease in heart failure related hospitalization and a 73% reduction of days spent in hospital from 646 days to 168 days. The move is towards solutions known as "home care", where patients are to be cared for, medically and paramedically, by remaining in their own homes. These remote monitoring solutions provide unquestionably higher quality of care and greater security than conventional practices and better quality of life for patients. They incorporate the most innovative technological aspects (monitoring and remote transmission of vital signs, detect falls, alarms, etc.) and organizational aspects necessary for the coordination of the different players contributing this "home care". These solutions are still widely at an experimental stage, especially to assess their economic viability.

Pilot projects, with various concepts and objectives, were born throughout the world: Gator Tech [3] for the USA, Prosafe [4] for France, the work of Ogawa [5] [6] in Japan or yet CarerNet [7] for England. Most recently, we have the systems based on ontologies proposed by [8-12]. These projects vary both in scale deployment and diseases monitored (daily activities, asthma, Alzheimer's, cardiovascular disease, falls, etc.). However, they all put back up relevant information on the evolution of the patient's health including information on daily activities.

Most of these projects include various sensors to monitor the person's home (medical sensors, motion sensors, infrared sensors etc..). Some, like the Gator Tech project and the work of Tamura, focus on the instrumentation of domicile to study the lifestyle of the occupant (electronic bathroom scales, ECG in the bathtub, intelligent floors for fall detection, etc.) and make his life easier. Other projects such as TelePat [13], Ailisa [14] [15], CarerNet adjoin the sensors and home automation physiological sensors to be placed directly on the person to bring up more detailed medical data and allow a finer tracking of changes in his condition.

All this information is daily backed up to monitor patients to early detection of any abnormalities, behavioral changes or vital signs, to raise an alert. The objective of such platforms is to monitor a large number of patients. If we take the single case of heart failure patients, actually they account for France about 1 million patients with more than 120,000 new cases per year. The amount of information stored and processed is designed logically to an explosion of their volume. This has prompted the community to build integrated systems where semantics and data are coupled. The challenge in these systems is to achieve semantic interoperability.

## II. KNOWLEDGE MODELING

### A. The main approaches

Today, databases cover most of biomedical information: patients administrative data, clinical chemistry, clinical diagnostics, images, or even genetic data. The use of this mass of information to improve care and patient safety is still very limited. Literature offers different approaches to address some

of the issues raised above, including: data warehouse approach as in BioWarehouse [16] and BioDWH [17] projects, the integration by views approach in Hemsys [18] and Tambis [19] projects or the so-called mashup in SRS [20] and Integr8 [21] projects. These different approaches offer methods and techniques to solve problems related to access to information regardless of its informative content, ie their semantics.

The increasing use of terminology, in the field of health, or ontologies in health information systems encourages the use of methodologies [22] and technologies from the Semantic Web community.

### B. Ontologies

Artificial intelligence has allowed knowledge to be represented in the form of a domain knowledge base and to automate their use in problem solving. These knowledge bases are generally not reusable which by the way limit their interest. To overcome this problem, the notion of ontology has been introduced [23]. An ontology is seen as a set of concepts for modeling knowledge in a given field. A concept may have several thematic senses. The concepts are linked by semantic relations, composition relations and inheritance. Many researchers have proposed definitions including:

- Guarino introduced the formal ontology notion, defined as a conceptual modeling: "An ontology is an agreement on a shared and possibly partial conceptualization" [24].

- The ontology is defined by Uschold [25] as a formal description of entities and their properties, relationships, constraints and behaviors. Furthermore, the authors introduced the notion of ontology explicit "An explicit ontology may take a variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meaning".

- Thomas R. Gruber [26] which describes ontology as an explicit specification of a conceptualization of modeling concepts and relationships between concepts: "An ontology is a specification of a conceptualization. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general".

- John F. Sowa [27] clarified this concept and defined ontology as a type catalog from the study of categories of abstract and concrete entities that exist or may exist in a domain: "The subject of ontology is the study of the categories of things that exist or may exist in some domain. The product of such a study, called ontology, is a catalogue of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. The types in the ontology represent the predicates, word senses, or concept and relation types of the language L when used to discuss topics in the domain D".

Christophe Roche [28] gave a simple and generic definition that encompasses and summarizes the above definitions "An ontology is a conceptualization of a domain to which are associated one or more vocabularies of terms. The concepts are structured into a system and participate in the meaning of terms. Ontology is defined for a particular purpose and expresses a view shared by a community. An ontology is expressed in language (representation) based on a theory (semantics) that guarantees the properties of the ontology in terms of consensus, consistency, reuse and sharing"

Ontologies are widely accepted as an appropriate form for the conceptualization of knowledge. They represent a basic step in the knowledge representation process which integrates terminology, taxonomy (organization of concepts) and description of relations among concepts and/or classes of concepts.

Using ontology enables appropriate organization of procedural knowledge and that can be beneficial for the implementation and maintenance of any complex system. Ontologies are reusable and facilitate interoperability among the application. They enable easier verification and comparison and ensure comparability of results coming from applications using the same ontology.

Ontologies can be described by meta-languages such as the Unified Modeling Language (UML), expressing the concepts in classes with attributes and operations as well as the interrelations in associations. HL7 Version 3 Normative Edition shows how to map the HL7 data types to the Object Management Group's (OMG) Unified Modeling Language (UML) [29].

### C. What does one represent in ontology?

Ontologies allow representing knowledge and the way to automatically handle it, while preserving their semantics. Knowledge is defined through concepts linked together by relationships. The ontology is then presented, usually in the form of a hierarchical organization of concepts.

Concepts are represented by a set of properties and could be equivalent, not connected or dependent. They can be linked by relations defined as a connection concept between entities, often expressed by a term or a literal symbol or other. We have two types of links: hierarchical and semantic. The hierarchical relationship resumes Hyperonymy / hyponymy structuring, while the semantic relationship links the concepts through a link, said part-whole, which corresponds to the Holonymie / meronymy structuring. A hierarchical relationship links a higher member, said the hypernym element, and a lower member, said hyponym element, having the same properties as the first element with at least one additional one.

As the concepts, relationships can have algebraic properties (symmetry, reflexivity, transitivity). To describe the concepts and relationships of ontology, it is expressed in a language and is based on formalism.

#### 1) The representation formalisms

Ontology, as described above, needs to be formally represented. Moreover, it must represent the semantic relations linking concepts. To this end, much formalism has been developed:

- The diagrams represent complex data structures. They are considered as a prototype describing a situation or standard

object. They provide a benchmark for comparing objects that we wishes to recognize, analyze or classify. The prototypes must consider all possible forms of expression of knowledge. A scheme is characterized by attributes (data structure), facets (the attributes semantics) and relationships (the inheritance semantics).

- Semantic networks represent a graph structure that encodes the knowledge and their properties. The nodes of the graph represent objects (concepts, situations, events, etc..) and edges express relations between these objects. These relations can be links "kind of" expressing the inclusion relation or links "is a" showing the relationship of belonging. It includes a set of concepts describing an area completely. The interest of these graphs is their non-ambiguity and ease of use. This prompted the designers of multiple applications to use them, whether in knowledge acquisition, information retrieval and reasoning about conceptual knowledge.

- A script is a data structure that contains knowledge about a situation and which combines representations. It can be seen as a set of elementary actions or references to other scenarios, ordered according to their sequence in time.

*2) Building ontologies*

The method chosen to build ontology should be strongly guided by the desired type of ontologies and objectives of its use.There are three types of methods for the construction of ontology: manual, automatic and semi-automatic. For the first, experts create a new ontology of a domain or extend an existing ontology. In the automatic method, the ontology is built by knowledge extraction techniques: concepts and relations are extracted and then verified by the inferences. Finally, the semi-automatic, ontologies are automatically built and used to extend ontologies that was built manually.

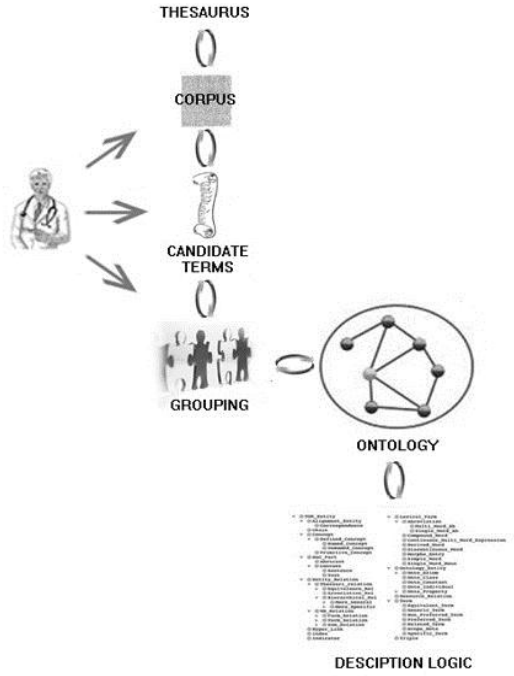



Figure 1. Steps for creating a medical ontology: *intervention at* different phases of the *ontology* development

For the medical field and chronic disease management, the creation of an ontology should go through the following steps (Figure 1):

- Establishment of a corpus of work from a thesaurus using a morpho-syntactic analysis for a list of candidate terms. A tool such as Syntex software, including working on verb phrases is particularly interesting. Furthermore, a study of the context of each candidate term would highlight additional concepts and / or to specify other relationships between concepts.

- Semantic analysis for validating candidate terms as a term of the domain by a medical expert. It would facilitate the grouping of terms validated in concepts, defining relations between concepts and between symptoms and function.

- Structuring by semantic groups.

- Finalization of the process in a language, based on description logics.

Process of designing the ontology begins after a language and a tool have been selected. There are two standard approaches to the ontology design: bottom-up approach (smaller parts of the ontology are constructed first and then later intagretaed) and top-down approach (design upper classes and the developp small pieces of the hierarchy). Though, probably the best way of creating an ontology is to combine both approaches in an iterative way.

## III. REQUIREMENTS FOR MONITORING PATIENTS WITH CHRONIC DISEASES

The home care solutions (Figure 2) support usually innovative technological aspects (monitoring and remote transmission of vital signs, detect falls, alarms ...) and organizational aspects necessary for the coordination of different factors contributing to remain at home [30]. These solutions are still largely at an experimental stage, in order to assess the relationship between cost, reliability, the medical service and economies of scale they are likely to make to the health system overall.

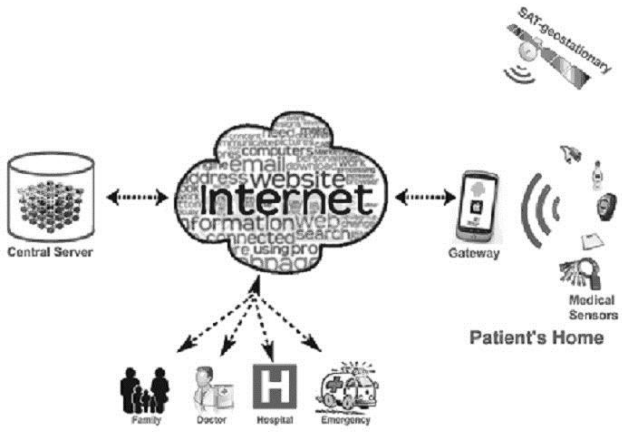



Figure 2. An evolving architecture, encompassing a full set of domestic and medical devices along with the analysis and interpretation of data

Most of these systems are designed and intended to follow a particular chronic disease (heart failure, respiratory failure, diabetes ...) and require fairly large technological equipment (sensors, computers located at home, set-top boxes for the transmission information ...). However, older people are multi-disease, with both of several chronic disorders related to age. Technically and economically it can't be considered to increase the monitoring systems number. It is therefore necessary to focus on the interoperability of these systems, so as to factor out the common elements, thereby reducing costs of deployment and operation.

Elderly patients are often multi-pathological so currently it is necessary to multiply the patient's home systems. This profusion of systems has little interest because most of them uses similar equipment to perform their measurements. Motion sensors for example are found in virtually all existing solutions.

To avoid this multiplication of equipment we must adopt architecture to pool them. Thus, even a motion sensor can be used by different applications. This pooling of equipment meets both an economic need but also a demand of patients seen in general who wish to limit the proliferation of such equipment in their homes.

In practice this mutualization and this consideration multi-phatologiques will translate by monitoring platforms necessarily evolving which can therefore integrate knowledge about various diseases. Ontologies represent then formalism well adapted to enable the integration of new knowledge and / or to make available the knowledge.

### A. *The heterogeneity of medical knowledge*

The information and resources used to consider and treat various diseases are necessarily heterogeneous and make their understanding and analysis very difficult. Meaning preservation of information shared is then an important problem. This is what is called semantic interoperability. A commonly accepted definition for semantic interoperability, *"it gives meaning to the information shared and ensures that this is common sense in all systems between which exchanges must be implemented"* [31-33]. Consideration of this semantics enables distributed systems to combine received information with local information and treat all consistently.

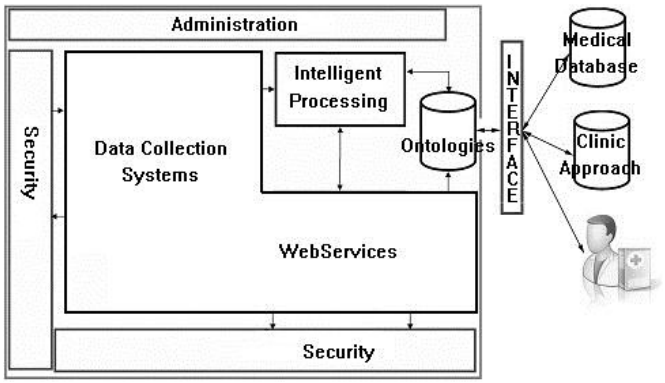



Figure 3. Ontologies are an effective way for the representation and the sharing of knowledge

To ensure semantic interoperability, information shared between systems (Figure 3) must first be described in a formal structure for preserving its semantics. This is a recurring problem in the field of knowledge engineering, where methodologies and techniques are proposed to collect, identify, analyze, organize and share knowledge between different entities. Among these techniques, ontologies are experiencing a rapid development over the ten years past and appear as an effective way for knowledge representation.

### B. Techniques for semantic interoperability.

A number of techniques have been proposed in the literature to achieve interoperability [34]. They are often used to allow data sharing between heterogeneous knowledge bases and for the re-use of these bases.

We can distinguish three main categories which are:

- The alignment of ontologies [35], for whom the goal is to find correspondences between ontologies. It is usually described as an application of the MATCH operator [65], whose input consists of a set of ontologies and output, formed correspondences between these ontologies

- The mapping of ontologies which allows, for example, to query heterogeneous knowledge bases using a common interface or transforming data between different representations.

- The merging of ontologies, which creates a new ontology, called the merged ontology with the knowledge of the original ontologies. The challenge then is to ensure that all correspondences and differences between ontologies are properly reflected in the resulting ontology.

Generally speaking, providing semantic interoperability among heterogeneous ontologies is still primarily a semi-automated process.

## IV. DISCUSSION

Ontologies are necessary to both facilitate goals semantic structuring with their relations and take into account the heterogeneity of knowledge in a growing field such as monitoring patients at home and especially patients with chronic diseases. Their increasing use in this area, leads to in significant availability of ontologies that drives us to think about when to re-use them.

It is then important to take into account both their popularity and also the simplicity of their implementation. The ratio of these two parameters allows us to gauge the interest to investigate their interoperability.

Ontology engineering and management have to encompass the entire ontology lifecycle: creation, coordination and merging [37]. Merging tools or alignment of ontologies allow the integration of information from a distributed environment or heterogeneous systems. It is essential to establish semantic correspondences between ontologies that describe this shared information.

The role of alignment tools [1] is to search for matches between the concepts of distinct ontologies, to allow the joint consideration of the resources they describe. This is to

combine the techniques and methods of matching linguistic, syntactic, semantic or structural. Reference [38] enriched ontologies based on thin semantic analysis of concept of labels and in the fact that regularities exist in the way of naming them. These naming conventions are used to establish mappings between these labels and axioms of the ontology, which makes semantic information explicit and then use it to automatically reason above.

Currently there exist a variety of heuristics and other techniques that can be utilised for semantic interoperability, but there is still plenty of scope for refinement and for providing fully automated frameworks.

## V. CONCLUSION

In order to provide a consistent solution in the field of medical telemonitoring, monitoring systems must take into account different pathologies in order to avoid duplication of equipment. They must therefore be open and scalable to allow the sharing and management of heterogeneous knowledge.

Ontologies are particularly suited for understanding, sharing and integrating information. However, various problems are still open, others appear: design method ontologies, representation and reasoning on ontologies, automatic generation of ontologies, ontology alignment and development, representation and data persistence based ontological systems integration based ontological design databases accessible from ontologies, integration of blur in ontologies, etc..

Beyond the issues raised by the heterogeneity of available data, the sequence of algorithmic processes that can exploit this data represents a scientific and technical challenge.

## REFERENCES

[1] Simon, P.; Acker, D. La place de la télémédicine dans l'organisation des soins. Rapport du Ministère de la santé et des sports – Direction de l'Hospitalisation et de l'Organisation des Soins. 2008.

[2] Lupon, J.; http://www.acmcb.es/files/425-933-DOCUMENT/Lupon-41010-7.pdf

[3] Helal, S. ; Mann, W. ; El-Zabadani, H. ; King, J. ; Kaddourra, Y.; Jansen, E. ; The Gator Tech Smart House: A Programmable Pervasive Space », In Computer, 2005, 38, 50-60.

[4] Chan, M.; Campo, E.; Estève, D. PROSAFE, a multisensory remote monitoring system for the elderly or the handicapped in Independent Living for Persons with Disabilities and Elderly People In Proceedings of the ICOST'2003, Paris, France, September 2003, pp. 89-95.

[5] Ogawa, R.; Togawa, T.; Attempts at monitoring health status in the home », In the Proceedings of the 1st Annual International Conference On Microtechnologies in Medicine and Biology, Lyon, France, October 2000, 52-556.

[6] Tamura, T.; Togawa, T.; Ogawa, M.; Yoda, M. Fully automated health monitoring system in the home, In Medical Engineering & Physics, 1998, 20, 8, 573-579.

[7] Williams, G.; Doughty, K.; Bradley, D. A. A systems approach to achieving CarerNet - an integrated and intelligent telecare system, In IEEE Transactions on Information Technology in Biomedicine, 1998, 2, 1, p. 1-9.

[8] Aniello Minutolo, Giovanna Sannino, Massimo Esposito, Giuseppe De Pietro. 2010. A rule-based mHealth system for cardiac monitoring. 2010 IEEE EMBS Conference on Biomedical Engineering & Sciences.

[9] Federica Paganelli, Dino Giuli. 2011. An Ontology-Based System for Context-Aware and Configurable Services to Support Home-Based Continuous Care. IEEE Transactions on Information Technology in Biomedicine, 2011, 15, 2.

[10] Miguel A. Valero, Laura Vadillo, Iván Pau, and Ana Peñalver. 2009. An Intelligent Agents Reasoning Platform to Support Smart Home Telecare. IWANN 2009, Part II, LNCS 5518. 679–686.

[11] F. Latfi, B. Lefebvre and C. Descheneaux. 2007. Le rôle de l'ontologie de la tâche dans un Habitat Intelligent en Télé-Santé. 1ères Journées Francophones sur les Ontologies JFO, Sousse, Tunisie, 18 - 20 Octobre 2007.

[12] F. Latfi, C. Descheneaux, B. Lefebvre. 2007. Habitat intelligent en télé–Santé : ontologie de l'équipement. FICCDAT, 16-19 Juin. Toronto, Canada, 2007.

[13] Boudy, J. ; Baldinger, J. ; Delavaut, F. ; I., B.; Dorizzi, B. ; Farin, I. Télévigilance médicale au service du patient à domicile In Proceedings of the ASSISTH'2007, Toulouse, France, November 2007, 1, 397-40.

[14] Noury N. AILISA : Plateformes d'évaluations pour des technologies de télésurveillance médicale et d'assistance en gérontologie, In Gérontologie et Société, 2005, 113, 97-119.

[15] Rialle, V.; Lamy, J.B.; Noury, N.; Bajolle, L. Telemonitoring of patients at home : a software agent approach , In Computer Methods and Programs in Biomedicine, 2004, 72, 3, 257-26.

[16] Lee, T.; Pouliot, Y.; Wagner, V.; Gupta, P.; Stringer-Calvert D. W.J.; Tenenbaum, J. D.; Karp, P. D. BioWarehouse: a bioinformatics database warehouse toolkit In BMC Bioinformatics, 7, 170, 2006.

[17] Töpel, T.; Kormeier, B.; Klassen, A.; Hofestädt, R. BioDWH: A Data Warehouse Kit for Life Science Data Integration, Journal of Integrative Bioinformatics, 2008, 5, 2.

[18] Pillai, S.; Gudipati, R.; Lilen, L. Design issues and an architecture for a heterogenous multidatabase system In Proceedings of the 15th ACM Computer Science Conference, St. Louis, Missouri, USA, February 1987.

[19] Goble, C. A.; Stevens, R.; Ng, G.; Beshhofer, S.; Paton, N. W.; Baker, P. G.; Peim, M.; Brass, A. Transparent access to multiple bioinformatics information sources In IBM Systems Journal, 2001, 40, 2, 532-551.

[20] Etzold, T.; Argos, P. SRS – an indexing and retrieval tool for flat file data libraries In Computer Applications in the Biosciences, 1993, 9, 1, 49-57.

[21] Kersey, P.; Bower, L.; Morris, L.; Horne, A.; Petryszak, R.; Kanz, C.; Kanapin, A.; Das, U.; Michoud, K.; Phan, I.; Gattiker, A.; Kulikova, T.; Faruque, N.; Duggan, K.; Mclaren, P.; Reimholz, B.; Duret, L.; Penel, S.; Reuter, I.; Apweiler, R. Integr8 and genome reviews: integrated views of complete genomes and proteomes *In Nucleic Acids Res*, 2005,. 33, 297-302.

[22] Blobel, B. & Oemig, F. Ontology-driven health information systems architectures. Stud Health Technol Inform 150, 195-199 (2009), doi:10.3233/978-1-60750-044-5-195.

[23] Charlet, J. ; Bachimont, B.; Troncy, R. Ontologies pour le web sémantique In Revue I3, 2004, Hors série Le Web sémantique, 43-63.

[24] Guarino, N.; Poli, R. Formal ontology in conceptual analysis and knowledge representation. In Special issue of the International Journal of Human and Computer Studies, 1995, 43, 5/6, 625–640.

[25] Uschold, M.; Gruninger, M. Ontologies: Principles, Methods and Applications. In Knowledge Engineering Review, 1996, 11, 2, 93 - 155.

[26] Gruber, T. R. A Translation Approach to Portable Ontology Specifications In Knowledge Acquisition, 1993, 5, 2, 199-221.

[27] Sowa, J. F. Building, Sharing, and Merging Ontologies. In http://users.bestweb.net/~sowa/ontology/ontoshar.htm, 2001.

[28] Roche, C. Terminologie et ontologie. In Revue Langages, 157, 2005.

[29] Health Level 7, Inc, http://www.hl7.org.

[30] Ahmed Benyahia, A. ; Hajjam, A. ; Hilaire, V.; Hajjam, M. Ontological architecture for management of telemonitoring system and alerts detection. In A. Hajjam (eds), eHealth and remote monitoring, ISBN 980-953-307-305-2, 2012.

[31] Charlet, J. L'Ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Habilitation à diriger des recherches, Université Paris 6, 2002.

[32] Jouanot, F. Un modèle sémantique pour l'interopérabilité de systèmes d'information. In INFORSID, 2000. 347–364.

[33] Vernadat, F. Interoperable enterprise systems: Architectures, methods and metrics. In Rapport technique, LGIPM, Université de Metz, France, 2007.

[34] Choquet, R.; Teodoro, D.; Mels, G.; Assele, A.; Pasche, E.; Ruch, P.; Lovis, C.; Jaulent, M.C. Partage de données biomédicales sur le web sémantique, Proceedings of the IC'2010, Nîmes, France, june 2010.

[35] Scharffe, F. ;Fensel, D. Correspondence patterns for ontology alignment. In Proceedings of the 16th EKAW '08, Berlin, Heidelberg, 2008, 83–92.

[36] Shvaiko, P.; Euzenat, J. Ontology matching: state of the art and future challenges**.** In IEEE Transactions on Knowledge and Data Engineering, 2012.

[37] Rebstock, M., Fengel, J., Paulheim, H. (2008) Ontologies-Based Business Integration. Springer-Verlag, Berlin.

[38] Fernandez-Breis, J. ; Iannone, L. ; Palmisano, I. ; Rector, A. ; Stevens, R. Enriching the gene ontology via the dissection of labels using the ontology pre-processor language. In Knowledge Engineering and Management by the Masses, 2010, 6317, 59–73.

AUTHORS PROFILE

**Amir HAJJAM** received his Ph.D. in Computer Science from University of Haute-Alsace (UHA) in 1990. He was with the UHA from 1990 to 2002. He joined the University of Technology of Belfort-Montbeliard (UTBM) in 2002 and performs research activity at the Systems and Transportation Laboratory (SeT). His current research interest includes heuristic mechanisms and Artificial Intelligence applied in a distributed dynamic framework to networks, telecommunications, intelligent transportation services and e-Health. He is responsible for the research area of e-Health. He is the author of 3 books and many international publications in refereed journals and conferences. He is Editorial Board Member of many international journals and serves on the steering committee of many workshops and conferences. He is an expert for various organizations such as the Association Nationale de la Recherche et de la Technologie (ANRT France), Agence Régionale de l'Innovation d'Alsace (ARI Alsace) and Centre National pour la Recherche Scientifique et Technique du Maroc (CNRST Morocco).

# E-commerce Smartphone Application

Abdullah Saleh Alqahtani
School of Computer Science, Engineering and Mathematics, Faculty of Science and Engineering,
Flinders University,
Adelaide SA 5001, Australia

Robert Goodwin
School of Computer Science, Engineering and Mathematics, Faculty of Science and Engineering,
Flinders University,
Adelaide SA 5001, Australia

***Abstract*— Mobile and e-commerce applications are tools for accessing the Internet and for buying products and services. These applications are constantly evolving due to the high rate of technological advances being made. This paper provides a new perspective on the types of applications that can be used. It describes and analyses device requirements, provides a literature review of important aspects of mobile devices that can use such applications and the requirements of websites designed for m-commerce. The design and security aspects of mobile devices are also investigated. As an alternative to existing m-commerce applications, this paper also investigates the characteristics and potential of the PhoneGap cross-mobile platform application. The results suggest that effective mobile applications do exist for various Smartphones, and web applications on mobile devices should be effective. PhoneGap and Spree applications can communicate using JSON instead of the XML language. Android simulators can be used for ensuring proper functionality and for compiling the applications.**

***Keywords- E-commerce ; PhoneGap ; M-commerce ; Smartphones ; Spree –commerce ; Ruby on Rails.***

## I. INTRODUCTION

The internet has changed many aspects of society, from business to recreation, from culture to communication and technology, as well as shopping and travelling. This new form of communication has provided new ways of doing business with the help of technological development. E-commerce is the new way of shopping and doing business. Technology has allowed companies to promote and sell their products on new markets, overcoming geographical borders as never before. Consumers have access to a wider market of products when they use wireless and internet technologies. Mobile devices with wide access to the Internet have allowed companies to reach consumers in more diverse ways, thus ensuring deep market penetration.

This study investigates the opportunities generated through mobile telephone access to the Internet [13]. Faster wireless networking standards allow wireless devices to use more e-commerce applications, and consequently, permit wider access to mobile commerce (m-commerce) [14]. M-commerce has been defined as "a special branch of e-commerce, in which mobile devices and their network connection medium are used to buy, sell, and promote products, services, and information" [20]. According to Koukia, Rigou and Sirmakessis (2006), wireless technologies have improved traditional e-commerce by "providing the additional aspects of mobility (of participation) and portability (of technology)." On this theme, mobile and e-commerce application developments are an important factor for the expansion of m-commerce among consumers. The technical characteristics of devices and corresponding applications, as well as Internet access facilities, are determining the level of acceptance of m-commerce and its development. Aspects like processing power, display and device size, mobile internet coverage, standardization and quality of devices, are only some of the important factors that decide the level of use of m-commerce, and consequently, the level of its development [15].

The purpose of designing interfaces for mobile applications should be to increase consumers' interest in using and dedication to m-commerce. Among the inhibiting factors is that m-commerce applications were developed based on e-commerce applications. The most important thing when designing such applications is to design the application in such a way that it does not distract the user from the main purpose of the application [15]. However, aspects concerning security and accessibility should not be neglected. Even though storing sensitive data such as medical, financial, or personal information on mobile devices can help people, the risks of losing such information or of unauthorized access are higher and should be considered when an m-commerce transaction begins[16].

This paper will review the latest trends in mobile and e-commerce applications and will develop an application architecture that describes the internal architecture of both web and mobile components. Moreover, the focus will be on developing a more sophisticated demonstration mobile application regime that will employ web-services to communicate with web servers. Furthermore this paper will discuss the main characteristics of devices used for m-commerce, the available survey design guidelines, and the important role of these characteristics for increasing the potential of m-commerce will be articulated.

This paper will also investigate the requirements of e-commerce applications and why normal websites are not suitable for mobile devices. More specifically this paper will analyse the characteristics required for websites so that they function properly on mobile devices. These characteristics are screen size, input device, task-based interfaces for mobile devices, m-loyalty, design aesthetics and website design. Other research questions relate to the reason why mobile native applications are preferred over mobile websites and

what can be achieved using PhoneGap cross-mobile platform applications. Finally, the paper will attempt to provide solutions for mobile application development and make recommendations for future directions.

## II. RESEARCH METHODOLOGY

A literature review on the two major issues of e-commerce mobile applications has been conducted, these being: firstly, the interface usability of mobile applications; and secondly, design and security considerations. The literature review concluded that mobile applications must effectively operate on different Smartphones and have the ability to use different ecommerce web applications through web services.

### A. Research Procedure

The main question answered here is: "How can we exploit the usability and security of e-commerce application(s) for mobile devices (m-commerce) with maximized mobile platform independence?"

The sub-questions indicated below should be answered in order to fully explain the main themes in this paper :

1) Why is mobility required for e-commerce application?
2) Why are normal websites not useful for mobile devices?
3) Why are mobile native applications preferred over mobile websites?
4) What can we achieve using the PhoneGap (cross-mobile platform application development framework)?
5) How will the initial application architecture use Spree-commerce, PhoneGap and web applications?
6) What are the different solutions for developing mobile applications?

In this paper the focus was also on developing a prototype mobile application. The following steps were taken in order to complete a demonstration application:

1) Understanding how the Spree commerce system was developed and how it works.
2) Improved the application architecture based on more research and actual development experience gained while developing the application.
3) Developed a more sophisticated demonstration mobile application that will communicate with web servers through web services.

a) Developed of the PhoneGap application and tested it on iOS and Android

Page constraints have limited the paper's scope to actual functionality of the application. This paper will therefore discuss basic functionality.

*Step 1*: In order to achieve the first goal based on the web server was set up a Spree-commerce Ruby on Rails application. Spree is basically an open source e-commerce system and it has the ability to deal with web services in obtaining a list of products, product details and cart system.

*Step 2*: A mobile application gap via PhoneGap (mobile application development framework). Then product listing and view page was initiated through managing data from web services

*Step 3*: The application was tested on different Smartphones as follow:

1) Test on iPhone Simulator
2) Test on iPad Simulator
3) Test on Android Simulator
4) Test on actual iPhone device by paying 100$ apple developer account

### B. Software development methodology

Based on the questions developed for this study a general search was conducted in the first phase of the research. Software architecture was developed based on this research. The results of the research enabled the initial scope of the research to be more precise and achievable. Various software development methodologies were utilized since software engineering is a diverse field and encompasses many diverse factors and contexts. Experimental software engineering was used in the initial development phase so that risk and uncertainty were reduced. An iterative software model was used as it is the best choice for prototype development.

### C. Application Architecture

The application architecture is explained in terms of how: the Spree web application is hosted on the Ruby on Rails equipped webserver. Web services have been built inside the Spree application; and the mobile phone native applications built using PhoneGap which communicates with the webserver through web services to obtain data and information. (Figure1). This diagram shows how the application architecture used the Spree web server and a native mobile application using PhoneGap which allows the same application to run in different mobile operating systems. This corresponds with the web server which uses web services to obtain data.

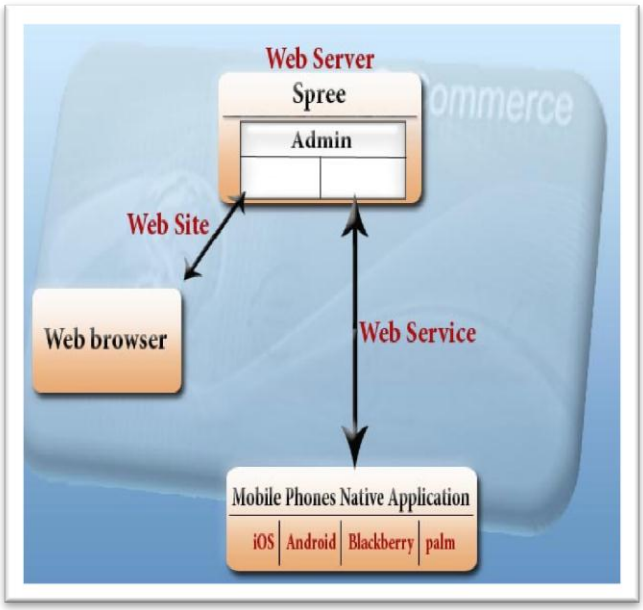



Figure 1: Application Architecture

## III. SCOPE

### A. Scope of application

The basic functionality delivered by the application features two main factors, product listing and a product detail page.

Two changes were made,

- Spree web services were modified in order to show complete data in our mobile application.
- The functionality needed so that the mobile application can list products by requesting web services provided by the Spree-commerce application was added.
- The functionality of the mobile application was improved to show complete information on a product by requesting a second web-service from the Spree-commerce application.

Best practices were employed for making the user interface of the mobile application functional.

## IV. RESULT

### A. User/System scenario

- The process starts when a user starts the mobile application.
- The mobile application requests the product listing web service.
- The web application (Spree-commerce application), which operates on remote web servers receives product listing web service requests.
- The web application finds the published products from the database and prepares JSON response.
- The mobile application receives JSON response and will convert it into HTML and render it. The mobile application also makes sure that all links in the product list should only work as AJAX.
- When the user selects a product the mobile application sends on AJAX request to the server for the product details web service.
- The server finds full details of the product and sends JSON response.
- The mobile application prepares HTML from JSON response for the display on the screen.

### B. iOS and Android Demo Application

The application was tested on Android and iOS (iPhone/iPad/iPod operating system). The application provided two web services. Figures 2 and 4 depict the web service which enables the user to see a list of products available. Figure 3 shows that when you tap (select) any of these products, a new detailed page relating to the product selected will open and display information such as image, colour and price.

#### 1) *iPhone simulator tested*

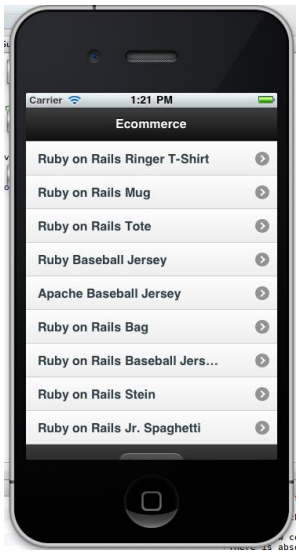



Figure 2: List of products on iPhone


Figure 3: Product detail screen on iPhone

#### 2) *Android simulator tested*

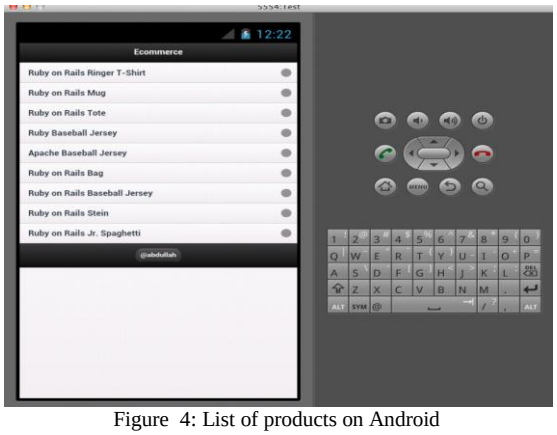



Figure 4: List of products on Android

## V. SET UP AND SETTING UP

### A. Setting up Spree web application

*1) Before starting the Spree application the following applications must be installed:*

*a) Unix bases OS*

*b) Mysql5*

*c) Ruby 1.8.7 or 1.9.2*

*d) Ruby on Rails 3.0.9*

*e) ImageMagick (library for image processing; used for generating different sizes of images)*

*2) Once these applications are installed then the following steps are needed:*

*a) Copy the ecommerce application code.*

*b) On UNIX terminal go to ecommerce application folder*

*c) Run bundle install [Make sure machine is connected to internet because this command will install all dependencies for the application]*

*d) rake db: create [ It will create MYSQL database]*

*e) rake db: migrate [ It will create all tables]*

*f) rake db: bootstrap [ It will load sample data in database]*

*g) Rails S [ run web application server]*

*h) Use frontend and admin panel of Spree application for exploration*

### B. Mobile applications

*1) Deploy to iOS:*

*a) Run Xcode and open iOS version of PhoneGap project*

*b) Deploy to actual iOS device*

- Add your device to apple provisioning portal (remember paid developer account is needed)
- Select device to be used in execution list and Click run button

*c) Or run on iPhone simulator*

- Select iPhone simulator in execution list and click run button

*2) Deploy to Android:*

*a) Download and install Eclipse Classic.*

*b) Download and install Android SDK..*

*c) Download and install ADT Plugin for Eclipse.*

*d) Download the latest copy of PhoneGap and extract its contents.*

*e) Setup Project.*

*f) Merge our HTML code to be I used.*

## VI. DISCUSSION RESULT

E-commerce functionality was provided by installing a Spree application on a web server. The research suggested that mobile phones such as Android-based Smartphones, iPad and iPhones were appropriate for installing the PhoneGap application on them. The research also illustrated that using JSON-based web services promoted data communication between PhoneGap and the Spree application. The native support of JSON in JavaScript and it being comparatively lightweight was the significant reason why this report suggested JSON should be selected instead of the XML data format.

The actual Spree application provided several web services but the research forewarned that these web services would not be feasible with every application. For instance, often there was too much unneeded information on a products listing web service. For this reason and that of network latency, not only did an application slow down but in response to the JSON object. The processing of JSON in JavaScript was increased. The research also suggested that the required data such as the ID, title, the short description and slug were obtained by overriding that particular web service.

As the research elaborated, depending on the screen size of mobile phones, it was often the case that image sizes for the Spree application were either too large or too small. Fortunately, the paperclip library that is available in the Spree application was able to regenerate all the images after their smaller sizes were introduced. The research also suggested that for the UI components of a Smartphone, jQuery and jQueryMobile should be employed. The research also illustrated the fact that no image was used for UI components but rather CSS3 features were used, making this library the preferred option. The research indicated that jQuery guidelines were used to create a robust code and jQunit, which is a jQuery-based unit testing library dedicated to unit testing.

According to the research a PhoneGap template was used to make an Xcode project once this functionality became functional on WebKit based browsers such as Chrome and Safari. CSS, HTML and JavaScript were embedded into this project template. The research suggested that iPhone and iPad simulators had to be used for testing until it was definite that they worked appropriately. Once they indeed worked appropriately, the application was then tested on the actual device by purchasing a $100 Apple developer account.

The research highlighted the fact that Android Smartphones followed a similar process. Android development and the Eclipse IDE plug-ins had to be installed in this case. The research suggested that an Android template was then used to make an eclipse project, merging CSS, HTML and JavaScript into the template as well. Ultimately, the research indicated that the Android simulator had to be utilized to compile and test the application. This was the process used to execute the application and ensure its success.

## VII. CONCLUSION AND FUTURE DIRECTIONS

Mobility of people and technologies are key factors in today's economy. Mobile applications are of utmost importance when companies market their products or services. Mobile phones have generated an incredible opportunity for accessing the Internet, while m-commerce has increased the level of using a mobile phone for business. Applications for electronic and mobile commerce are, however, developed

sufficiently to cover all aspects of the market. The ideal applications do not distract the user from his/her intent and they provide on appropriately level of security, accessibility and speed. Screen size, input device, urgency, task based interfaces, and design are important aspects that developers have to consider when designing mobile phone applications. Such applications should be implemented on different operating systems, whether using JavaScript, CSS3 or HTML5, or combining these three together. The installation of a Spree application allows partial web services to function because not all applications may work with these web services. The device screen size proved to be of great importance but the Spree application solved these issues. Simulators for iPhone and iPad represent an excellent way to test applications.

### A. Future directions

In order to bring this research from its current analysis on the basic model to an advanced level, the points listed under the scope of excluded and discussed below must also considered.

Payment transactions need to be made through a more secure and safe channel like Pay-Pal, which will ensure a safe and sound transaction system. Transaction payments should be password protected or should have a PIN code that confirms its reliability. Even if the gadget used - for example a cell phone - is lost, stolen or otherwise incapable of being used, consumers would not have to worry about theft or mismanagement of money. Furthermore, they can obtain a password to improve protection and security.

Products listed on the product information pages should have paged display pictures, which will be more appealing to the viewer. The product should be listed with information such as the product specifications: price, part number, technical details, features and packaging details. Yet some product information pages do not provide this information which is very inconvenient for the consumer and does not present a professional image.

The applications should allow users to change the web services' URLs. Settings should be made flexible so that users can improvise and access the web services they require by changing settings. The application should allow the user to add multiple sources for the same web server by cell phone so that the user can code flex, which means that a user can develop and deploy cross-platform Internet applications being run on cell phones. This implies that a mobile application should be able to locate a list of products from different remote applications, which provide the same web services being supported by a mobile application.

## REFERENCES

[1] Ngai, E.W.T. & A. Gunasekaran (2007). A review for mobile commerce research and applications. *Decision Support Systems*, vol. 43, no. 1, pp. 3-15.

[2] Cyr, D., M. Head & A. Ivanov (2006). Design aesthetics leading to m-loyalty in mobile commerce. *Information & Management*, vol. 43, no. 8, pp. 950-963.

[3] [Sumita, Ushio & Jun Yoshii (2010). Enhancement of e-commerce via mobile accesses to the Internet. *Electronic Commerce Research and Applications*, vol. 9, pp. 217-227.

[4] Lee, Ching-Chang, Hsing Kenneth Cheng & Hui-Hsin Cheng (2007). An empirical study of mobile commerce in insurance industry: Task–technology fit and individual differences. *Science Direct*, vol. 43, 2007, pp.95-110.

[5] Wu, Jen-Her & Yu-Min Wang (2006). Development of a tool for selecting mobile shopping site: A customer perspective. *Science Direct*, vol. 5, pp. 192-200.

[6] Wu, Jen-Her & Shu-Ching Wanga (2005). What drives mobile commerce? An empirical evaluation of the revised technology acceptance model. *Information & Management*, vol. 42, pp. 719-729.

[7] Yung-Ming, Li & Yung-Shao Yeh (2010). Increasing trust in mobile commerce through design aesthetics. *Computers in Human Behavior*, vol. 26, pp. 673-684.

[8] Chang, Yung-Fu & C.S. Chen (2005). Smart phone—the choice of client platform for mobile commerce. *Computer Standards & Interfaces*, vol. 27, pp. 329-336.

[9] Barnes, Stuart & Eusebio Scornavacca (2007). Chapter 7: The emergence of mobile commerce. In Stuart Barnes, *E-Commerce and V-Business: Digital Enterprise in the Twenty-First Century*, 2nd ed. Oxford, U.K.: Butterworth-Heinemann, pp. 157-178.

[10] Lehtinen, Toni (2011). Native versus Web - which approach is best for mobile apps?, *mobilesolutions*, White Paper, 2011, pp. 2-6.

[11] Power, Mark (2011). Mobile Web Apps: A Briefing Paper. Centre for Educational Technology & Interoperability Standards, March. Mark Power@CETIS.

[12] Descartes (2010). Why PhoneGap, 4p., *Descartes* website. Retrieved July 28, 2011 from http://www.phonegapmobileappdev.com/wp-content/uploads/Why-PhoneGap.pdf.

[13] Soriano, M. & D. Ponce (2002). A security and usability proposal for mobile electronic commerce. *Communications Magazine*, vol. 40, no. 8, pp. 62-67.

[14] Buranatrived, J. & P. Vickers, P (2002). An investigation of the impact of mobile phone and PDA interfaces on the usability of mobile-commerce applications. *Proceedings of the IEEE 5th International Workshop on Networked Appliances*. Liverpool: pp. 90-95.

[15] Koukia, Spiridoula, Maria Rigou & Spiros Sirmakessis (2006). The Role of Context in m-Commerce and the Personalization Dimension. In *International Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops.* Hong Kong: pp. 267-276.

[16] Leavitt, N. (2010). Payment Applications Make E-Commerce Mobile. *Computer*, vol. 43, no. 12, pp. 19-22.

[17] Github Inc. (2011). Github Social Coding: spree, *Github Social Coding* website. Retrieved Jun 28, 2011 from https://github.com/spree/spree.

[18] Spree (n.d.). The World's Most Flexible E-Commerce Platform, *Spree* website. Retrieved August 3, 2011 from http://spreecommerce.com/.

[19] Nitobi (2011). Home page, *PhoneGap* website. Retrieved August 20, 2011 from http://www.phonegap.com/.

[20] Kurvosky, Stan, Vladimir Zanev & Anatoly Kurkovsky (2005). SMMART: using context-awareness in m-commerce. In MobileHCI '05: Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Services & Devices. New York: ACM, pp. 383-384.

[21] PocketHacks.com (2011). iOS page, *All about Windows Mobile, Windows Phone 7 and Android devices, latest news, freeware apps and hacks.*. Retrieved August 20, 2011 from http://pockethacks.com/ios/.

[22] ImageMagick Studio LLC (2011). About Imagemagick. *Imagemagick.org* website. Retrieved October 5, 2011 from http://www.imagemagick.org/script/index.php.

[23] Nitobi (2011). How PhoneGap Works. *PhoneGap* website. Retrieved August 20, 2011 from http://www.phonegap.com/about.

[24] TechTerms.com (2010). Android. *TechTerms* website. Retrieved September 17, 2011 from http://www.techterms.com/definition/android.

[25] PC Magazine (2011). Definition of AJaX. *PC Magazine* website. Retrieved August 11, 2011 from http://www.pcmag.com/encyclopedia_term/0,2542,t=AJaX&i=55346,00.asp#fbid=cGwEB1ODoJ6.

[26] The jQuery Project (2011). Homepage, *jQuery Mobile Framework* website. Retrieved September 5, 2011 from http://jquerymobile.com/.

[27] Internet.com. (2011). jQuery, in *Webopedia* website. Retrieved September 7, 2011 from http://www.webopedia.com/TERM/J/jQuery.html.

[28] About.com (2011). JSON: JavaScript Object Notation, *About.com* website. Retrieved August 3, 2011 from http://javascript.about.com/library/bljson.htm.
[29] Motive Ltd. (2010). The Motive Internet Glossary: MySQL. *Motive Glossary* website. Retrieved August 3, 2011 by http://www.motive.co.nz/glossary/mysql.php.
[30] Howard, Alexander B. (2006). Definition: Ruby on Rails (RoR or Rails), *SearchSOA.com* website. Retrieved July 3, 2011 from http://searchsoa.techtarget.com/definition/Ruby-on-Rails.
[31] Linux Information Project (2005). Unix-like Definition [updated June 18, 2006], *Linfo* website. Retrieved July 1, 2011 from http://www.linfo.org/unix-like.html.

# SW-SDF Based Personal Privacy with QIDB-Anonymization Method

Kiran P
Research Scholar
VTU,Belgaum
Karnataka, India

Dr Kavya N P
Prof & Head
Dept of MCA,RNSIT
Bangalore,Karnataka,India

***Abstract*— Personalized anonymization is a method in which a guarding node is used to indicate whether the record owner is ready to reveal its sensitivity based on which anonymization will be performed. Most of the sensitive values that are present in the private data base do not require privacy preservation since the record owner sensitivity is a general one. So there are only few records in the entire distribution that require privacy. For example a record owner having disease flu doesn't mind revealing his identity as compared to record owner having disease cancer. Even in this some of the record owners who have cancer are ready to reveal their identity, this is the motivation for *SW-SDF* based Personal Privacy. In this paper we propose a novel personalized privacy preserving technique that over comes the disadvantages of previous personalized privacy and other anonymization techniques. The core of this method can be divided in to two major components. The first component deals with additional attribute used in the table which is in the form of flags which can be used to divide sensitive attribute. Sensitive Disclosure Flag (SDF) determines whether record owner sensitive information is to be disclosed or whether privacy should be maintained. The second flag that we are using is Sensitive Weigh (SW) which indicates how much sensitive the attribute value is as compared with the rest. Second section deals with a novel representation called Frequency Distribution Block (FDB) and Quasi–Identifier Distribution Block(QIDB) which is used in anonymization. Experimental result show that it has lesser information loss and faster execution time as compared with existing methods.**

***Keywords- Privacy Peserving Data Mining(PPDM);Privacy Preserving Data Publishing(PPDP); Personal Anonymization.***

## I. INTRODUCTION

Personal information present in different organizations can be used by research for understanding patterns there by achieving betterment of the community. For example a personal detail of the patient is present in different hospitals, this information can be used by researchers to understand the patterns for a particular disease and hence improve the identification of the diagnosis. The raw data present in hospitals contain detailed information regarding the patient like name, address, DOB, zip code, symptoms & disease. From this raw data, details regarding name and address which are considered personal are removed before it is given to Data Recipient and this information is also called Microdata. This microdata however contains details like zip, DOB that can be linked with other external publicly available data bases for re-identification of sensitive value.

This re-identification of the record by linking public data to Published data is called as linking attack. For example consider the details of the patient Published by the hospital in table 1, which does not contain details regarding name, address and other personal information. The attacker can use the publicly available external data base shown in table 2 and join these details with table 1 thereby personal details can be revealed. The query may look like

```
SELECT NAME, DISEASE
FROM VOTERS_TABLE AS V, PAIENT_TABLE AS P
WHERE V.ZIP=P.ZIPAND V.AGE=P.AGE;
```

The result of this query gives me entire details regarding sensitive information i.e. disease and the identity of the individual which is of great concern because the individuals are not ready to share their sensitive information. The join may give me a value <RAMA, Gastric ulcer > for zipcode 48677 & age 26 and is called Record Level Disclosure. The approaches used by researchers to mask sensitive data from Data Recipients come under a category called Privacy Preserving Data Publishing (PPDP). Attributes present in Published Patient Data that can be linked to external publicly available data bases like ZIP, DOB,… are called Quasi-Identifier (Q) attributes.

TABLE 1. PATIENT PUBLISHED DATA

| ZIP Code | Age | Disease |
|---|---|---|
| 48677 | 26 | Gastric ulcer |
| 48602 | 28 | Stomach cancer |
| 48678 | 32 | Flu |
| 48685 | 36 | Flu |
| 48905 | 42 | Flu |
| 48906 | 46 | Flu |
| 48909 | 43 | Flu |
| 48673 | 48 | Heart Disease |
| 48607 | 55 | Heart Disease |
| 48655 | 58 | Stomach_cancer |

Modification of data is done in such a way that the resultant table has duplicated records there by restricting the disclosure. Indirectly there must be more than one link to the

external data base and is done by using generalization [1, 2, 3, 4]. Once the table is generalized various methods were used to check the property of duplication and distribution. To measure this Samarati and Sweeney [6,7] introduced k-anonymity. A table satisfies k-anonymity if every record in the table is indistinguishable from at least k − 1 other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table. In other words each group of quasi identifier values must have at least k-1 records and can be cheeked by linking a record in released data to multiple records publicly available data base. Table 3 shows a 2-anonymus generalization for table 1. Let us assume that the attacker uses the publicly available data base and finds that Rama's zip code is 48677 and his age is 26 and wants to know the disease of Rama, the attacker observes the anonymized table 3 from which attacker understands that 48677 & 26 has been generalized to 486** & [20-30] which can be linked to two records of published table and hence the disease cannot be inferred. In this table <486**,[40-50],Heart Disease> has been suppressed and is not considered for publication. Similarly if the attacker tries to infer Sita's disease who is related to group 3 but since the entire group contains the same sensitive attribute the attacker infers that his disease is Flu. This leakage of sensitive value leads to Attribute Level Disclosure. This happens if all the diseases indicated in a group are related to the same disease. To overcome this *l*-diversity [8] was defied. An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute. A table is said to have *l*-diversity if every equivalence class of the table has *l*-diversity. *l*-diversity also has the disadvantage that it suffers from skewness and similarity attack. To overcome this t-closeness was defined [9]. In this technique distribution of sensitive attribute must be equal to the anonymized block. This suffers from information loss

### A. Motivation

Major disclosures that take place are record level and attribute level to avoid this various anonymity techniques have been proposed in literature. Among them most important are k-anonymity, l-diversity, t-closeness, but each of them have several drawbacks as indicated above it includes data utility loss and sensitivity disclosure. To overcome this author in [5] had indicated a method personalized privacy preservation which takes in to account record owners privacy requirement. In [5] the record owner can indicate his privacy by indicating in terms of a guarding node. The values of it are based on a sensitive hierarchy which is framed by Data Publisher. The core of this technique is to divide the sensitive value based on importance so that more privacy is given to those values and data utility is improved. The drawback of this method is that it may require several iterations based on the guarding node, sensitive attribute is also generalized which has larger information loss. The most important drawback is that distribution of sensitive attribute has not been taken in to account while anonymization.

### B. Contribution and paper outline

In this paper we propose a novel privacy preserving technique that over comes the disadvantages of [5] and other anonymization techniques. The core of this method can be divided in to two major components. The first component deals with additional attribute used in the table which is in the form of flags. Sensitive Disclosure Flag (SDF) determines whether record owner sensitive information is to be disclosed or whether privacy should be maintained. The second flag that we are using is Sensitive Weigh (SW) which indicates how much sensitive the attribute is. SDF is dependent on SW.

TABLE 2. EXTERNAL VOTERS DATA BASE

| Name | ZIP Code | Age |
|---|---|---|
| Rama | 48677 | 26 |
| Laxman | 48677 | 35 |
| Suresh | 48602 | 28 |
| Nagesh | 48602 | 22 |
| Anuma | 48678 | 32 |
| Sita | 48905 | 42 |
| Kushal | 48909 | 43 |
| Vihan | 48906 | 46 |
| | . | |
| | . | |

TABLE 3. 2-ANONYMUS TABLE

| ZIP Code | Age | Disease |
|---|---|---|
| 486** | [20-30] | Gastric ulcer |
| 486** | [20-30] | Stomach cancer |
| 486** | [30-40] | Flu |
| 486** | [30-40] | Flu |
| 489** | [40-50] | Flu |
| 489** | [40-50] | Flu |
| 489** | [40-50] | Flu |
| 486** | [40-50] | Heart Disease |
| 486** | [50-60] | Heart Disease |
| 486** | [50-60] | Stomach_cancer |

SDF can be easily obtained from the individual when he/she is providing her data. SW can be based on the prior knowledge of sensitive attribute. General privacy methods provide the same level of security for all sensitive attributes which has been overcome in this method by the use of SDF and SW. The flag SDF=0 means that the record owner is not ready to disclose his sensitive attribute whereas SDF=1 doesn't mind revealing his sensitivity. SW is indicated by the publisher for those Sensitive attribute where privacy is at most important. For example record owner who has Flu or Gastritis doesn't mind revealing his identity as compared to a record owner who has Cancer. The value of SW=0 is used when the sensitive attribute is a common disease like Flu or Gastritis and SW=1 for sensitive attribute like Cancer which is not common. For SW=0 default value of SDF=1 & if SW=1 SDF values are accepted from record owner.

Second section deals with a novel representation called Frequency Distribution Block (FDB) and Quasi–Identifier Distribution Block (QIDB) used for measuring the distribution. FDB contains distribution of every disease with respect to original private data. For every record with SW=1 and SDF=0 QIDB is created. There will be multiple QIDB

blocks. These blocks are used to ensure that distribution of FDB is matched with individual QIDB.

In section II we have indicated Model and Notations used in our Personalized Privacy. Personalized Privacy Breach has been discussed in section III. Section IV gives the QIDB-Anonymization Algorithm. Experiment in section V has been analyzed. Related work has been discussed in section VI. Last section deals with conclusion and future work.

TABLE 4. SW FOR DISEASES

| Disease | SW |
|---|---|
| Gastric ulcer | 0 |
| Stomach cancer | 1 |
| Flu | 0 |
| Heart Disease | 1 |

TABLE 5. PATIENT PUBLISHED DATA WITH SW & SDF

| ZIP Code | Age | Disease | SW | SDF |
|---|---|---|---|---|
| 48677 | 26 | Gastric ulcer | 0 | 1 |
| 48602 | 28 | Stomach cancer | 1 | 0 |
| 48678 | 32 | Flu | 0 | 1 |
| 48685 | 36 | Flu | 0 | 1 |
| 48905 | 42 | Flu | 0 | 1 |
| 48906 | 46 | Flu | 0 | 1 |
| 48909 | 43 | Flu | 0 | 1 |
| 48673 | 48 | Heart Disease | 1 | 1 |
| 48607 | 55 | Heart Disease | 1 | 0 |
| 48655 | 58 | Stomach_cancer | 1 | 1 |

TABLE 6. FREQUENCY DISTRIBUTION BLOCK

| Disease | Probability |
|---|---|
| Gastric ulcer | 0.1 |
| Stomach cancer | 0.2 |
| Flu | 0.5 |
| Heart Disease | 0.2 |

## II. MODEL AND NOTATION FOR PERSONALIZED PRIVACY

Let $T$ be a relation containing private data about a set of individuals. there are four categories of attributes in $T$ i) unique Identifiers $UI_i$ which can be used for identification of a person and is removed from $T$ ii) quasi identifiers $Q_i$ whose values can be used for revealing the identity of a person by joining $Q_i$ with publicly available data iii) sensitive attributes $S_i$ which is confidential or sensitive to the record owner. iv)Non quasi identifiers $NQ_i$ which do not belong to the previous three categories.

Objective of our approach is to find a generalized table $T^*$ such that distribution of each *QIDB* is approximately equal to the diversity of the overall distribution which is there in *FDB*. For simplicity the entire quasi identifiers are represented as Q and their values as q. similarly we assume there is a single sensitive attribute $S$ and its value is $s$. Relation $T$ is made of $n$ number of tuples $T=\{t_1,t_2,\ldots,t_n\}$. Record owner information can be retrieved by referring as $t_i.s$ to indicate sensitive value and $t_i.q$ for quasi identifier value $1 \leq i \leq n$.

### A. Requirement for personal privacy

DEFINITION 1 (SENSITIVE WEIGHT) For each tuple $t \in T$, its *sensitive weight* is added. This value is taken from Relation $W(d,sw)$ where $d$ disease and $sw$ sensitive weight. $W$ contains $k$ records.

$$t_i.sw=\{ w_j.sw \text{ if } w_j.d=t_i.s \; 1 \leq j \leq k \} \; \forall \; 1 \leq i \leq n$$

For example table 4 shows the sw value for each disease. This distribution is taken from Table 1.

DEFINTION 2 (SENSITIVE DISCLOSURE FLAG) for each tuple $t \in T$, its *sensitive Disclosure Flag* is indicated as *t.sdf*.

$$t_i.sdf=\{1 \quad \text{if } t_i.sw=0$$

$$ud \quad t_i.sw=1 \; \}\forall \; 1 \leq i \leq n$$

*ud* represents *user defined* and the value is either *0* or *1*. $t_i.sdf=0$ then user is not ready to disclose his information and $t_i.sdf=1$ then user is ready to disclose his information. In table 5 value of sw and sdf are indicated assuming that sdf value is accepted from record owner for SW=1. We can also observe that if sw=0 its correspondent sdf is initialized to 1 indicating that the sensitivity of this record is not of much relevance.

### B. Thresholds for Personalized Privacy

Threshold values are defined for various dimensions of personalized privacy to improve the overall performance of generalization, suppression and disclosure.

i)$TH\rho_n$ minimum number of records in T.

ii)$TH\rho_{iter}$ maximum number of iterations that must be performed . it indicates the amount of generalization & Height(VDH)

iii)$TH\rho_{suppr}$ minimum number of sensitive values for suppression.

iv)$TH\rho_{disc}$ minimum number of sensitive values for disclosure.

v) $TH\rho_{acct}$ minimum threshold that can be added or subtracted.

Since we are considering the distribution aspect we can indicate different threshold values. The first value indicates the minimum number of tuples that must be present for applying anonymization which was never defined in the previous representations. $TH\rho_{iter}$ based on the knowledge of the height of Value domain hierarchy. The larger the value of $TH\rho_{iter}$ higher the generalization and consequently information loss is more. $TH\rho_{suppr}$ indicates the minimum number of sensitive distribution that may be there in QIDB for removal of that block after $TH\rho_{iter}$. $TH\rho_{disc}$ indicates the threshold value that can be added or subtracted to each frequency distribution for each disease such that it is equivalent to the distribution FDB. The frequency of QIDB block and FDB will not be exactly same so while checking the distribution of each disease is checked whether the frequency in that $qidb.v.s \pm TH\rho_{acct}$ always $TH\rho_{disc} > TH\rho_{acct}$.

### C. Additional Block Creations for personal privacy

DEFINITION 3 (FREQUENCY DISTRIBUTION BLOCK) Distribution of each $w_j.d$ with respect to the original distribution $t_i.s$ is saved in relation *FDB(d,p)* where *d* indicates disease and *p* indicates probability distribution of it. Each *p* for *d* is calculated by mapping each *d* in *T* (values of $t_i.s=fdb_u.d$) to the total number of tuples in *T* i.e. *n*, $\forall\ 1 \le u \le k$ . let us assume there are *m* records in the relation.

DEFINITION 4 (Quasi–Identifier Distribution Block ) for each $t_i.s$ where $t_i.sw=1$ & $t_i.sdf=0$ a new *QIDB* is created containing $t_i.s\ \forall\ 1 \le i \le n$. The relation *QIDB.V(q,s)* where $qidb.v_l.q=t_i.q$ & $qidb.v_l.s=t_i.s$. Let us assume there are *dn QIDB* blocks.

For example Table 6 shows the frequency distribution of each disease. This distribution shows that the disease flu is a common disease so its frequency is more, around 50% in the published data. The same distribution is maintained in each of the QIDB. In the first iteration two blocks of QIDB will be created for the qasi value <48602,28> and <48607,55> since its SW=1 & SDF=0 which is shown in table 7 & 8.

TABLE 7. QIDB.1 CONTENTS

| ZIP Code | Age | Disease |
|---|---|---|
| 48602 | 28 | Stomach cancer |

TABLE 8. QIDB.2 CONTENTS

| ZIP Code | Age | Disease |
|---|---|---|
| 48607 | 55 | Heart Disease |

### D. Functions For Personal Privacy

DEFINITION 5 (GENERALIZATION) A general domain of an attribute *T.Q* is given by a generalization function. Given a value *t.q* in the original domain, function returns a generalized value within the domain.

For each $t \in T$ we use $t^*$ to represent its generalized tuple in $T^*$.we denote it as *G(T)*

This is similar to earlier representations let us assume that Domain Generalization Hierarchy and Value Generalization Hierarchy are defined for each Quasi Identifiers. The distance vector of quasi attributes has also been generated. In figure 1 Value and Domain Generalization Hierarchy of zipcode has been indicated. Age is also generalized similarly. Distance vector is calculated which is shown in figure 2.

DEFINITION 6 (CHECK FREQUENCY) for any *QIDB*, we check *CF(QIDB.V )* wither *QIDB.V* frequency of distribution is equal to the frequency distribution in *FDB.* It is done as follows

Let c be the no of records in *QIDB.V*. for each *UNIQ(qidb.$v_l$.s)* find total no of mappings which match *qidb.$v_l$.s* to the no of records i.e. *c* in *QIDB.V*, thus CF will return true if

$\forall\ 1 \le u \le m$ such that $fdb_u.d=qidb.v_l.s$

$$fdb_u.p \approx \frac{UNIQ(qidb.v_l.s)}{c} \pm TH\rho_{acct}$$

this is checked in every iteration if a QIDB satisfies the frequency distribution then this block will not be considered for the next iteration.

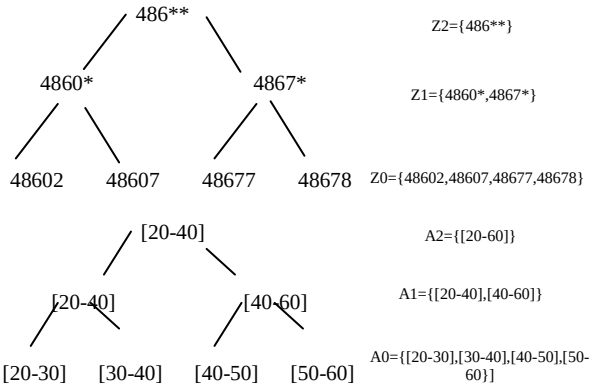



Figure 1. An example of Value and Domain generalization hierarchy for zipcode and Age

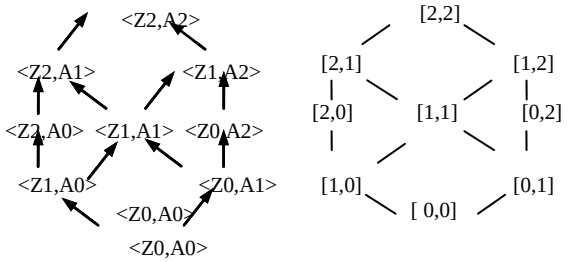



Figure 2. Hierarchy DGH<Z0,A0> and corresponding Hierarchy of distance vectors

DEFINITION 7 (SUPPRESSION) After $TH\rho_{iter}$ *iterations, SUPP(QIDB.v)* suppress the block if it satisfies the following condition

$\forall\ 1 \le u \le m$ such that for every $fdb_u.d=qidb.v_l.s \wedge fdb_u.d=w_j.d \wedge w_j.sw=1\ \forall$ j $1 \le j \le k$

$$Count(qidb.v_l.s) \le TH\rho_{suppr}$$

DEFINITION 8 (DISCLOSURE) After $TH\rho_{iter}$ *iterations, DIS(QIDB.v)* adds additional records if it satisfies the following condition

$\forall\ 1 \le l \le c$ such that for every $fdb_u.d=qidb.v_l.s \wedge fdb_u.d=w_j.d \wedge w_j.sw=1$ for some j $1 \le j \le k$

$$\frac{UNIQ(qidb.v_l.s)}{c} \approx TH\rho_{disc} \pm fdb_u.p$$

## III. PERSONALIZED PRIVACY BREACH

Consider an attacker who attempts to infer the sensitive data of a record owner *x*. the worst case scenario assumes that the adversary knows *Q* of *X*, therefore the attacker observes only those tuples $t^* \in T^*$ whose *Q* value $t_i^*.q$ covers *x.q* for all *i* such that $1 \le i \le n$. These tuples form a *Q-group*. That is, if $t_i^*$ and $t_{ip}^*$ are two such tuples then $t_i^*.q=t_{ip}^*.q$ for all *i* such that $1 \le i \le n$.if this group is not formed the attacker cannot infer sensitive attribute of *x*.

DEFINITION 9 (REQUIRED Q-GROUP/ ACT(X)). Given an individual *x*, the Required Q-group *RG(X)* is the only Q-group in *t** covers *x.q*. let us assume *ACT(X)* refers to those records which are generalized to *RG(X)*.

*ACT(X)* is unknown to the attacker. To obtain *ACT(X)*, the attacker must find some external data base *EXT(X)* that must be covered in *RG(X)*.

DEFINITION 10(EXTERNAL DATA BASE EXT(X)) *EXT(X)* are set of individuals whose value is covered by *RG(X)*

In general ACT(X) $\subseteq$ *EXT(X)*

The attacker adopts a combinational approach to infer sensitive attribute of *x.* let us assume that *x.s* is present in one of $t_i^*$ and the repetition of *x* is not present. The possible reconstruction of the *RG(X)* includes

*r* distinct record owners $x_1,x_2,x_3,\ldots,x_r$ who belong to *EXT(X)* are taken but there can be only *y* in *RG(X).* this can be understood by the probabilistic nature and can be indicated as perm(r,y). perm(r,y) is Possible Reconstruction(PR) that can be formed by using r owners and y mappings. Breach Probability (BP) indicates the probability of inferred knowledge. Let us assume *ACTN* indicates actual number of records with sensitive attribute that can be inferred to x.

$$BP=\frac{ACTN}{perm(r,y)}$$

*BP* will decide the privacy parameter, *BP* is 100% then x can be inferred if it is very low than the inference will be very much difficult for the attacker.

## IV. *QIDB*-ANONYMIZATION ALGORITHM

In this algorithm we are using a sequential processing of quasi values since the assumption is that in each region usually the distribution of sensitivity is approximately same. The algorithm is as follows

Algorithm *QIDB*-Anonymization

Input: private data *T* with *SW-SDF*, threshold values $THp_n, THp_{iter}, THp_{suppr}, THp_{disc}, THp_{acct}$ and initialized *FDB(d,p)*

Output: publishable table $T^*$

1. if ($n < THp_n$) then return with *1*

2. for every $t_i.s$ where $t_i.sw=1$ & $t_i.sdf=0$ a new *QIDB* is created containing $t_i.s$ *and* $t_i.q$ $\forall\ 1 \le i \le n$.

3.ini_itr=0, accept_flag=0 and gen=first G(T)

4.while (ini_itr< $THp_{iter}$and accept_flag=0)

4.1.QIDB blocks are removed if CF() returns *true* then check the number of QIDB if it is equal to zero then accept_flag=1

4.2.itr=itr+1 and gen=next G(T)

5. if accept_flag=0 then invoke supp() & dis()

6. check number of QIDB if it is equal to zero accept_flag=1

7. publish T* if accept_flag=1

The resultant anonymization after applying Personal Anonymization of one of the QIDB with $THp_{acct}$ =0.1 block is shown in Table 9.

TABLE 9. RESULTANT SW-SDF BASED QIDB-ANONYMIZATION WITH $THp_{ACCT}$ =0.1

| ZIP Code | Age | Disease |
|---|---|---|
| 486** | [20-40] | Stomach cancer |
| 486** | [20-40] | Gastric ulcer |
| 486** | [20-40] | Heart Disease |
| 486** | [20-40] | Flu |
| 486** | [20-40] | Flu |

## V. EXPERIMENTS

In this section we try to evaluate the effectiveness of our technique as compared to k-anonymity and l-diversity. We have used a standard dataset used in the literature[7,8,9] for our experiment. We have considered Americal adult dataset of 400 records, with the following quasi attributes Age, Education , Maritial status & Occupation. The attribute age is numerical and the rest of the attributes are categorical. The sensitive attribute income has been converted to disease. Probability is used to find SDF value for SW=1.

We have defined and used generalization hierarchy for each qasi identifier and distance vector is generated which has been used in our algorithm. The maximum height of our generalization hierarchy is 10. Information loss parameter is shown in figure 3. Less the information loss better is the data quality. Minimal distortion (MD)is based on charging penalty for each value which is generalized or suppressed. Each hierarchy is assigned a penalty when it is generalized to the next level with in the domain generalization hierarchy. MD is shown in figure 4. In our experiment we have used a penalty of 10 for every generalization. This Discernibility Metric (DM) calculates the cost by charging a penalty to each tuple for being indistinguishable from other tuples which is shown in figure 5. Execution time is shown in figure 6. For our experiment the threshold values $THp_n$=400, $THp_{iter}$=10, $THp_{suppr}$=1, $THp_{disc}$=0.01 and $THp_{acct}$=0.1 was used. Experiment was conducted using Matlab 7 in which our algorithm out performs k-anonymity and l-diversity.

## VI. RELATED WORK

Different methods of PPDM exist, among them the most important are Randomization Method [13], Data Swapping [14], Cryptographic Approach [15] and Data Anonymization. Data Anonymization is considered as one of the most important anonymization technique since it has lesser information loss and higher data utility. There are different anonymization algorithms has been proposed in literature [1, 3, 4, 6, 10, 11, 12]. Initial anonymization algorithm was called k-anonymity [6] but the drawback of this approach is that it is prone to record level disclosure. To overcome this

disadvantage *l*-diversity[8] was proposed. Disadvantage is that it is prone to Skewness and Back ground Knowledge Attack. *t*-closeness[9] is used to overcome the disadvantages of l-diversity but it has larger information loss. Personalized Privacy[5] was added on to anonymization which gave lesser information loss. This is the motivation of our approach.

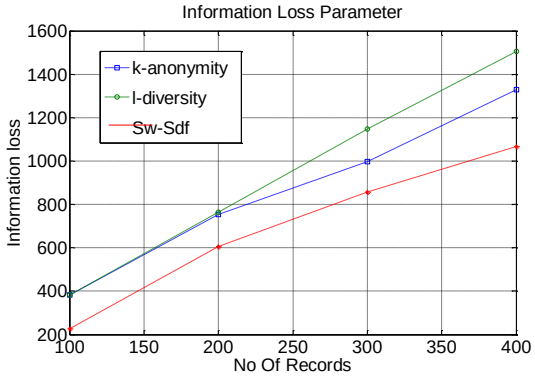



Figure 3. Information Loss Of *SW-SDF* Personal Anonymization As Compared With K-Anonymity & L-Diversiy

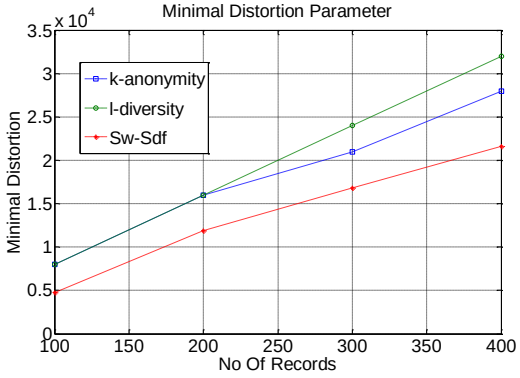



Figure 4. Minimal Distortion Parameter Of SW-SDF Personal Anonymization As Compared With K-Anonymity & L-Diversiy

## VII. Conclusions and FutureWork

Personalized privacy is an important research direction in PPDP since its data quality and execution time is less. Usage of *SW* not only improves the indication of sensitivity as the entire records do not require privacy but also improves the data utility. *SDF* is an additional flag which once again improves data utility with in *SW* record since some of the record owners are ready to reveal their identity. Thus the combination of *SW-SDF* is a better option for personalized privacy as compared to just using a guarding node.

*QIDB* based anonymization allows different quasi group to be generalized independently. In this approach each quidb block is checked for the frequency distribution of sensitive value approximately equal to the frequency distribution of the sensitive value in original contents thereby improving privacy.

It also overcomes record linkage, attribute linkage and even probabilistic attack. This approach works well when the frequency distribution of a particular sensitivity is concentrated within a region of individual pattern.

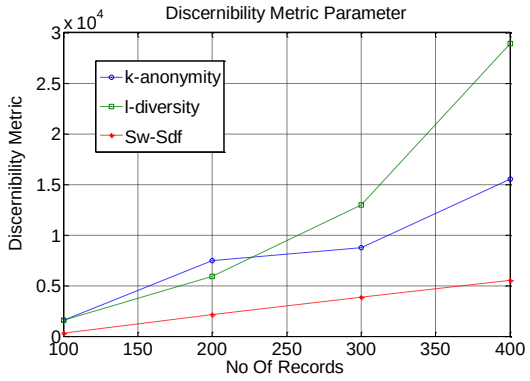



Figure 5. Discernibility Metric Parameter of SW-SDF personal anonymization as compared with k-anonymity & l-diversiy

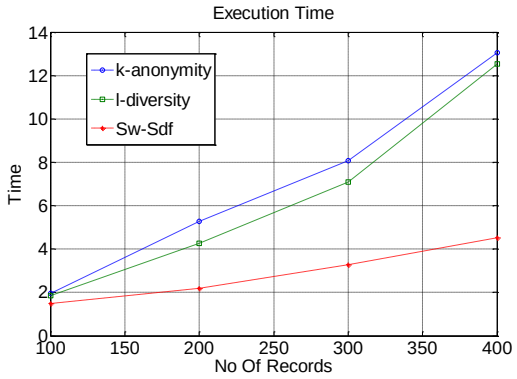



Figure 6. Execution Time of SW-SDF personal anonymization as compared with k-anonymity & l-diversiy

There are several future research directions along the way of analyzing *SW-SDF* personal privacy with *QIDB* anonymization. First we haven't considered the effect of Sequential Release and Multiple Release of published data. Research on giving different Weight on sensitivity can be considered. In this approach we have used sequential processing of records to check the generalized record matches with *QIDB* generalized value if they are same then it would be included in the block. Instead of sequential processing alternative methods can be looked in to. This method can be extended to unstructured schema and multi-dimensional data.

## References

[1] Lefevre K., Dewitt D. J. and Ramakrishnan R., "Incognito: Efficient full-domain k-anonymity". In Proceedings of ACM SIGMOD. ACM, New York, pp. 49–60, 2005.

[2] Fung B. C. M., Wang K. and Yu P. S., "Anonymizing classification data for privacy preservation". IEEE Trans. Knowl. Data Engin, pp. 711–725,2007.

[3] Fung B. C. M.,Wang K. and Yu P. S., " Top-down specialization for information and privacy preservation". In Proceedings of the 21st IEEE International Conference on Data Engineering (ICDE). 205–216, 2005.

[4] Iyengar V. S., "Transforming data to satisfy privacy constraints". In Proceedings of the 8th ACMSIGKDD. ACM, New York, pp. 279–288,2002.

[5] Xiao X. and Tao Y., "Personalized privacy preservation". In Proceedings of the ACM SIGMOD Conference.ACM, New York, 2006.

[6] P. Samarati, "Protecting Respondent's Privacy in Microdata Release". IEEE Trans. on Knowledge and Data Engineering (TKDE), vol. 13, no.6, pp. 1010-1027, 2001.

[7] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy". Int'l J.Uncertain. Fuzz., vol. 10, no. 5, pp. 557-570, 2002.

[8] Machanavajjhala A, Gehrke J, Kifer D and Venkitasubramaniam M, "l-diversity: Privacy beyond k-anonymity". In Proceedings of the 22nd IEEE International Conference on Data Engineering(ICDE), 2006.

[9] Ninghui Li , Tiancheng Li , Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l–Diversity". ICDE Conference, 2007.

[10] R. Bayardo and R. Agrawal. "Data privacy through optimal k-anonymization". In ICDE, pp. 217–228, 2005.

[11] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian multidimensional k-anonymity". In ICDE, 2006.

[12] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection". In *ICDM*, pp. 249–256,2004.

[13] Agrawal D. Aggarwal C. C., "On the Design and Quantification of Privacy- Preserving Data Mining Algorithms". ACM PODS Conference, 2002.

[14] Fienberg S.,McIntyre J., "Data Swapping: Variations on a Theme by Dalenius and Reiss". Technical Report, National Institute of Statistical Sciences, 2003.

[15] Pinkas B.,"Cryptographic Techniques for Privacy-Preserving Data Mining". ACM SIGKDD Explorations, 2002.

# Integration of data mining within a Strategic Knowledge Management framework:

## A platform for competitive advantage in the Australian mining sector

Sanaz Moayer
Murdoch University
Perth, Australia

Scott Gardner
Murdoch University
Perth, Australia

**Abstract— In today's globally interconnected economy, knowledge is recognised as a valuable intangible asset and source of competitive advantage for firms operating in both established and emerging industries. Within these contexts Knowledge Management (KM) manifests as set of organising principles and heuristics which shape management routines, structures, technologies and cultures within organisations.When employed as an integral part of business strategy KM can blend and develop the expertise and capacity embedded in human and technological networks. This may improve processes or add value to products, services, brands and reputation. We argue that if located within a suitable strategic framework, KM can enable sustainable competitive advantage by mobilising the intangible value in networks to create products, processes or services with unique characteristics that are hard to substitute or replicate. Despite the promise of integrated knowledge strategies within high technology and professional service industries, there has been limited discussion of business strategies linked to Knowledge Management in traditional capital intensive industries such as mining and petroleum. Within these industries IT-centric Knowledge Management Systems (KMS) have dominated, with varying degrees of success as business analysis, process improvement and cost reduction tools.**

**This paper aims to explore the opportunities and benefits arising from the application of a strategic KM and Data Mining framework within the local operations of large domestic or multinational mining companies, located in Western Australia (WA). The paper presents a high level conceptual framework for integrating so called *hard*, ICT and *soft*, human systems representing the explicit and tacit knowledge embedded within broader networks of mining activity. This Strategic Knowledge Management (SKM) framework is presented as a novel first step towards improving organisational performance and realisation of the human and technological capability captured in organisational networks. The SKM framework represents a unique combination of concepts and constructs from the Strategy, Knowledge Management, Information Systems, and Data Mining literatures. It was generated from the Stage 1- Literature and industry documentation review of a two stage exploratory study. Stage 2 will comprise a quantitative case based research approach employing clearly defined metrics to describe and compare SKM activity in designated mining companies.**

***Keywords- Knowledge Management (KM); data mining, sustainable competitive advantage; Strategic Knowledge Management (SKM) framework; integration; hard and soft systems; Australian mining organisation.***

## I. INTRODUCTION:LINKING BUSINESS STRATEGY AND KNOWLEDGE MANAGEMENT

In a complex and challenging world, organizational success depends on the ability to configure corporate assets to reflect rapidly shifting markets and environmental conditions (Hafizi & Nor Hayati, 2006). In recent years the competitive significance of tangible assets such as financial capital, technology or inventories has declined in the disrupted, globally interconnected markets of early 21st century. This has led to an increased focus on strategic deployment of unique, hard to imitate, intangible assets such as expert knowledge which as a basis for sustainable competitive advantage (Hafizi & Nor Hayati, 2006). (Shih, Chang, & Lin, 2010).

Knowledge is actionable information which helps people to make better decisions and to be more creative in their approach to a range of problem solving activities. Many organizations seek to identify, stocks of knowledge embedded in their human, information and communication networks and link them to value adding flows, using Knowledge Management. Knowledge Management(KM) is defined by Jashapara (2011, p14) as- "The effective knowledge processes associated with exploration, exploitation and sharing of human knowledge (tacit and explicit) that use appropriate technology and cultural environments to enhance an organization's intellectual capital and performance". Knowledge Management (KM) processes help organizations to define, select, organize, distribute, and transfer information, knowledge and expertise retained in the organization's memory in an unstructured manner (Turban & Leidner, 2008).Effective KM improves operational efficiency, enhances products and services and creates customer satisfaction (Lee, 2009). Knowledge Management has many potential benefits such as retaining expertise, capturing and sharing best practice, corporate support, improved customer service, better decision making, increased profitability and competitive advantage (Duvall, 2002). Over the past fifteen years KM has been increasingly recognised in the strategy, management and information systems literature, with senior managers treating it as a focal point for improving organizational performance, adding value to goods and services, building brands and reputation.

While KM is often associated with professional services, biotechnology, IT and internet businesses, it also has significant potential for adding value and reducing costs in mining and allied industries. Mining organizations employ many skilled knowledge professionals in science, engineering, and technology, including geologists and other earth scientists. Computer scientists also have a key role in mining organizations working with managers from different functional areas to integrate and exploit the knowledge capacity of these companies (Fernandez, 2010). Employing high value knowledge workers and retaining knowledge in-house has a strong economic justification. The migration towards Strategic Knowledge Management systems and practices is inevitable part of this process (Fernandez, 2010).

## II. Converting Data Into Actionable Knowledge

In the IT literature Data Mining (DM) is often represented uncritically as a powerful tool for knowledge discovery. As noted by (Lee, 2009) it is helpful for discovering patterns of data and creating new information. Arguably Data Mining gives organisations the ability to exceed their goals and enables decision makers to deploy the results (Noonna, 2000), for improving business performance in highly competitive environments. These statements point to the practical benefits and applications of data mining as a knowledge generation and decision support tool. However there has been little discussion in the literature of how DM processes can be integrated into dynamic business strategies supported by KM organizing principles, management routines, structures, information systems and organisational culture. Malhotra highlights the dynamic connectivity between strategy, change, human and information systems stating that- "Knowledge Management caters to the critical issues of organisational adaptation, survival, and competence in the face of increasingly discontinuous environmental change…(It) seeks a synergistic combination of the data and information processing capabilities of information technology and the creative and innovative capacity of human beings" (Malhotra 2000&2001 in Haslinda and Sarinah 2009).

Just how to achieve these synergies between hard (technological) and soft (human) systems remains one of the central questions yet to be addressed in organisational studies. Managing the interface between people and technology within complex organisations often manifests as a 'black art' for even the most adept computer scientist or organisational psychologist. The ability to frame or ask the right questions and arrive at satisfactory answers is deeply rooted in our own worldviews. It is subject to our interpretation of KM concepts drawn from philosophy and the information, cognitive, social sciences. As observed by Liebowitz and Megbolugbe (2003) converting data into actionable information or useful knowledge in organisations is not a simple, mechanistic process. It is dependent on heuristic rules and the broader ontologies or worldviews of key actors within the system (Liebowitz & Megbolugbe, 2003, p. 189).The Strategic Knowledge Management (SKM) framework presented in this paper (Figure 1) acknowledges this complexity while outlining key elements and broad interrelationships, which subject to further empirical investigation may advance KM and DM practice.

## III. Objectives, Scope And Method Of The Paper

The aim of this paper is to present a novel Strategic Knowledge Management (SKM) framework (Figure1), as a platform for building competitive advantage within a mining industry context. The defining characteristics and claimed benefits of Knowledge Management and Data Mining are discussed in relation to the SKM framework incorporating four related perspectives on strategic management. These characteristics have been derived from a limited review of the Strategic Management, Knowledge Management and Data Mining literatures, with relevant academic sources and reports profiling the Australian mining industry. This review of concepts and industry data has been used to produce a high level representation of strategic management and Data Mining processes applicable to the Australian based operations of large mining companies. It represents Stage 1 of two stage study combining a a high level SKM framework generated from a review of relevant strategy, KM and DM concepts with a quantitative case based research method. The SKM framework in this paper is presented for peer and industry stakeholder feedback prior to developing a detailed Stage 2 model employing hard metrics. These will be used to describe measure and compare Strategic Knowledge Management (SKM) activities displayed in three large mining organisations.

## IV. Four Views Of Strategy And Competitive Advantage

Strategy is the scope for an organisation to gain benefits and advantages, with available resources in a challenging environment in the long-term (Johnson, Scholes, & Whittington, 2005). Strategic management is the art and science of formulating, implementing and evaluating functional decisions that empower the organisation to gain its goals and objectives (David, 2011). The long term performance of organisation is defined by a set of managerial decisions or strategic choices (Hunger & Wheelen, 2003).

There a number of interrelated views of strategy identified as the basis of competitive advantage in the strategic management literature. These include: variations on Porter's (1980) economic perspective or Market-Based View (MBV); Freeman and McVea's (2001) political perspective or Stakeholder-Based View, (SBV); Barney's (1993) internal human, structural and capital asset capability perspective, or Resource-Based View (RBV); and more recently the portfolio of expertise or Knowledge-Based View (KBV) popularised by Spender (1996),Grant (1996). Both RBV and KBV have subsequently been elaborated and linked the theory of dynamic capabilities (Mouritsen, Larsen, & Bukh, 2005).Dynamic capabilities theory focuses on the firm's ability anticipate and adapt to dynamic, discontinuous or disruptive market conditions. This is achieved when managers synergistically combine portfolios of knowledge assets with organizational learning routines, sense making and strategy to process market signals and anticipate emerging conditions (Choo, 1998). These actions serve to renew organizational structures and systems whilst generating unique assets as a basis for competitive advantage. This follows the Schumpeterian logic of competition based on "creative destruction of existing resources and novel combinations of

new functional competences" (Pavlou and El Sawy, 2004 in Easterby-Smith and Prieto, 2008, p 236).

All four strategic perspectives and elements of dynamic capabilities theory incorporated into the SKM (Figure 1), are particularly applicable to multinational mining companies, which can exercise control over local, regional, and global markets and supply chains. The local operations of these larger companies can also develop a strong internal human resource profile and workforce capability through attractive salary packages or significant investments in workforce planning, training, information technology expertise and infrastructure. The Market-Based View focuses on achieving an attractive position within a designated industry. (Poser, 2003). It can help mining companies to exercise strategic choice and identify which factors of production should be prioritised to gain competitive advantage in specific industry structures or market segments (Porter, 1980). The Stakeholder-Based View acknowledges the political dimensions of strategy. It highlights the importance of working with constituents or stakeholders to facilitate the achievement of business goals and competitive advantage through informed decision making. It proposes that managers formulate and implement political processes that identify, classify and build productive relationships with people who have stake in the business (Freeman & McVea, 2001); (Gardner, 2001). The Resource-Based View emphasises that organisational performance depends on internal resource configurations and capabilities including physical resources, human resources, and organisational resources (David, 2011). By extension the Knowledge-Based View is built on the logic of the Resource-Based View. It revisits many tenets of knowledge conversion and creation. It also identifies organizational learning and management routines as potential sources of competitive advantage (Jashapara, 2011). Taking the Knowledge-Based View (KBV) of strategy, knowledge is a valuable resource and basis for competitive advantage in organizations. We argue that this view is particularly applicable to the emerging knowledge and high technology sectors and the traditionally capital intensive industry sectors, such as mining and petroleum.

## V. The Australian Mining Industry

Mining is a major industry in Australia. One third of the word's mineral resources are produced in Australia (Nimmagadda & Dreher, 2009). The Australian mining sector generated revenue of about $138.8 billion in 2006-7 growing to $203.9 billion in 2011-12. Mining is expected to generate about 8.0% of Australia's GDP in 2012 with a profit forecast of $58.3 billion. (IBIS World, 2012). These statistics indicate that the mining industry plays important role in maintaining revenue growth in Australia within a global context of economic slowdown or contraction. No other industry in Australia has gained a superior significance in economic development terms (Fernandez, 2010).

The scope of the mining industry includes all operations for extracting minerals or hydrocarbons. Coal, oil and gas, metal ore, and non-metallic mineral commodities are products of this industry (IBIS World, 2012). Exploration, drilling, production, and marketing are significant business functions in the resource industry. In recent years the major resource companies and primary contractors have increasingly recognised relationships with suppliers, customers, regulators and other stakeholders as critical determinant of firm and industry performance. According to Richards (2009), suppliers in particular can drive organisations to produce new services in different ways (Richards, 2009). The knowledge-base of suppliers is an important element to increase performance and maintain the competitive advantage of firms in the mining industry (Urzúa, 2011).

## VI. Data mining and its application in mining industries

Data Mining (DM) is a technique for identifying patterns and relationships between data in large databases (Lee, 2009). It also informs strategic and operational decisions in organisations through dashboards, and interrogation or scanning of relational databases. Data mining aids organisational problem solving by employing programs that can search for patterns and relationships without human intervention (Paddock & Lemoine, 2012, p. 4). Giudici (2003, p.2) offers a more complete definition of data mining as: " ...the process of selection, exploration, and modelling of large quantities of data to discover regularities or relations, that are at first unknown with the aim of obtaining clear and useful results for the owner of the database".

Data Mining encompasses major tasks such as data exploration, data archaeology, data pattern processing, data dredging, information harvesting, and knowledge extraction (Lee, 2009). Data mining technology is becoming a significant aspect of strategy for many organisations. It has become major component of (often complex multi-interface) enterprise decision support systems (Brusilovsky & Brusilovskiy, 2008).

In general, business problems can be categorised as structured or unstructured. Statistical analysis is useful for overcoming structured problems and DM is often employed to deal with unstructured problems. This capability to interpret problem characteristics and dimensions makes DM potentially compatible with the human cognitive processes required to generate useful context specific, knowledge and address complex problems. This is consistent with the logic of gaining competitive advantage through unique processes, products and services that are hard to replicate. As noted by (Brusilovsky & Brusilovskiy, 2008) the strategic strength of DM resides in the ability to deal with unstructured problems because competitors are not familiar with the characteristics of, or solutions to, these kinds of problems (Brusilovsky & Brusilovskiy, 2008, p. 131).Data Mining clearly has the potential to deliver significant benefits if framed within broader KM enabling architecture and aligned through heuristic questioning and iteration to business goals and an unfolding strategy process. The different characteristics of KM enabling architecture are outlined in the models below. These key elements can most usefully be incorporated into the SKM framework are then briefly discussed.

## VII. Relevant Models Of Knowledge Management

Senior managers, KM and IT specialists within the mining industry must choose an appropriate model which fits the

strategic goals, processes and changing environment of their organisations. Table 1 below illustrates the different dimensions of knowledge, creation; individual cognition and shared learning captured in some of the more widely cited models from the KM literature: (Dalkir, 2005, pp. 49-72); (Haslinda & Sarinah, 2009, pp. 189-196); (McAdam & McCreedy, 1999, pp. 95-98).

The decision to incorporate any of these KM elements into an organisational Knowledge Management systems and practices is context or domain dependent. According to Sanchez and Heene (1997) organisational knowledge and learning cannot be understood from a purely (top down) strategic perspective, so organisations should also generate (bottom up) KM activities based on analysis of the context in which organisation's knowledge is applied (Sanchez & Heene, 1997, p. 12). All the models outlined above contribute a perspective or position on the nature of knowledge, knowledge as an asset, and knowledge as a capability, knowledge enabling structures, cultures, or leadership practices germane to the SKM framework in Figure 1. However Nonaka's (1995) knowledge spiral model and Hedland and Nonaka's (1993) KM framework are the most pertinent to our discussion of integrating strategy, KM structures and processes, with data mining activities. These models focus on surfacing, combining and actioning tacit knowledge (based on human cognition) and explicit knowledge (repositories of data and information) to add value in organisations. This is achieved through the SECI (Socialisation, Externalisation, Combination and Internalisation), knowledge conversion process. The SECI process is in turn enabled by Ba - Nonaka and Tackeuchi's concept of a safe space (or cyberspace). This supports conversion of knowledge assets into value added products, processes or services, enabled by Information and Communication Technology (ICT) infrastructure, management and teamwork practices (Nonaka & Takeuchi, 1995). The converted knowledge assets are simultaneously carried up through the organisational structure in a dynamic spiral to inform senior management decision making and support the strategy process. This model is most applicable within project based industries like mining which typically rely on matrixes superimposed on functional structures to align staff expertise and capacity with business requirements. We propose that SKM can overcome these limitations by using clearly articulated organising principles to drive KM and Organisational Learning (OL) activities. These are embedded in management routines, which continuously align human, and technology interactions, structures, cultural norms and values with dynamic changes in the competitive environment.

## VIII. Towards a Strategic Knowledge Management (SKM) framework for the Australian Mining Industry

Figure 1, below, illustrates the integration of data mining practices into a SKM framework applicable to global and Australian based mining operations. It presents a high level guide to exploiting the knowledge embedded in human and ICT networks to create process efficiencies, improve decision making and by extension, productivity for large multinational miners domiciled in Australia. SKM is premised on the idea that data mining should not be conducted in isolation from a broader KM strategy that incorporates the following elements:

1)Simultaneous application of interrelated strategic perspectives notably: The Market Based View paying attention to product, price or supply chain concerns; The Resource-Based View focusing on how to build human capability and physical asset capacity; The Stakeholder Based View concerned with building relational capital and the salient stakeholders who can affect or are affected by the goals and activities of the firm; and finally the Knowledge-Based View which emphasises the importance of managing human networks, knowledge portfolios, stocks and flows as a key determinant of organisational performance and sustainable competitive advantage. Adoption of these strategic perspectives and organisational learning processes links the firm's dynamic capabilities to KM and day to day management practices. As such SKM supports a dynamic strategy process which calibrates internal capability with changes in the external environment.

2) A knowledge enabling architecture based on reciprocal hard and soft system organising principles. This living architecture is generated from different ontological positions, heuristics, and taxonomies. Its design elements are comprised of transparent organising principles, shared goals and priorities negotiated between key internal actors representing the tension between commercial, humanistic and technologically orientated worldviews. These in turn drive management routines and practices, information and communications infrastructure design (including DM tools), organisational structure, culture and reward systems. Using iterative action learning loops these first and second order design elements are continuously re-configured to support knowledge creation and adaptability to dynamic competitive conditions.

3) Alignment mechanisms: The proposed SKM approach to action learning is consistent with Nonaka's knowledge spiral and tacit to explicit continuous knowledge conversion model. Both are generative bottom up approaches. They are aim to align individual and team behaviours with organisational structures, rewards and management routines and broader KM based strategy. This requires ongoing dialogue between DM and information systems specialists, vendors and senior managers. This is something akin to the *Ba* or a safe physical or virtual space for knowledge sharing. High trust, open protocol environments of this type is essential for effective knowledge sharing, problem solving, and identification of common ground between senior managers and IT specialists. This safe space and common ground allows for the surfacing and testing of different ontological viewpoints and creation of shared heuristics. This precedes the dialogue on shared organising principles which once agreed can lead practical discussion of how best to manage data, information and unique knowledge assets to achieve competitive advantage for the organisation.This preliminary framework will be further refined based on peer review, tested and empirically validated through application to KM and data mining systems and practices in three West Australian mining organisations.

A broad inventory of data mining tasks and Strategic Knowledge Management processes will be created for each firm as part of the process for testing the model.

| *Model* | *Features* |
|---|---|
| The von Krogh and Roos Model of organisational Epistemology (Von Krogh & Roos, 1995) | Individual knowledge |
| | Social knowledge |
| The Nonaka and Takeuchi Knowledge Spiral Model (Nonaka & Takeuchi, 1995) | Knowledge creation |
| | Knowledge conversion (Socialisation, externalisation, combination, internalisation), '*Ba'* safe space, Knowledge assets. |
| Hedlund and Nonaka's Knowledge Management Model (Hedlund & Nonaka, 1993) | Articulated knowledge- Individual |
| | Tacit knowledge- Individual |
| | Articulated knowledge- Group |
| | Tacit knowledge- Group |
| | Articulated knowledge- Organisation |
| | Tacit knowledge- Organisation |
| | Articulated knowledge- Inter- Organisational Domain |
| | Tacit knowledge- Inter- Organisational Domain |
| The Choo Sense-making KM Model (Choo, 1998) | Sense making |
| | Knowledge creation |
| | Decision making |
| The Wiig Model for Building and Using Knowledge (Wiig, 1993) | Public Knowledge |
| | Shared experience |
| | Personal knowledge |
| The Boisot knowledge category Model (Boisot, 1987) | Propriety knowledge |
| | Personal knowledge |
| | Public knowledge |
| | Common sense |
| The Boisot I-Space KM Model (Boisot, 1998) | Codified- Uncodified |
| | Abstract- Concrete |
| | Diffused- Undiffused |
| Skandia Intellectual Capital Model of Knowledge Management (Chase, 1997); (Roos & Roos, 1997) | Equity |
| | Human Capital |
| | Customer Capital(Customer Base, Relationsips, Potential) |
| | Innovation Capital |
| | Process Capital |
| Demerest's Knowledge Management Model (Demerest, 1997) | Knowledge construction |
| | Knowledge embodiment |
| | Knowledge dissemination |
| | Use |
| Frid's Knowledge Management Model (Frid, 2003) | Knowledge Chaotic |
| | Knowledge Aware |
| | Knowledge Focused |
| | Knowledge Managed |
| | Knowledge Centric |
| Stankosky and Baldanza's Knowledge Management Framework (Stankosky & Baldanza, 2001) | Learning |
| | Leadership |
| | Organisation, structure & culture |
| | Technology |
| Kogut and Zander's Knowledge Management Model (Kogut & Zander, 1992) | Knowledge Creation |
| | Knowledge Transfer |
| | Process & Transformation Of Knowledge |
| | Knowledge capabilities |
| | Individual "Unsocial sociality" |
| Complex Adaptive System Model of KM (Bennet & Bennet, 2004) | Creating new ideas |
| | Solving problems |
| | Making decisions |
| | Taking actions to achieve desired results |

Table 1: Overview of Widely Cited Knowledge Management Models

## IX. CONCLUSION

Australia is the one of the largest producer of mineral resources and the mining industry plays a significant role in national revenue growth. Finding new ways to identify and activate the knowledge capabilities embedded in human and technological networks is a critical concern for mining organisations seeking increased efficiencies, productivity and competitive advantage in Australian and global markets. The SKM framework and discussion presented in this paper represents a useful point of departure for companies pursuing this goal.

## REFERENCES

[1] IBIS World. (2012, May). Retrieved from Mining in Australia: Market Research Report: http://www.ibisworld.com.au/industry/default.aspx?indid=55

[2] Barney, J. B. (1991). Firm resources and sustained competitive advantage. Journal of Management, 17(1), 99-120.

[3] Bennet, A., & Bennet, D. (2004). Organizational survival in the new world: the inteligent complex adaptive system. A new theoary of the firm. Burlington: MA:Elsevier Butterworth-Heinemann.

[4] Boisot, M. (1987). Information and Organizations. The Manager as Anthropologist, , Fontana/Collins, London.

[5] Boisot, M. (1998). Knowledge assets: Securing competitive advantage in the information economy. Oxford University Press, USA.

[6] Brusilovsky, D., & Brusilovskiy, E. (2008). White Paper: Data mining: the means to competitive advantage. retreived from http://www.bisolutions.us/web/graphic/BISolutions-DATA-MINING-THE-MEANS-TO-COMPETITIVE-ADVANTAGE.pdf.

[7] Chase, R. (1997). The Knowledge based Organization: An International Survey. Journal of Knowledge Management, 1(1).

[8] Choo, C. (1998). The knowing organization. New York: Oxford University Press.

[9] Dalkir, K. (2005). Title Knowledge management in theory and practice . Amsterdam ; Boston : Elsevier/Butterworth Heinemann.

[10] David, F. R. (2011). Strategic Management concepts and cases. Prentice Hall.

[11] Demerest, M. (1997). Understanding knowledge management. Journal of long range Planing, 30(3), 374-84.

[12] Duvall, M. (2002). Knowledge Management Vendors Go Vertical. Intranet journal.

[13] Easterby-Smith, M., & Prieto, I. M. (2008). Dynamic Capabilities and Knowledge Management: An integrative role for learning. British Journal of Management, 19, 235-249.

[14] Fernandez, M. C. (2010). Knowledge-intensive service activities in the success of the Australian mining industry. The Service Industries Journal, 30(1), 55-70.

[15] Freeman, R., & McVea, J. (2001). A stakeholder approach to strategic management. .

[16] Frid, R. (2003). A Common KM Framework For The Government Of Canada: Frid Framework For Enterprise Knowledge management. Canadian Institute of Knowledge Management, Ontario.

[17] Gardner, R. S. (2001, May). Stakeholder and Reputation Management as a tool for effective governance in corporatised government agencies. Journal of Contemporary Issues in Business and Government, 7(1), 29-40.

[18] Grant, R. M. (1996). Toward a knowledge-based theory of the firm. Strategic Management Jurnal, 17, 109-22.

[19] Hafizi, M. A., & Nor Hayati, A. (2006, September). KNOWLEDGE MANAGEMENT IN MALAYSIAN BANKS: A NEW PARADIGM. Journal of Knowledge Management Practice, Vol. 7, No. 3, September 2006, 7(3).

[20] Haslinda, A., & Sarinah, A. (2009). A Review of Knowledge Management Models. The Journal of International Social Research, 2(9), 187-198.

[21] Hedlund, G., & Nonaka, I. (1993). Models of Knowledge Management in the West and Japan. Implementing Strategic Process, Change, Learning and, 117-44.

[22] Hunger, D., & Wheelen, T. L. (2003). Essentials of Strategic Management. Prentice Hall.

[23] Jashapara, A. (2011). Knowledge management an integrated approach (second ed.). Sydney: Prentice Hall.

[24] Johnson, G., Scholes, K., & Whittington, R. (2005). Exploring corporate strategies (7 ed.). Financial Times Prentice Hall.

[25] Kogut, B., & Zander, U. (1992). Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. Organization Science, 3(3), 383-97., 3(3), 383-397.

[26] Lee, M.-C. (2009, September). The Combination of Knowledge Management and Data mining with Knowledge Warehouse. International Journal of Advancements in Computing Technology, 1(1), 39-45.

[27] Liebowitz, J., & Megbolugbe, I. (2003). A set of framework to aid the project manager in conceptualizing and implementing knowledge management initiatives. International Journal of Project Management, 21, 189-98.

[28] McAdam, R., & McCreedy, S. (1999). A critical review of knowledge management models. The learning organization, 6(3), 91-100.

[29] Mouritsen, J., Larsen, H. T., & Bukh, P. N. (2005). Dealing with the knowledge economy: intellectual capital versus the balanced scorecard. Journal of Intellectual Capital, 16(1).

[30] Nimmagadda, S. L., & Dreher, H. (2009). On issues of data warehouse architectures-managing australian resources data. IEEE International Conference on Digital Ecosystems and Technologies.

[31] Nonaka, I., & Takeuchi, H. (1995). The knowledge-creating company: How Japanese companies create the dynamics of innovation. New York: Oxford University Press.

[32] Noonna, J. (2000, Jul). Data Mining strategies. Retrieved from information management: http://www.information-management.com/issues/20000701/2367-1.html?zkPrintable=true

[33] Paddock, J., & Lemoine, K. (2012). Business Intelligence and Data Mining Review.

[34] Porter, M. E. (1980). Competitve Strategy: thechniques for Analyzing Industries and Competitors. New York: Free Press.

[35] Poser, T. B. (2003). Poser,T. B.(2003). The Impact of Corporate Venture Capital: Potentials of Competitive Advantages for the Investing Company. Germany: DUV.

[36] Richards, J. (2009). Mining, society and sustaiable world. Springer.

[37] Roos, G., & Roos, J. (1997). Measuring your Company's Intellectual Performance. Journal of Long Range Planning,, 30(3), 413-26.

[38] Sanchez, R., & Heene, A. (1997). Strategic learning and knowledge management. John Wiley & Sons, Inc.

[39] Shih, K.-H., Chang, C.-J., & Lin, B. (2010). Assessing knowledge creation and intellectual capital in banking industry. Journal of Intellectual Capital, 11(1), 74-89.

[40] Spender, J. C. (1996). Making knowledge the basis of a dynamic theory of the firm. Strategic Management Journal, 17, 45-62.

[41] Stankosky, M., & Baldanza, C. (2001). A systems approach to engineering a KM system. Unpublished manuscript.

[42] Turban, E. D., & Leidner, e. a. (2008). Information Technology For Management.

[43] Urzúa, O. (2011, June 29). World-class suppliers to the Mining Industry Supporting the development of a knowledge-based. Retrieved from http://www.corfo.cl/transferenciatecnologica/PDF/Technology_Transfer_Presentation_BHP_Billiton_Santiago_2011.pdf.

[44] Von Krogh, G., & Roos, J. (1995). Organizational Epistemology. New York:St.Martin press.

[45] Wiig, K. M. (1993). Knowledge management foundations: thinking about thinking: how people and organizations create, represent, and use knowledge. Arlington, TX: Schema Press.

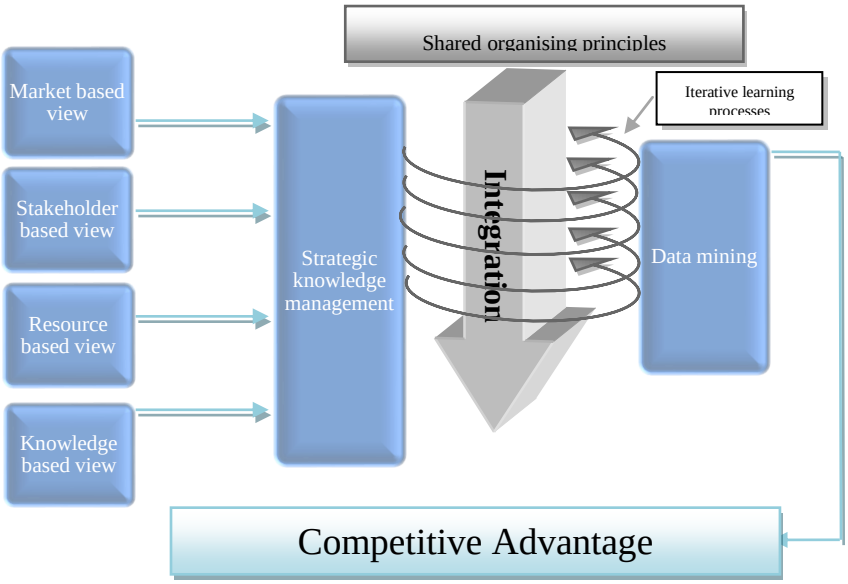



Figure 1: Creating competitive advantage through integration of Data Mining and Strategic Knowledge Management

# Managing Changes in Citizen-Centric Healthcare Service Platform using High Level Petri Net

Sabri MTIBAA
LI3 Laboratory / University of Manouba
National School of Computer Sciences
2010 Manouba, Tunisia

Moncef TAGINA
LI3 Laboratory / University of Manouba
National School of Computer Sciences
2010 Manouba, Tunisia

***Abstract*— The healthcare organizations are facing a number of daunting challenges pushing systems to deal with requirements changes and benefit from modern technologies and telecom capabilities. Systems evolution through extension of the existing information technology infrastructure becomes one of the most challenging aspects of healthcare and the adaptation to changes is a must. The paper presents a change management framework for a citizen-centric healthcare service platform. A combination between Petri nets model to handle changes and reconfigurable Petri nets model to react to these changes are introduced to fulfill healthcare goals. Thanks to this management framework model, consistency and correctness of a healthcare processes in the presence of frequent changes can be checked and guaranteed.**

***Keywords- Healthcare; requirements changes; evolution; information technology; healthcare service platform; handle changes; reconfigurable Petri nets; consistency.***

## I. INTRODUCTION

The traditional method of receiving healthcare required a patient to visit their doctor's office, or a hospital; now doctors, hospitals, and healthcare ecosystems are increasingly brought to the patient. The baby boomer generation is entering a new stage of their life; they are adopting and demanding better delivery models for quality and access to care [1]. This has caused a paradigm shift in the healthcare system.

The new culture is patient centered utilizing care coordination that is focused on successful outcomes that depend on new innovations, technology and meaningful data, including collection, delivery, ease of use, intelligence, and reporting [2]. As human life expectancy continues to increase and aging populations lead to higher health treatment costs, telehealth is on the top of the regulators agenda.

The focus on the climbing cost to deliver and maintain quality healthcare is no longer in the peripheral view, but a clear line of sight for the patients, healthcare providers, regulators, and payers. This brings new requirements for healthcare professionals to share information, communicate and collaborate in real time from multiple locations, because medicine is a collaborative science.

Communications become a strategic asset for a strongly needed healthcare transformation technology. It must be deployed to this field in order to ensure better context for medical decisions, reduce administrative costs and improve patient safety by reducing errors. The healthcare community has recognized the need to transform from the current hospital centralized treatment-based mode to prevention-oriented comprehensive healthcare mode in which hospitals, communities, families and individuals are closely involved. The new mode needs to provide individuals with intelligent health information management and healthcare services. It allows them to enjoy medical prevention and healthcare services in their daily life.

The advancement of information technology (IT) brings more opportunities for innovations in the healthcare area. The use of service oriented technologies such as SOA, Web Services (WS) allows service providers to reduce and simplify integration process, to abstract network capabilities (e.g., call control, presence, location, etc.), and create personalized and blended services (both internally and with 3rd party partners) [3]. These technologies facilitate the construction of service systems with higher reusability, flexibility, extensibility, and robustness.

Cloud computing is evolving as an important IT service platform with its benefits of cost effectiveness and global access. Built upon Enterprise Service Bus (ESB) as an integration backbone [4], this paper presents a novel citizen-centric healthcare service platform. One of the much-touted potentials of this platform is the ability to construct healthcare composite Web services on demand, relieving telecom operators from the intricate details of how technologies work so they can focus on the business aspects.

As healthcare services aggregated in the proposed healthcare service platform increases, the complexity of managing changes will grow and the manual management of changes becomes not practical [5]. One of the greatest promises of the healthcare platform is ability to self-adapting to guarantee goals achieving. Abstracting changes in business relationships claim a framework to manage changes without any impact. For instance, the proposed healthcare platform needs to achieve the plug-in/plug-out Web services with little overhead while guaranteeing properties. These properties can be classified either *functional* or *non-functional*. Functional properties address the functionalities that healthcare platform have to fulfill. The non-functional properties refer to events surrounding the functional properties.

Change management is a critical component in the deployment of healthcare platform. We identify two main approaches dealing with changes: top-down and button-up [6]. A top-down approach focuses on changes that are usually business mandated. These changes are motivated by the business goal, and do not consider the uncertainty of the underlying member services. The second type of changes is referred to as bottom-up changes because Web service providers are the initiators of changes. Bottom-up changes are initiated by the member services. These changes are initiated in the Web service environment, and eventually translate into top-down changes. A service operation may become unavailable and trigger dependencies services and users in order to replace this service. In this paper, we will concentrate on this aspect.

In this paper, we present a conceptual module for management of bottom-up changes using Petri nets. In our work, we use Petri nets to model handling and adaptive changes in healthcare platform. We model changes using Petri nets because of their applicability to a Web services composition modelling.

The behaviour of a composite Web services is described by the evolution of its Petri net model [7]. As the Petri net evolves, the system goes through different safe and unsafe states that can be completely defined by the marking of a Petri net model. Furthermore, Petri nets map directly to our change specification. They also preserve all the details of our change specification while modelling the changes accurately. For instance, Petri nets can easily represent the safe an unsafe states of web services composition. They represent changes between these states as transitions. In addition, the use of reconfigurable Petri nets allows us to incorporate our mapping rules into the Petri net model. This allows us to completely model our change specification.

The remainder of this paper is organized as follows. In Section 2, we use the healthcare platform architecture of and a scenario from this domain to motivate our work. It will also be used as a running example. Section 3 presents a bottom-up specification of changes. In Section 4, we describe our change management model which is based on Petri nets. Finally, we conclude in section 5.

## II. HEATHCARE SERVICE PLATFORM ARCHITECTURE

In this section, we present the global context of our work and an overview about service oriented architecture, cloud computing and enterprise service bus. Then, the healthcare services platform architecture is exposed and some basic concepts and definitions are explained.

### A. Context and Background

Below we have summarized a few key notions and technologies that should be of significant value to the design of healthcare architecture.

Service Oriented Architecture (SOA): In this IT architecture, applications and more discrete software functions are network-based, loosely coupled and available on demand to authorized users or to other applications or services. Although SOA is not a new concept, the emergence of Web services as a standard way to expose, describe, access and combine services has given new life to this approach to computing.

The key idea of SOA is the following: a service provider publishes services in a service registry [7]. The service requester searches for a service in the registry. He finds one or more by browsing or querying the registry. The service requester uses the service description to bind service. These ideas are shown in Fig. 1.

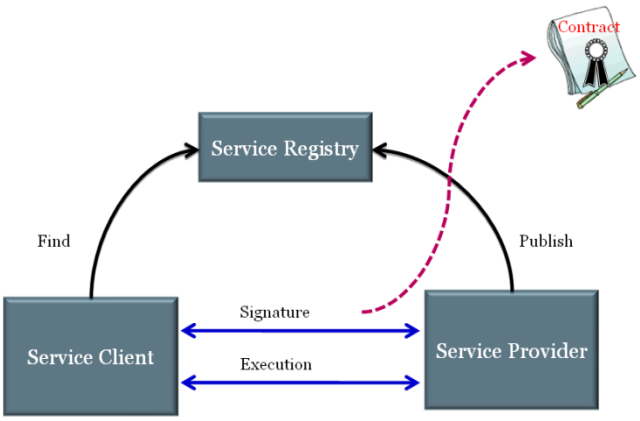



Fig. 1. Reference Architecture Of Web Services- SOA

*1) Cloud Computing:* Cloud computing called also *utility computing* refers to an IT service model and platform that provides on-demand based IT services over the internet (see Fig. 2). The five essential characteristics are: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service [8]. The three services models include:

- SaaS (Software as a Service) which delivers software service on demand, such as, salesforce.com – Customer Relationship Management (CRM) service and Google Gmail;
- PaaS (Platform as a Service) which provides the computing platform for companies to deploy and customize business applications on demand, such as, Google App Engine and Microsoft's Azure;
- IaaS (Infrastructure as a Service) which offers data center, infrastructure hardware and software resources on demand, such as, Amazon Elastic Compute Cloud (EC2) and VMware vCloud Datacenter. Both of these resources provide virtual computers for renters to run their business applications.

The four major deployment models include: private cloud, public cloud, community cloud, and hybrid cloud. Companies normally adopt different service models and deployment models depending on their unique business processes and demands on IT services.

Cloud computing today is an evolution and application of modern ICT including server virtualization, autonomic computing, grid computing, server farm, network storage, and web service.

2) *Enterprise Service Bus:*

ESB is one piece of an infrastructure that might help facilitating the implementation of a SOA, but it is not a perquisite. There are many aspects of an ESB that fit well with the SOA model, and denying its possible usefulness would be counterproductive, but the two are not completely inter-dependent [4].

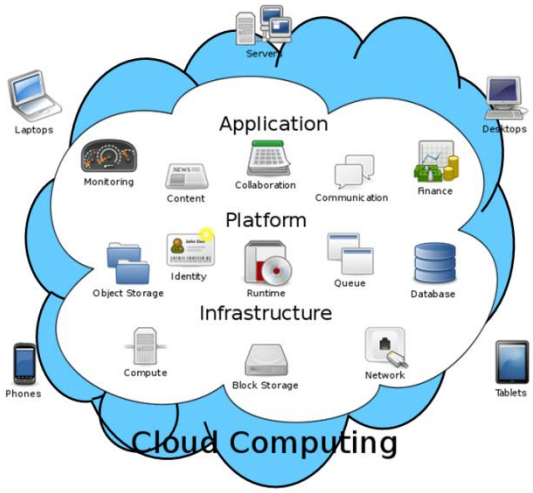



Fig. 2. Cloud Computing Architecture

Fig. 3 Depicts The Base Functional Elements Within An ESB. It Includes:

- Data transformation.
- Application adapters.
- Automation of processes.
- Transformation.
- Routing.
- Messaging.
- Event triggering.

If we consider some of these functional elements it can be seen that items such as application adapters fall neatly into the product category, while routing and messaging are more of an architectural consideration.

*B. System Architecture*

First, we present our Healthcare Service Platform (HSP). It intends to provide personalized healthcare services for the public. The healthcare value chain is complex. It consists not only of healthcare providers, but also of payers (government, employers and patients), fiscal intermediaries, distributors and producers of pharmaceuticals and devices [9].

The HSP does not attempt to address this complete value chain. It focuses on the delivery of healthcare services. It is an end-to-end reference architecture that focuses on meeting the needs of citizens, patients and professionals. Its architectural diagram is given in Fig. 4.

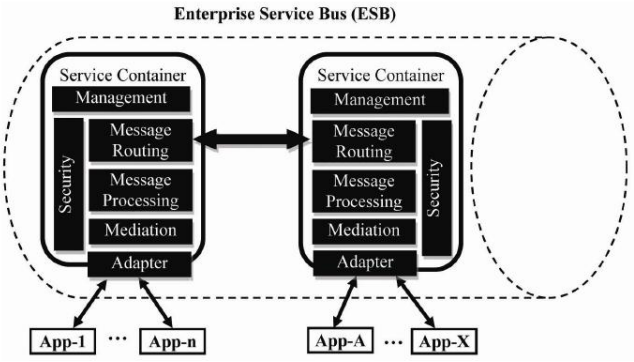



Fig. 3. ESB architecture

We distinguish three main components, i.e. body sensor networks (BSN), IaaS cloud, healthcare delivery environment.

- BSN: according to circumstances and personalized needs, appropriate health information collection terminals (i.e. sensors) are configured for different individuals. BSN is used to provide long term and continuous monitoring of patients under their natural physiological states. It performs the multi-mode acquisition, integration and real-time transmission of personal health information anywhere [10].
- IaaS cloud: modern healthcare is information driven. Healthcare providers are making progress in building an integrated profile of patients. This data sits in systems throughout the enterprise including the HER and many other electronic systems throughout the enterprise and community [11]. This component achieves the rapid storage, management, retrieval, and analysis of massive heath data. It mainly includes *Electronic Medical Record* (EMR) repository. It considers also personal health data acquired from BSN.
- Healthcare delivery environment: it includes a personal health information management system. It replaces expensive in-patient acute care with preventative, chronic care, offers disease management and remote patient monitoring and ensures health education/wellness programs.

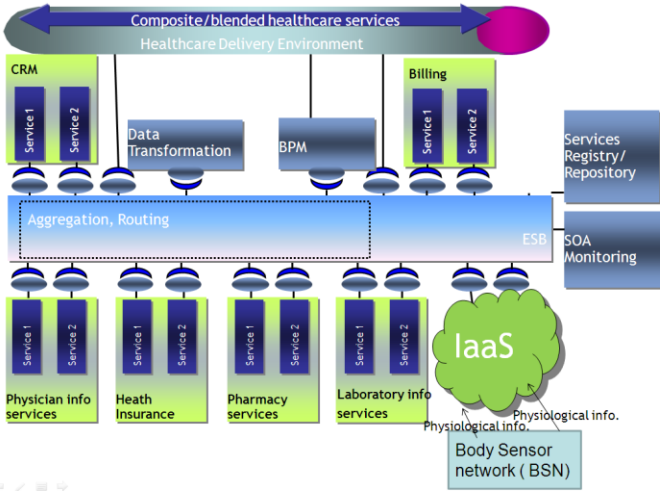



Fig. 4. HSP Architecture

### C. Healthcare Web Services Provided by HSP

In PHISP, we adopt the design idea of SOA and Web service technology for its design and implementation. The majority of its functional modules are developed and packaged in the form of services [8]. Here, we overview some of them as follows.

- *PhysInfoWS*: this service can acquire some general physiological signals such as body temperature, blood pressure, and saturation of blood oxygen, electrocardiogram, and some special physiological signals according to different sensor deployment for different users. User's ID number is required.
- *EnvInfoWS*: for a unique ID number, this service can acquire temperature, humidity, air pressure and other environmental information for this user.
- *SubjFeelWS:* it can acquire the user subjective feelings, food intake, etc., and the information is often provided by the user from the terminal.
- *CoronaryDiagWS:* it can analyze the information according to a series of analysis models, which are built for coronary heart disease, and then produce preliminary diagnostic results.
- *AssessmentWS:* this service can assess the status of the patient's health risk based on the diagnostic results and the EMR information of the patient.
- *EmrWS:* this service can output the user's medical history information.
- *GeoWS*: it can return the user's location.
- *EmerWS*: it can raise an alarm to the user in case of illness.
- *GuideWS*: it can provide the patient with preventive measures especially items that need attention.

### D. Healthcare Service Scenario

A way to motivate and illustrate this work, we presents an example of healthcare service scenario. We distinguish three main layers: service, business and HSP. The service layer consists of available web services, and the business layer represents the Web service like operations typically ordered in a particular application domain. We refer to the selected services as member services (see Fig. 5).

Key healthcare environment objectives include:

- **Allowing people to stay in their homes to an older age.** By doing this, we can reduce the economic burden of dedicated care facilities and improve quality of life for a substantial proportion of the aging population.
- **Using televisions to keep in touch.** Another use of camera technology is in conjunction with an IPTV set-top box and connection back to a Contact Center.
- **Using wireless toys for always-on monitoring and communications**. The wireless home network itself enables a new class of device that has significant healthcare implications.

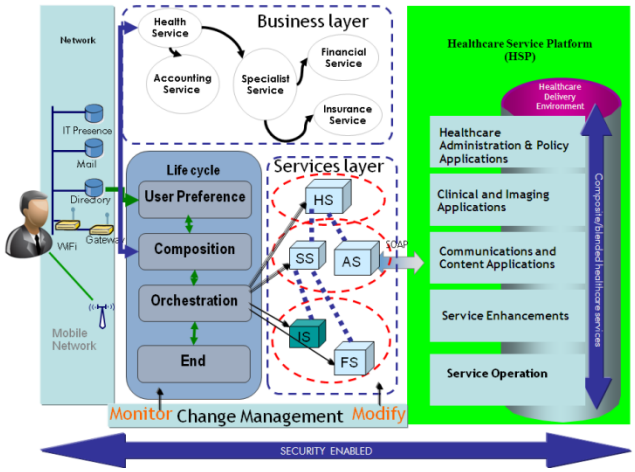



Fig. 5. Healthcare Service Scenario Based On HSP

Let us assume that a *citizen* establishes a need for a business objective (healthcare service). Typically, he starts with formulating the business strategy (or goal). During the planning, some services can be identified: *HealthService*, *AccountingService*, *SpecialistService*, *FinancialServcie* and *InsuranceService*. Second, the *senior citizen* develops a specification listing the services to be composed through a graphical interface. We assume that *HS*, *AS*, *SS*, *FS* and *IS* are selected and orchestrated. The third step is the orchestration where member services that match the specified high level configuration are selected and invoked.

We describe here the ideal scenario: the *senior citizen* subscribe to *HealthService*. Then all information regarding who contacts it and when are forwarded to *AccountingService*. *HealthService* forwards also the received data to *SpecialistService* in charge of checking the received values. After analyzing the received values, the team sends a confirmation or an adjustment of the medication doses. The *FinancialService* and *InsuranceService* are executed to finalize the process.

### E. Modeling Healthcare process using High Level Petri Nets

The modeling of healthcare process is as crucial as the implement of healthcare service platform. The formal representation of real healthcare process with Hierarchical Petri-Nets is easy to understand. Fig. 6 shows a Hierarchical Petri Net that describes our healthcare service scenario offered by HSP.

From the above, the first part of hierarchical healthcare process net N with refinable transition named 'Health Service is shown below. It is refined with the attachment of web services net N'. The planned web service is the assembly of the set of web services presented previously (*PhysInfoWS, EnvInfoWS, SubjFeelWS, CoronaryDiagWS, AssessmentWS, EmrWS, GeoWS, EmerWS, GuideWS).*

However some changes to member services may handle some inconsistency in the composition and orchestration. Each

service layer change presents a functional and non-functional change that may happen in a member service.

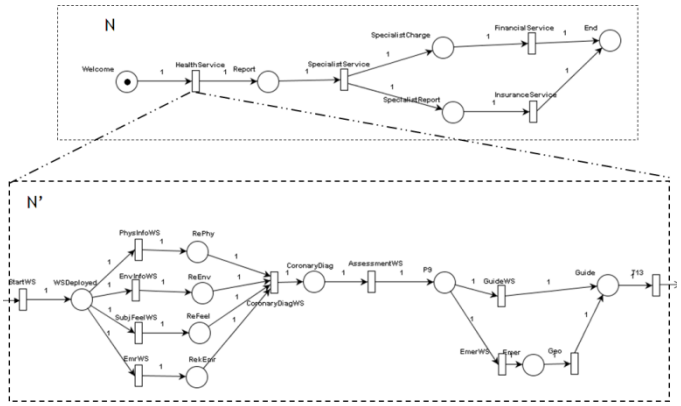



Fig. 6. Healthcare Process Modeled By Hierarchical Petri Nets

For example, in case of non-availability of a *SpecialistService*, a change management is required to ensure that the healthcare system is remaining profitable.

## III. CHANGE SPECIFICATION

Managing bottom-up changes is highly dependent on the services that compose the system. Therefore, it is quite important to define the changes that may occur to web services and then map them to into the system level.

In this section, we present bottom-up changes. We define the set of handling changes. Handling change ( $\theta$ ) is defined for changes that occurred at the service level (for example Web service availability) while adaptive changes ( $\Omega$ ) are related to changes at the business level (for instance the selection of alternative service).

### A. Changes Overview

For each change, a transition will be associated between two states: precondition and postcondition. In our scenario presented, a precondition for SS unavailability is that it was available and the postcondition is that it has switched to unavailable state.

Handling changes will be modeled using Petri nets. Our classification of triggering changes is based on the traditional approaches from the fields of software engineering and workflow systems [7]. A handling change is initiated at the service level such as the operations, the availability, etc. Therefore, we can distinguish several handling changes based on Web service properties.

The Web service properties can be sorted into two categories: functional and non-functional. Fig. 7 shows the handling changes: functional and non-functional.

- *Non-functional changes:* we assume that the non-functional parameters represent the dependability and response aspects associated with a member service. Service dependability can be set to two possible values (i.e., available or unavailable). Alternatively, service cost values may take more than two possible values.

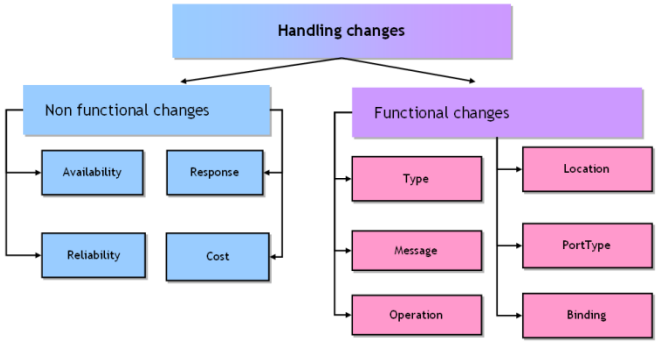



Fig. 7. Handling Changes Categories

- *Functional changes*: this category of changes is related to a service WSDL description [12]. We consider functional changes as a combined execution of elementary operations: *remove* and then *add*. We distinguish two different functional changes: structural and behavioral (see Fig. 6). Structural changes are related to the operational aspects of a Web service. For example, a structural change in a healthcare service can be a consequence of change in the operations offered to a citizen. Functional changes to a member Web service occur when its WSDL description is modified.

- *Adaptive changes:* Adaptive changes may occur at the composition and orchestration levels. Fig. 8 shows the adaptive changes considered in our model. In our scenario, when the healthcare system is interrupted by a change in *SS*, it reacts to the change after suspending execution. This may be accomplished by raising a fault, compensating for the change at the composition layer, and calling of an alternate service.

For example, if *SS* becomes unavailable, business layer will be search for equivalent service to continue execution to ensure that there is no high level impact on user demands.

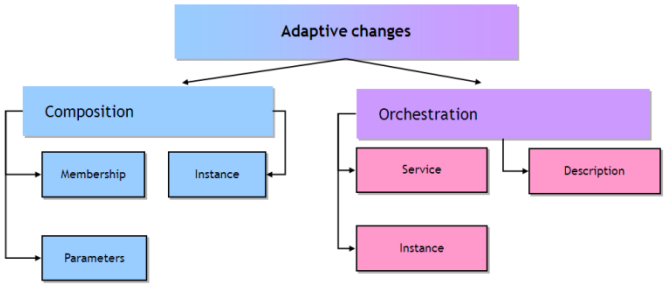



Fig. 8. Handling Changes Categories

Now, we will discuss the impact of $\theta$ changes to the healthcare business layer. A mapping details how change instances in one layer correspond to changes in another layer. These mapping must remain consistent in the presence of frequent changes.

Handling changes have a reactive impact on the business layer. For instance, a $\theta$ change in availability maps to $\Omega$ change of change instance.

## IV. CHANGE MODEL

In this section, we introduce a change model to accurately identify eventual types of changes in a composite Web services.

### A. Handling Changes Model using Petri Nets

Petri nets or PN are a well-founded process modeling techniques that have formal semantics. They have been used to model and analyze several types of processes including protocols, and business processes.

Visual representations provide a high-level, yet precise language, which allows reasoning about concepts at their natural level of abstraction. Services are basically a partially ordered set of changes. Therefore, it is a natural choice to map it into a Petri net. Moreover, the semantics delivered by Petri nets can be used to model the standard behavior of composite Web services described by BPEL [7].

We formalize the change model for triggering changes by introducing Petri-Net-Handle (PNH) which is defined as follows.

The algebraic structure of PNH = ($P$, $T$, $F$, $P_0$, $P_n$) if the following conditions hold:

- $F \subseteq (P \times T) \cup (T \times P)$
- $P \cap T = \varnothing$
- $P \cup T \neq \varnothing$
- $P_i \in P$
- $P_0 \in P$

where:

- $P$ is a finite set of places representing the states of Web service.
- $T$ is a finite set of transitions representing changes to Web service.
- $F$ is called the web services action flow.
- $P_0$ is the input place, or starting state of the Web service
- $P_n$ is the output place, or the ending state of the Web service

Fig. 9 represents the model of non-functional changes to Web services. It is composed from five places and four transitions. *PS* is the initial place of $PNH_N$. It represents the initial state of the Web Service. *PS* consists of four tokens, each representing one of the four non-functional changes. The token corresponding to change is fired each to represent dynamic evolution. If more than one change occurs, the corresponding token for each change type is fired.

For instance, if a member services (i.e Web service) becomes unavailable, the transition will be enabled and the corresponding token will be fired.

The subnet representing dependability changes in $PNH_d$= ($P_d$, $T_d$, $F_d$ , $P_{0d}$, $P_{nd}$), where $P_d$={*PS,PS'Re,PS'A*} and $T_d$={*TRe,TA*},. The place *PS* is corresponding to the state of available and reliable service. *PS'A* represents a service that becomes unavailable.

Table I. gives summary about non-functional changes.

TABLE I. NON FUNCTIONAL CHANGES

| Change | $\theta$ | ***Pre*** | ***Post*** |
|---|---|---|---|
| alterAvailibility | $\theta_A$ | PSA | PS'A |
| alterReliability | $\theta_R$ | PSR | PS'R |
| alterCost | $\theta_C$ | PSC | PS'C |
| alterResponsivenss | $\theta_{\mathrm{Re}}$ | PSRe | PS'Re |

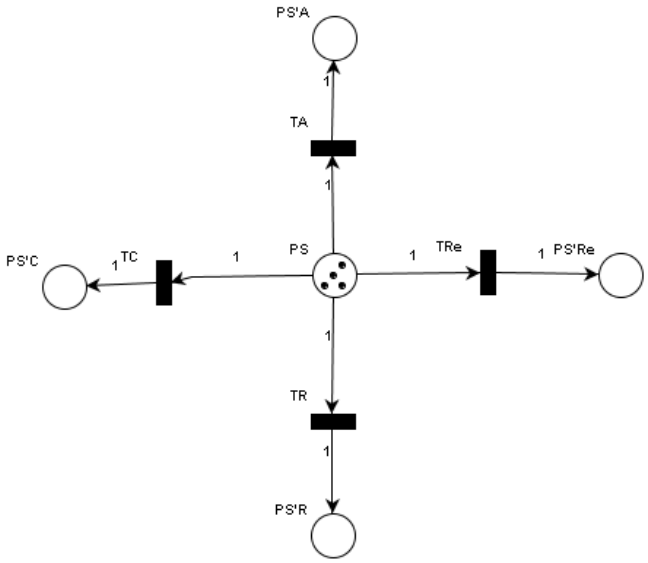



Fig. 9. Handling non-functional changes categories

When a service becomes unavailable, the token (representing the availability property) is moved from *PS* to *PS'A*. A similar behavior is observed when the service becomes unreliable. *TA* represents alterAvailibility, and *TRe* represents alterReliability.

The same logical structure can be applied for functional changes.

### B. Modeling Adaptive Changes with Reconfigurable Petri Nets

We have surveyed extensions of Petri nets for modeling reactive changes. Reconfigurable Petri-nets provide formalism for modeling these changes. It is a class of high level Petri nets.

They support internal and incremental description of changes over a uniform description. Reconfigurable Petri nets are an extension of Petri nets with local structural modifying rules performing the replacement of one of its subnets by other subnets [13]. The tokens in a deleted place are transferred to the created one.

We formalize the change model for adaptive changes by introducing PNAC (PNAC) which is defined as follows. The algebraic structure of PNAC= (P, T, F, R, I) where:

- $P$= $\{P_1 \ldots, P_n\}$ is a non-empty and finite set of places
- $T$=$\{T_1,...,T_n\}$ is a non-empty and finite set of transitions disjoint from $P(P \mathrm{I}\ T = \Phi)$

- *F:* $(P \times T)\,\mathrm{Y}\,(T \times P) \rightarrow$ IN is a weighted flow relation. A rewriting rule is a map $r : P_1 \rightarrow P_2$ whose domain and co domain are disjoint subsets of places $P, \mathrm{P}_1 \subseteq \mathrm{P}, \mathrm{P}_1 \mathrm{I}\ \mathrm{P} = \Phi$
- $R=\{r_1,..,r_n\}$ is a finite set of structure modifying structure rules.
- *I* represent the initial state: the first configuration of composition in business layer. The domain of *I* is $HCE_0$.

Table I. gives summary about adaptive changes.

TABLE II. ADAPTIVE CHANGES

| Change | $\Omega$ | *Pre* | *Post* |
|---|---|---|---|
| alterState | $\Omega_{ST}$ | VEST | VE'ST |
| alterServiceInstance | $\Omega_{S}$ | VES | VE'S |
| alterCost | $\Omega_{C}$ | VEC | VE'C |

We consider the scenario containing five places corresponding to adaptive changes:

- $HCE_S$ is the set of places {$HCE_0$, $HCE_1$, $HCE_2$, $HCE_3$, $HCE_4$} where S represents alterState.
- $HCE_v$ is the set of places {$HCE_5$, $HCE_6$, $HCE_7$, $HCE_8$, $HCE_9$} where V denotes alterServiceInstance.
- $HCE_w$ is the set of places {$HCE_{10}$, $HCE_{11}$, $HCE_{12}$, $HCE_{13}$, $HCE_{14}$} where W denotes alterOrder.

Fig. 10 shows a PNAC representing initial statechange in service orchestration. The adaptive changes using Reconfigurable Petri net representing modification on service state (see Fig. 11), removal of service (see Fig. 12), and addition of service (see Fig. 13) are presented.

### C. Change Management Framework

We use our Petri net change specification as the basis for handling changes in our healthcare environment. The framework of change management is divided into two modules: detection and reaction.

After the change specification is defined, we begin the management. Detecting the respective changes is the first step of change management.

All changes identified in the handling changes models are subject to detection. Detection involves an agent that monitors the Web service. Each change type has an associated set of rules for detection. For example, a *SpecialistService* may change the input parameters (i.e required information provided by patient), when this change occurs, the healthcare system must detect this change using some predefined detection rules.

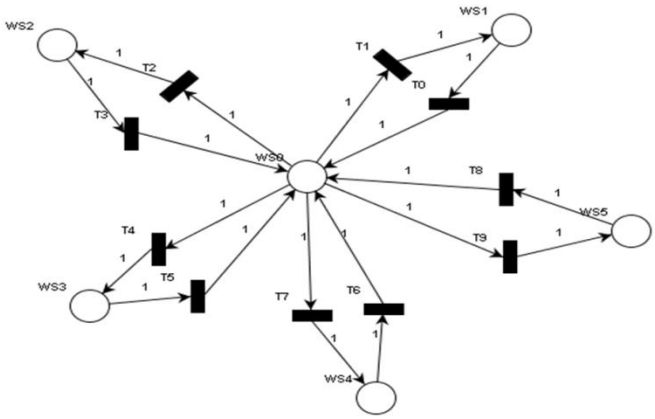



Fig. 10. Reconfigurable Petri Nets For Reactive Changes- Initial State

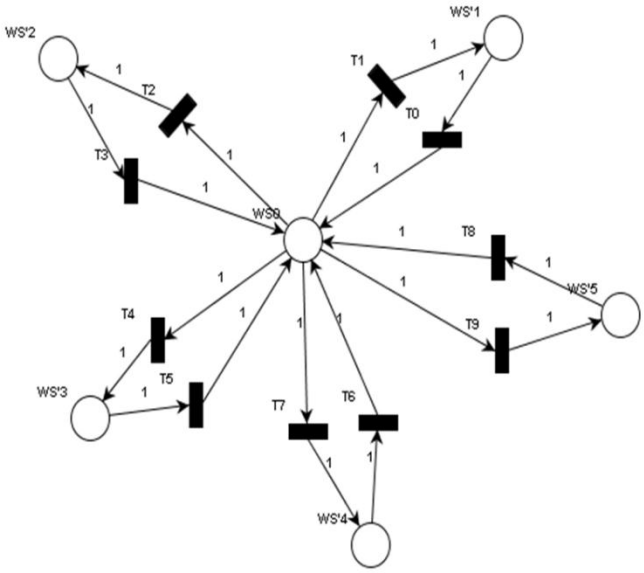



Fig. 11. Reconfigurable Petri Nets For Reactive Changes- After Change

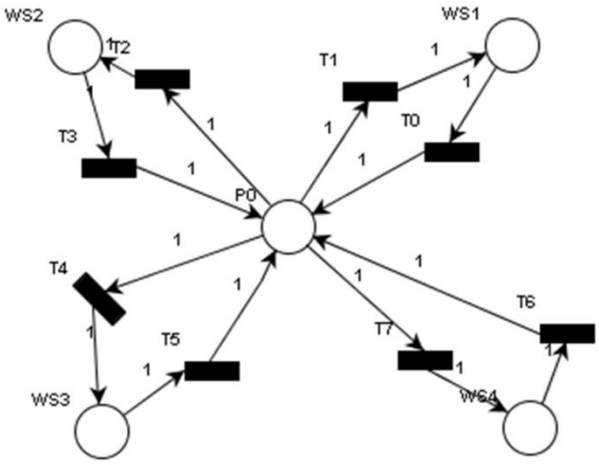



Fig. 12. Reconfigurable Petri Nets For Reactive Changes- After Removal Of Service

Each detected change must be forwarded to monitoring service; and then the composition strategy must be updated. The notification and polling mechanism are mainly the techniques to awareness that a change has occurred. These techniques require that a monitoring service periodically send "Refresh" and "Alive" messages to detect unavailable services and also renew membership.

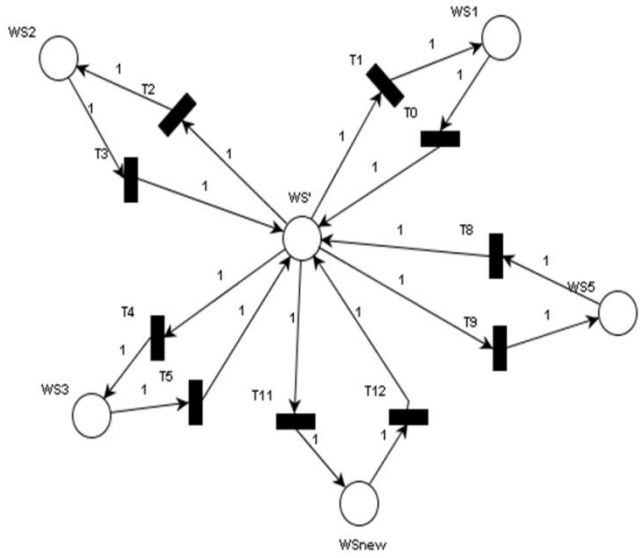



Fig. 13. Reconfigurable Petri Nets For Reactive Changes- After Addition Of New Service

The changes are detected at the service layer and represented as an incidence matrix. Some rules are identified for detecting functional and non functional changes.

We define a rule for mapping the change into the defined Petri-net: first of all, the current service state is corresponded to a set of precondition places in the triggering Petri net and the updated service state as the set of postcondition places in the triggering Petri net. Then, a comparison between the values precondition and postcondition places of the Petri net is done. Depending on result retuned, a token is placed in the respective precondition place. This token will enable the change transition only if a difference is found.

Let us consider the example where the service WS service change availability (due to maintenance reasons) and the other attributes remain constant. In this case, we map the change into the non-functional Petri net. The service agent responsible of monitoring of WS will generate the incidence matrix corresponding to the Petri net model. This agent is in interaction with healthcare service platform to detect effective changes in the execution environment and map it into a Petri net model defined in this section ( for example, unavailability of a WS). The Centralized agent is the module that reacts to these changes and purpose a reconfiguration in the execution environment to guarantee consistency and correctness of the healthcare processes. Fig. 14 shows the different modules designed for management framework to detect and react to different changes in healthcare services.

Based on the information sent by service agent we define how to execute adaptive change. After receiving the matrix indicating the change that occurred, the handling change is mapped to the appropriate reactive change. We can list some considered reaction techniques in our system:

- In case of add, the newly service member will be considered. It can be taken into account in load balancing context or as back-up alternative.
- In case of unavailability of a service, if it is critical then the orchestration will be paused. Since a heartbeat is activated to check the status of the service; the orchestration will wait the service availability otherwise orchestration will exit (depending on configurable heartbeat number).

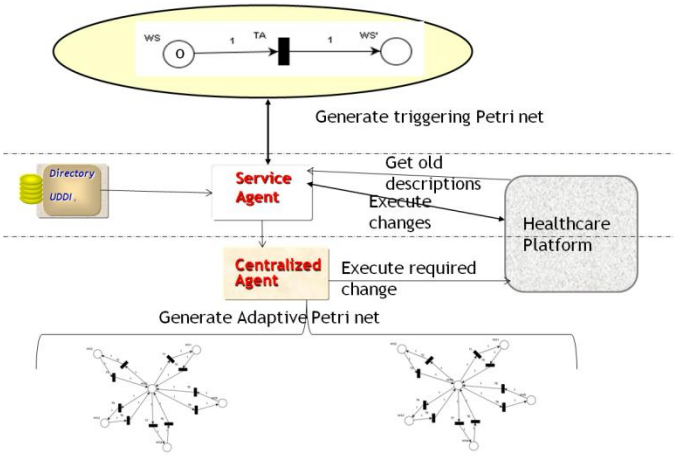



Fig. 14. Adaptive Petri Net Generator Module In Change Management Framework

## V. Conclusion And Future Work

In this paper, we first presented a novel architecture of healthcare services platform. Second, we exposed the bottom-up approach focusing on handling changes that may occur in this system and then mapped to adaptive changes. We used a formal change model based on High Level Petri Nets to accurately represent these changes.

Future work includes an extension of change management framework. We plan to include a top-down approach to specifying changes. A full simulation prototype taking into account priority in changes and the estimation of their frequency based on measures represent actions planned to enhance actual healthcare platform.

## References

[1] M. d. Reuver, and T. Haaker, "Designing viable business models for context-aware mobile services," Telematics and Informatics, Vol. 26 , pp. 240–24, 2009.

[2] V. .s. .Kondapall, J. Raju, K. R. Rao, and K. tanuja ," Patient-Centric-Integrated EHA OF Physician Pratice Portal Through Cloud Computing Technology," International Journal of Computer Science & Communication Networks, ,Vol 2, pp. 172-18, 2012.

[3] N. Kryvinska, C. Strauss, and P. Zinterhof, "Next Generation Service Delivery Network as Enabler of Applicable Intelligence in Decision and Management Support Systems Migration Strategies, Planning Methodologies, Architectural Design Principles," Studies in Computational Intelligence, Vol. 352, pp. 473-502, 2011.

[4] L. Chen," Integrating Cloud Computing Services Using Enterprise Service Bus (ESB)," Business and Management Research,Vol.1, No. 1, pp. 26-31, 2012.

[5] L. Baresi, and L. Pasquale., "Adaptive Goals for Self-Adaptive Service Compositions," *Proceedings of Web Services (ICWS)*, pp. 353 - 360, 2010.

[6] S. Akram, A. Bouguettaya, X. Liu, A. Haller, and F. Rosenberg, "A Change Management Framework for Service Oriented Enterprises," International Journal of Next-Generation Computing (IJNGC), Vol. 1. No. 1, 2010.

[7] S. Mtibaa, and M. Tagina "An Automated Petri-Net Based Approach for Change Management in Distributed Telemedicine Environment," Journal of Telecommunications,Vol.15, No. 1, pp. 1-9, 2012.

[8] P. Wang, Z. Ding, C. Jiang, and M. Zhou, "Web Service Composition Techniques in a Health Care Service Platform," *IEEE International Conference on Web Services*, pp. 355-362, 2011.

[9] A. Azees," Challenges and Performance Enhancement in Could Computing," IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS),Vol.1, No. 1, pp. 49-54, 2011.

[10] M. Brenner, M. Grech, and M. Unmehopa, "An enabler gateway for service composition using SIP" 1*3th International Conference on Intelligence in Next Generation Networks*, , pp. 1-6, 2009.

[11] V. .s. .Kondapall, J. Raju, K. R. Rao, and K. tanuja ," Patient-Centric-Integrated EHA Of Physician Pratice Portal Through Cloud Computing Technology," International Journal of Computer Science & Communication Networks, ,Vol 2, pp. 172-18, 2012.

[12] A. Chehri, and H. T. Moutah, "Survivable and Scalable Wireless Solution for E-health and Eemergency Applications," *EICS4Med*, pp. 25-29, 2011.

[13] P. Shanchen, L. Yin, H. Hua and L. Chuang, "A Model for Dynamic Business Processes and. Process Changes," Chinese Journal of Electronics, Vol. 20, No. 4, 2011.

AUTHORS PROFILE

**Sabri Mtibaa,** He is currently a Ph.D. student in the National School for Computer Sciences of Tunis, Tunisia (ENSI). He received the master degree from High School of Communication of Tunis, University of Carthage, Tunisia (Sup'Com) in 2008. His current research interest includes web service composition using Petri nets as well as system verification and QoS aware.

**Moncef Tagina**, He is a professor of Computer Science at the National School for Computer Sciences of Tunis, Tunisia (ENSI). He received the Ph.D. in Industrial Computer Science from Central School of Lille, France, in 1995. He heads research activities at LI3 Laboratory in Tunisia (Laboratoire d'Ingénierie Informatique Intelligente) on Metaheuristics, Diagnostic, Production, Scheduling and Robotics.

# Software Architecture- Evolution and Evaluation

S.Roselin Mary
Department of Information Technology,
Hindustan University,
Chennai, India.

Dr.Paul Rodrigues
CTO,
WisdomTree Software Solutions,
Chennai, India.

***Abstract*— The growth of various software architectural frameworks and models provides a standard governing structure for different types of organizations. Selection of a suitable framework for a particular environment needs much more detailed information in various aspects and a reference guide of features should be provided. This paper brings out the history of software architecture with a new evolution tree. It also technically analyses well known frameworks used in industries and other governmental organizations and lists out the supportive tools for them. This paper presents the comparative chart that can be used as a reference guide to understand top level frameworks and to further research to enable and promote the utilization of these frameworks in various environments.**

***Keywords- Framework; Software Architecture; Views.***

## I. INTRODUCTION

Architecture is playing a vital role to reveal the complexity of a given system. Number of steps will be increased when the system becomes complex. Planning should be done in a detailed manner when the system becomes complex. Architecture comprises the combination of process and product of planning, designing and constructing space to reflect functional, social and aesthetic considerations [21]. Planning for buildings and complexity behind this will be interrelated in civil engineering. As customers and constructors have their own views in a particular subject, the architecture should solve it in a unique manner by covering all of them [13].

Likewise the same concept in software is called software architecture. The term and concept of Software architecture was brought out by the research work of Dijikstra in 1968 and David parnas in 1970's. The interconnected basic building components and the views of end user, designer, developer and tester are needed to build a complicated, critical system. The design and implementation of the high-level structure of the software are the backbone of software architecture. The architectural elements will be interconnected in well-known manner to get the major functionality and performance requirements of the system and to obtain non-functional requirements such as reliability, scalability, portability, and availability [12]. Software frameworks points out the suitable places in the architecture where specific functionality can be adapted by application programmers [17]. A software framework provides an abstraction where generic functionality can be selectively overridden or specialized by user code. The overall development time will be cut into minimum as it concentrates on the low level details of a working system. So, the designers and programmers can concentrate only on the software requirements. [7].

The rest of the paper is organized as follows. Section II briefly describes the history of Software architecture and the figure Fig.1 given below clearly portrays the evolution. Section III classifies the frameworks. Section IV and V summarizes and compares the different frameworks.

## II. HISTORY OF SOFTWARE ARCHITECTURE

The basic principles of 'software architecture' have been applied since the mid 1980's and it crossed various stages from algorithm's era by borrowing the concepts from others to get a shaped form. In 1928, An Algorithm was formulated to solve the problem by the finite sequence of instructions. Von Neumann developed 'Flow Chart' that has a visual representation of the instruction flow, to plan computer programs in 1947 by inheriting the idea from the flow process chart(1921) and multi flow chart(1944)which were used mostly in the area of electrical engineering. But, there is a gap to point out the flow of control. So, 'Control Flow Diagram' (CFD) was developed in the late 1950's to describe the control flow of a business process and program. This was not enough to view the complex systems. The high level view of the work and immediate access of particular points can't be represented using this diagram. So, to reveal the entire system by dividing into blocks, 'Block Diagram' was developed in late 1950's. A specific function for each block and the connection between blocks will be shown in a diagram.

The introduction of abstraction concept became a booster in the field of software architecture. It made a revolution and tremendous growth to that area. By that way, data structures that have similar behaviour, data structures that have similar behaviour, certain data types and modules of one or more programming languages that have similar semantics are grouped in the late 1960's.This was happened by the introduction of Abstract data types. It again leads to 'Modular Programming' that introduces the concept of separate parts called modules in software in 1968. Separation of concerns with the logical boundaries between components is called as modules.

In 1977, 'Three Schema Approach' that adopts layered architecture based on the modular programming was developed. It is used to build information systems using three different views in systems development. Here an application will be broken into tiers and developers have to modify a

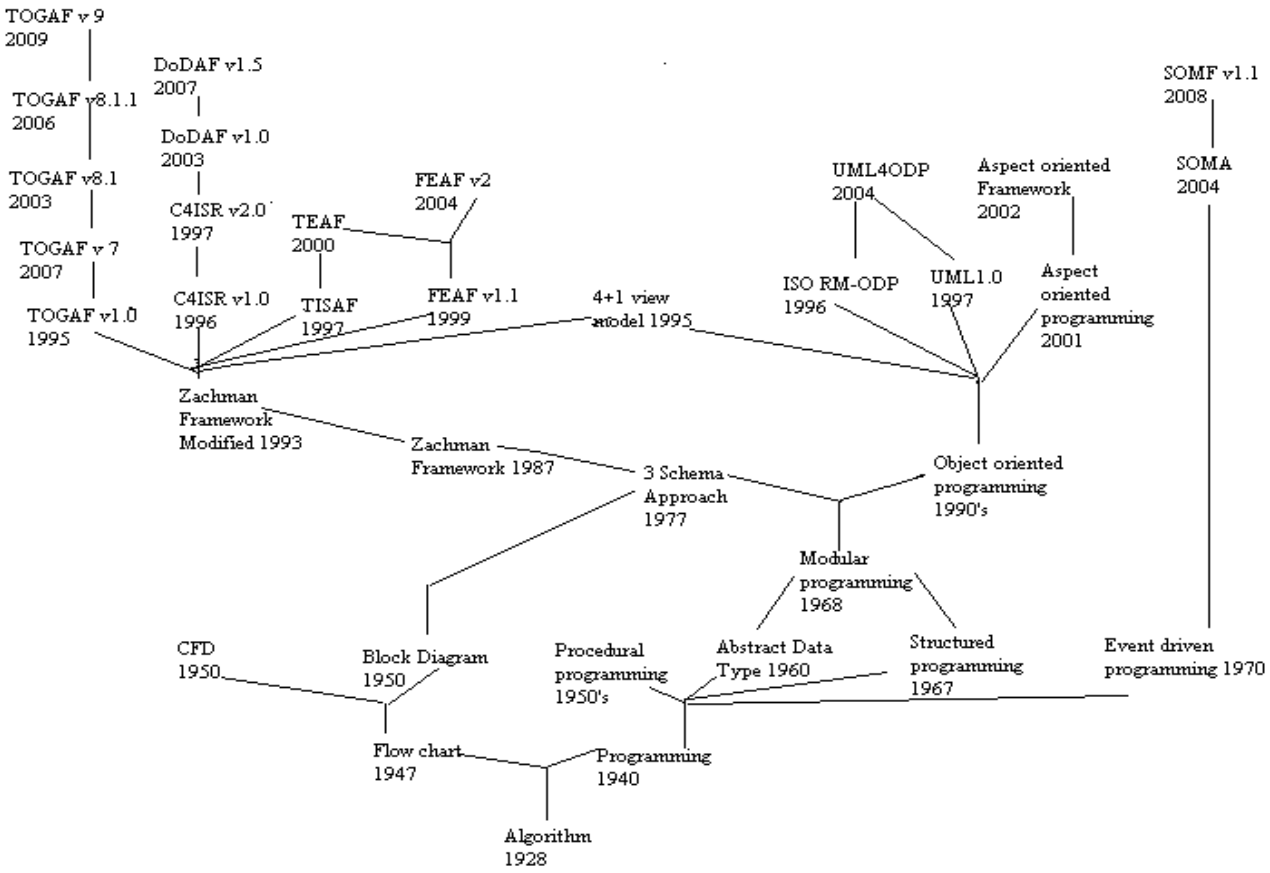



Figure 1. Evolution of software Architecture

specific layer not to rewrite the entire application over. Flexible and reusable applications can be developed using this scheme.

Later, based on this three tier approach, a layer of six perspectives was introduced in 1987 by John Zachman .That is called as 'The Zachman Framework' which still plays an important role in the era of 'Enterprise Architecture' and influenced frameworks DODAF, TOGAF, TEAF and FEAF. In 1993Zachman released the modified version of Zachman Framework with more number of views. In 1995, 4+1 view model was developed by Kruchten.

Views are used to analyze the complex systems, and to list out the problem elements and the solution. A view of a system suppresses details. It focuses on specific concerns of the system. It provides a simplified model [13] [12].

U.S Government encouraged the researchers to develop the frameworks for defense side applications and it leads to the C4ISR Architecture Framework in 1996. 'The Department of Defense Architecture Framework (DODAF)' was released n 2003, which restructured the C4ISR framework ver2.0 [19] [6].

The restructured C4ISR framework ver2.0 was released as, 'The Department of Defense Architecture Framework (DODAF)' in 2003[19] [6]. 'The Open Group Architecture Framework (TOGAF)' was developed by the members of open architecture forums in 1995. Recently in 2009, TOGAF Version 9 was released [15].

To integrate its myriad agencies and functions under single common and enterprise architecture, the 'Federal enterprise Architecture Framework (FEAF)' was developed in 1999 by the Federal Government [18].

'Treasury Enterprise Architecture Framework (TEAF)' was developed to support the Treasury's business processes in terms of products of the US Department of Treasury and published in July 2000 [20].

A reference model RM-ODP was developed by Andrew Herbert in 1984. It combines the concepts of abstraction, composition and emergence on the distributed processing developments. By including the set of UML profiles in the ODP and UML4ODP was released in 2004[10].

In 2001, Aspect oriented programming boom out by inheriting the principles of OOPS. And, it leads to the Aspect oriented software development in later 2002.

IBM announced 'Service Oriented Modeling Architecture (SOMA)' in 2004 opposing the distributed processing and Modular programming. It is the first publicly announced SOA related methodology. In addition to this, to provide tactical and strategic solutions to enterprise problems, the SOMF ver 1.1 was released by Michael Bell [4][5].

This section clearly portrays that Zachman framework paves a way to build so many frameworks on it. The application of UML on RM-ODP derives a new framework. This analysis invokes why not to develop new frameworks by combining some existing technology to yield a better framework. The frameworks dealt in the next sections are most widely used for the commercial and Government departments. So, it is necessary to classify and compare them.

## III. Classification Of Frameworks

Classification is the problem of identifying which of a set of categories a new observation belongs to. As the frameworks were developed under the interests of different field masters, they were influenced by various perspectives. So, it is necessary to classify them as whether they are developed by standard bodies or individual interests or by private agencies.

The frameworks developed by standard bodies fall under the standard category and others fall under nonstandard category. And also they are subcategorized based on their usage in commercial or Government purpose.

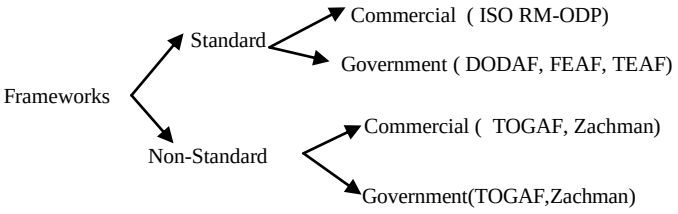



Figure 2. Classification

Frameworks developed and used for the Government departments and for Defense side applications are classified under the Government frameworks. Frameworks used for commercial purpose are classified under the commercial frameworks.

The Open Distributed model ISO RM-ODP falls under the standard and commercial frameworks. DODAF, FEAF and TEAF which were developed for the U.S Government agencies are coming under the standard and government frameworks.The well accepted and most widely used frameworks, TOGAF and Zachman frameworks are used by both the commercial and government agencies.

Even though TOGAF and Zachman frameworks are falling under non-standard category, mapping of these frameworks to DODAF, FEAF and other standard frameworks yielded good products in the industry. The classification described in this section will be very much useful for the customer to choose the suitable framework quickly for his organization based on the job nature also. The next subsection deals the comparison parameters that can be used by the customer to choose the appropriate tool. The following section analyses the well-known frameworks and lists out their criteria.

## IV. Evaluation Of Various Frameworks

In this paper, we have taken the survey of few frameworks which are most widely used. The parameters used for comparison in existing surveys are not suitable for a customer to choose the tool. So, the methodologies, techniques and tools used in these frameworks are considered for the comparison.

### A. Zachman Framework

The Zachman Framework describes the complex thing in different ways using different types of descriptions. It provides thirty-six categories to describe anything completely.

*1) Views / Viewpoints:* It has six different views to facilitate each player to view the system in their own particular way.

- Planner's View (Scope)
- Owner's View (Enterprise or Business Model)
- Designer's View (Information Systems Model)
- Builder's View (Technology Model)
- Subcontractor View (Detailed Specifications)
- Actual System View

*2) Domain:* It mainly focuses on Categorizing Deliverables [8].

*3) Origin:* This framework is well suited for Manufacturing Industries [8].

*4) Focus:* It focuses mainly on Business process.

*5) Phase of SDLC:* In the Design stage or planning stage, it can be used [8].

*6) System development methodology:* Organization's own methodology can be followed.

*7) System modeling Technique:* OMG-Model driven Architecture, Organization's own technique

*8) Business Modeling Technique:* BPML is used for this framework.

*9) Advantages :*

- Provides improved professional communications within community [22].
- Understanding the reasons for and risks of not developing any one architectural representation [22].
- Provides variety of tools and/or methodologies [22].
- Developing improved approaches [22].

*10) Weakness:*

- It may lead to more documentation depending on the cases [2]
- It may guide to a process-heavy approach to development [2].
- It isn't well accepted by all the developers [2].
- It seems in its first appearance as a top-down approach to developers. [2].
- It is to be biased towards traditional and data-centric techniques. [2].

### B. NATO Architecture Framework / C4ISR / DoDAF

The Department of Defense Architecture Framework (DoDAF) provides the organization of an enterprise architecture (EA) into consistent views. It is well suited for large complicated systems and interoperability challenges. "Operational views" used here are to deal with the external customer's operating domain.

*1) Views / Viewpoints:* DoDAF provides multiple views, each of which describes various aspects of the architecture. DoDAF defines the following views:

- Overarching All View (AV).
- Operational View (OV).
- Systems View (SV).
- Technical Standards View (TV).

*2) Domain:* It mainly focuses on operating domain [8].

*3) Origin:* This framework is developed for Defence [8].

*4) Focus:* It focuses mainly on Architecture data and Business process.

*5) Phase of SDLC:* It is used in a Process or Planning stage [8].

*6) System development methodology:* The Framework does not advice the use of any one methodology. It depends on the organization's decision.

*7) System modeling Technique:* If the system to be developed is larger, then UML tools are likely the best choice.

*8) Business Modeling Technique:* IDEF Family

*9) Advantages:*

- Defines a common approach for describing, presenting, and comparing DoD enterprise architectures [19].
- Common principles, assumptions and terminology are used [19].
- Across the organizational boundaries architecture descriptions can be compared [19].
- Deployment costs and reinvention of same system can be reduced. [9].

*10) Weakness:*

- No common ontology of architecture elements [1].
- Baseline (current) and objective (target) architectures are not addressed [1].
- How the architectures can be used to measure effectiveness is not dealt [1].
- Business-financial plans are not addressed. [1].

### *C. TOGAF*

The Open Group Architecture Framework (TOGAF) provides a comprehensive approach to the design, planning, implementation, and governance of enterprise information architecture.

*1) Views / viewpoints :* TOGAF identifies many views to be modeled in an architecture development process. The architecture views, and corresponding viewpoints come under the following categories:

- Business Architecture Views
- Information Systems Architecture views
- Technology Architecture views
- Composite views

*2) Domain:* It mainly focuses on Business, data and applications [8].

*3) Origin:* This framework is developed due to the motivation in Defence side framework.

*4) Focus:* It focuses mainly on Business process, Data, applications and Technology.

*5) Phase of SDLC:* It is used in a Process or Planning stage [8].

*6) System development methodology:* Rational Unified process (RUP) is used as a system development Methodology.

*7) System modeling Technique:* UML, BPMN are widely used in TOGAF system modeling.

*8) Business Modeling Technique:* IDEF is used for business modeling in TOGAF

*9) Advantages :*

- Increased transparency of accountability [24].
- Controlled risk [24].
- Protection of assets [24].
- Proactive control [24].
- Value creation [24].

*2) Weakness:*

- Lots of Detail [16].
- Planning methods and governance framework [15].
- Weak on Information Architecture [15].
- Can lead startup efforts into too much too soon [16].

### *D. TEAF*

Treasury Enterprise Architecture Framework (TEAF) was developed by the US Department of the Treasury and published in July 2000. It is based on the Zachman Framework.

The Treasury Enterprise Architecture Framework (TEAF) supports Treasury's business processes in terms of products. This framework guides the development and redesign of the business processes for various bureaus.

*1) Views / Viewpoints: It provides four different views.*

- Functional Views
- Information View
- Organizational View
- Infrastructure View

*2) Domain:* It has a domain on Business processes [20][8].

*3) Origin:* This framework is developed for Treasury department [20].

*4) Focus:* It focuses mainly on Business process.

*5) Phase of SDLC:* It is used in a communication or Planning stage [8].

*6) System development methodology:* It does not refer any specific methodology. It depends on the organization's decision.[23].

*7) System modeling Technique:* Flow chart, UML can be used in TEAF.

*8) Business Modeling Technique:* IDEF, ERD can be used as business modeling techniques.

*9) Advantages :*

- Provides the guidance to the treasury bureaus and offices in satisfying OMB and other federal requirements [20].

- Support Treasury bureaus and offices based on their individual priorities and strategic plans [20].
- Leads to Treasury-wide interoperability and reusability [20].

*10) Weakness:*

The TEAF does not contain a detailed description of how to generate the specification documents (work products) that are suggested for each cell of the TEAF Matrix [14].

### *E. FEAF*

Federal Enterprise Architecture (FEA) was developed for the Federal Government to provide a common methodology for information technology (IT) acquisition, use, and disposal in that Federal government. It was built to develop a common taxonomy and ontology to describe IT resources. The FEAF provides documenting architecture descriptions of high-priority areas. It guides to describe architectures for functional segments in multi-organization manner of the Federal Government.

*1) Views / Viewpoints:* Like zachman framework, FEAF is also having five different views in its framework.

- Planner's View (Scope)
- Owner's View (Enterprise or Business Model)
- Designer's View (Information Systems Model)
- Builder's View (Technology Model)
- Subcontractor's View (Detailed Specifications)

*2) Domain:* It has a domain on provision of services [8].

*3) Origin:* This framework is well suited for Enterprise Architecture planning.

*4) Focus:* It focuses mainly on Business process, Data, Application and Technology.

*5) Phase of SDLC:* It is used in a Communication or Planning stage [8].

*6) System development methodology:* RUP (Rational Unified process) is followed in FEAF.

*7) System modeling Technique:* UML is used as a system modeling tool in FEAF.

*8) Business Modeling Technique:* BPML is the technique used in FEAF.

*9) Advantages :*

- Serve customer needs better, faster, and cost effectively [18].
- Promote Federal interoperability [18].
- Promote Agency resource sharing [18].
- Reduced costs for Federal and Agency [18].
- Improve ability to share information [18].
- Supports capital IT investment planning in Federal and Agency [18].

*10) Weakness:*

- The Federal Government could risk allocating too much time and resources to an enterprise architecture description effort yielding potentially little return at significant cost [18].
- The Federal Enterprise Architecture program requires technical and acquisition expertise [18].
- The Federal IT community must keep its eyes on the basic principles rather than near-term objectives and achievements [18].
- The Federal Government has to pay up-front for the right to exercise options in the future [18].
- Concern over territoriality and loss of autonomy may impede the Federal Enterprise Architecture effort due to long-term, realignment of Agency functions and responsibilities [18].
- It is hard to have common, cross-Agency models and standards to ensure interoperability [18].

### *F. ISO RM-ODP*

The ISO Reference Model for Open Distributed Processing provides a framework standard to support the distributed processing in heterogeneous platforms. Object modeling approach is used to describe the systems in distributed environment.

*1) Views / Viewpoint:* The five viewpoints described by RM-ODP are:

- The enterprise viewpoint
- The information viewpoint.
- The computational viewpoint
- The engineering viewpoint
- The technology viewpoint

*2) Domain:* It has a domain on information sharing in distributed environment.

*3) Origin:* This framework is well suited for major computing and telecommunication companies.

*4) Focus:* It focuses mainly on Business process, Technical Functionality and Solution.

*5) Phase of SDLC:* It is used in a Processing and communication stage.

*6) System development methodology:* Object oriented method and IAD can be followed here [3].

*7) System modeling Technique:* UML, OMG (Model driven Architecture) are used as system modeling techniques [3].

*8) Business Modeling Technique:* BPMN is used as business modeling technique in RM-ODP.

*9) Advantages :*

- It provides lot of details for the analysis phases of the development of applications [3].
- It provides the platform to integrate the requirements from different languages consistently. [3].

It provides a set of established reasoning patterns to identify the fundamental entities of the system and the relations among them. It provides the appropriate degrees of abstraction and precision for building useful system specifications [3].

TABLE 1 COMPARATIVE CHART FOR FRAMEWORKS

| S/N | Frame work / TERMS | ZACHMAN FRAMEWORK http://zachmaninternational.com | DoDAF Cio-nii.defense.gov/docs/DoDAF_Volume_II.pdf | TOGAF http://www.opengroup.org/architecture/ | TEAF www.treas.gov/cio | FEAF www.cio.gov/documents/fedarch1.pdf | ISO RM-ODP http://www.rm-odp.net/ |
|---|---|---|---|---|---|---|---|
| 1 | **No of Views** | Six | Four | Four | Four | Five | Five |
| 2 | **Domain** | Categorizing deliverables | Operating domain | Business, Data and Applications | Business processes | Provision of services | information sharing in distributed environment |
| 3 | **Origin** | In- Manufacturing | Defence | Defence | Treasury Department | Enterprise Architecture planning | major computing and telecommunication companies |
| 4 | **Focus** | Business process | Architecture Data, Business process | Business process, Data, Applications, Technology | Business processes | Business process, Data, Applications& Technology | Business process, Technical functionality & Solution |
| 5 | **Phase of SDLC** | Planning (Design) | Process/Planning | Process/Planning | planning / communication | Planning & communication | Processing & communication |
| 6 | **System development methodology** | Organization' own methodology | Organization' own methodology | RUP | Organization' own methodology | RUP | Object oriented method, IAD |
| 7 | **System modeling technique** | OMG-Model driven Architecture, Organization's own technique | UML | UML, BPMN | Flow chart, UML | UML | UML, OMG(Model driven Architecture) |
| 8 | **Business model technique** | IDEF | IDEF Family | IDEF Family | IDEF, ERD | BPML | BPMN |
| 9 | **Advantages** | • Improving professional Communications<br>• wide variety of tools<br>• improved approaches For Architectural representations | • common Approach<br>• common principles, assumptions and terminology<br>• comparable architecture descriptions across organizational boundaries<br>• reduction of deployment costs | • increased transparency of accountability<br>• controlled risk<br>• protection of Assets<br>• proactive Control<br>• value creation | • satisfying OMB<br>• support individual<br>• Priorities and strategic Plans<br>• interoperability and reusability | • serve customer needs better, faster and cost effectively<br>• promote federal Interoperability<br>• provide agency resource sharing<br>• reduced costs<br>• improve ability to share information<br>• support Federal and agency capital IT investment planning | • improved requirement collection and analysis phases<br>• consistently integrated requirements expressed in separate languages<br>• set of already established reasoning patterns<br>• used for building robust, efficient and competitive applications<br>• backed by industrial products with enough acceptance |
| 10 | **Weakness** | • documentation heavy approach<br>• process heavy approach to development<br>• seems like Top down Approach<br>• biased towards traditional, data centric techniques | • No common ontology Of architecture elements<br>• not addressing baseline and objective architectures<br>• not addressing capabilities and measures of effectiveness<br>• lack of business financial artifacts | • lots of detail<br>• planning methods and governance framework<br>• weak on information Architecture<br>• can lead startup efforts into too much too soon | • No detailed description of Specification document for each cell<br>• Missing the techniques for creating specification document | • little return at significant cost<br>• need technical and acquisition expertise<br>• need a watch on future<br>• less future Maneuverability<br>• loss of autonomy may impede effort<br>• difficult to ensure interoperability | • problem of inter-view Consistency<br>• Not a truly guaranteed cross-view checks<br>• No precise notion of Consistency |
| 11 | **Tools** | • Adaptive EA Manager<br>• Mega V6.1<br>• SystemArchitect V10<br>• Simon Tool | • EA Webmodeler<br>• Corporate Modeler Enterprise Edition 10<br>• SystemArchitect V10<br>• Metis product family | • System Architect 10<br>• Metastorm ProVision EA V6.0<br>• IDS Scheer<br>• EA Webmodeler | • EA Webmodeler<br>• Corporate Modeler Enterprise Editionv10<br>• FEAMS V0.2<br>• Metis product family | • Adaptive EA Manager<br>• Flashline4<br>• FEAMS V0.2<br>• SystemArchitect V10 | • ConsVISor<br>• TINA<br>• Simon Tool<br>• MagicDraw |

- It provides a set of mechanisms and common services to build robust, efficient and competitive applications, interoperable with other systems [3].

*10) Weakness:*

RM-ODP has the problem of inter-view and inter-view consistency. A number of cross-view checks to be done to maintain the consistency. These checks don't guarantee the consistency [11].

## V. COMPARATIVE CHART OF FRAMEWORKS

Table 1 given above describes the comparison between all the discussed frameworks. It has precise data for the user with the additional information of available supportive tools.

## VI. CONCLUSION

This paper presents an overview of software architecture and reviewed the evolution of software architecture. By seeing the evolution tree, one can easily understand the evolution. Well known frameworks have been studied and discussed in detail in this paper. It summarizes the frameworks based on the industry side criteria and it discusses the benefits and drawbacks of each framework. The comparative chart included here clearly figures out the frameworks and it can be used as the reference guide also. It will invoke the user to choose the right framework for their industry, organization and business based on their requirement. Users can easily identify the supporting tools available for their frameworks. All the frameworks analyzed here are mainly focusing on business and IT solutions. In future ancient Indian architecture styles can be mapped to the familiar Frameworks to yield new frameworks to focus on quality.

## REFERENCES

[1] Alessio Mosto,: DoD Architecture Framework Overview [Online]. Available : www.enterprise-architecture.info/Images/.../DODAF.ppt (2004, May.)

[2] Ambler, S. Extending the RUP with the Zachman Framework.,http://www.enterpriseunifiedprocess.com/essays/ZachmanFramework.html (2007).

[3] Antonio Vallecillo,: RM-ODP: The ISO Reference Model for Open Distributed Processing. ETSI Informática, Universidad de Málaga , www.enterprise-architecture.info/Images/Documents/RM-ODP.pdf

[4] Bell, Michael,: "Introduction to Service-Oriented Modeling", in Service-Oriented Modeling: Service Analysis, Design, and Architecture. Wiley & Sons, (2008)

[5] Buckalew P M.: Service Oriented Architecture.,http://www.pmbuckalew .com/soa.htm (2009).

[6] Cris Kobryn and Chris Sibbald.,: Modeling DODAF Complaint Architectures., www.uml-forum.com/…/White_Paper_Modeling_ DoDAF_ UML2.pdf (2004, Oct. 25).

[7] HighBeam Research,:Software Framework., http://www.reference.com /browse/SoftwaRRre_framework (2008).

[8] Jaap schekkerman.: A comparative survey of Enterprise Architecture Frameworks. Institute for Enterprise Architecture Developments, Capgemini.,, www.enterprise-architecture.info (2003).

[9] Jim.,: Applicability of DODAF in Documenting Business Enterprise Architectures, http://www.thario.net/2008/08/applicability-of-dodaf-in-documenting.html (2008, Aug. 9).

[10] Juan Ignacio.; UML4ODP PLUGIN – User guide Version 0.9, Atenea Research Group, Spain., http://issuu.com/i72jamaj/docs/uml4odp_plugin (2009).

[11] Mark Maier and Eberhardt Rechtin, "Architecture Frameworks" in The art of systems architecting,2nd ed.CRC Press, Florida, pp. 229-250.(2000).

[12] Philippe Kruchten: "Architectural Blueprints – The "4+1" View model of software Architecture" IEEE Softw. Vol. 12, pp. 42-50, (1995)

[13] Roger Session: A Comparison of Top Four Enterprise –Architecture Methodologies, ObjectWatch, Inc., www.objectwatch.com/white_ papers.htm(2007, May)

[14] Susanne Leist, Gregor Zellner.,: Evaluation of Current Architecture Frameworks . University of Regensburg, Germany, www.dcc.uchile.cl/~vramiro/d/p1546-leist.pdf (2006).

[15] The Open Group.: Module 2 TOGAF9 Components. www.opengroup. org/togaf/ (2009).

[16] Tim Westbrock.,: Do Frameworks Really Matter? , EADirections, ww.eadirections.com/.../ EAdirections%20Frameworks%20Breakout% 20updated.pdf (2007, Oct. 24).

[17] Tony C Shan ,: "Taxonomy of Java Web Application Frameworks" in Conf. Rec. 2006 IEEE Int. Conf. e-Business Engg., pp. 378-385.

[18] U.S. Chief Information officers ( CIO) Council. :Federal Enterprise Architecture Framework Version 1.1 .,www.cio.gov/documents/ fedarch1.pdf (1999, Sep).

[19] U.S. Dept. of Defense.: DoD Architecture Framework Version 1.5. cio-nii.defense.gov/docs/DoDAF_Volume_II.pdf (2007, Apr. 23).

[20] U.S. Treasury Chief Information officer Council.: Treasury Enterprise Architecture Framework Version 1, www.treas.gov/cio (2000, Jul.).

[21] Will Conely: About Architecture . http://www.ehow.com/about_ 4565949_architecture.html (2009).

[22] Zachman, J. A.,: "A Framework for Information Systems Architecture". IBM Syst. J. Vol. 26, No. 3.pp. 276 – 292, Apr. (1987.)

[23] "Treasury Enterprise Architecture Framework" , en.wikipedia.org/.../ Treasury_Enterprise_Architecture_Framework

[24] "What is TOGAF?" ,http://www.articlebase.com/ information-technology-articles/what-is-togaf-626259.html

## AUTHORS PROFILE

Mrs. S. Roselin Mary is presently working as an Assistant professor in the department of Computer science and engineering, Anand Institute of Higher technology at Chennai. She obtained her B.E(CSE) degree from Madurai Kamaraj University and M.Tech(CSE) from Sastra University, Thanjavur. She is now pursuing her research in Hindustan University, chennai; India.She has 8 years of teaching experience.

Dr.Paul Rodrigues is a Chief Technology Officer, WSS at chennai. He has many years of experience in industries and teaching. He guided many Ph.d's and published more than 25 papers in various journals. He was formerly professor in Kalasalingam college of engineering, Tamilnadu and worked as Dean (IT & MCA )in Hindustan University.

# A hybrid Evolutionary Functional Link Artificial Neural Network for Data mining and Classification

Faissal MILI
Applied Economics and Simulation,
Faculty of Management and Economic Sciences,
5100 Mahdia, Tunisia

Manel HAMDI
International Finance Group Tunisia , Faculty of
Management and Economic Sciences of Tunis, Tunisia
2092, El Manar Tunisia.

***Abstract*— This paper presents a specific structure of neural network as the functional link artificial neural network (FLANN). This technique has been employed for classification tasks of data mining. In fact, there are a few studies that used this tool for solving classification problems. In this present research, we propose a hybrid FLANN (HFLANN) model, where the optimization process is performed using 3 known population based techniques such as genetic algorithms, particle swarm and differential evolution. This model will be empirically compared to FLANN based back-propagation algorithm and to others classifiers as decision tree, multilayer perceptron based back-propagation algorithm, radical basic function, support vector machine, and K-nearest Neighbor. Our results proved that the proposed model outperforms the other single model.**

***Keywords- component Data mining; Classification; Functional link artificial neural network; genetic algorithms; Particle swarm; Differential evolution.***

## I. INTRODUCTION

Classification task is a very important topic in data mining. A lot of research ([1], [2], [3]) has focused on the field over the last two decades. The Data mining is a knowledge discovery process from large databases. The extracted knowledge will be used by a human user for supporting a decision that is the ultimate goal of data mining. Therefore, classification decision is our aim in this study. A various classification models have been used in this regard. M. James [4] has employed a linear/quadratic discriminates techniques for solving classification problems. Another procedure has been applied using decision trees ([5], [6]). In the same context, Duda et al. [7] have proposed a discriminant analysis based on the Bayesian decision theory. Nevertheless, these traditional statistical models are built mainly on various linear assumptions that will be necessary satisfied. Otherwise, we cannot apply these techniques for classification tasks. To overcome the disadvantage, artificial intelligent tools have been emerged to solve data mining classification problems. For this purpose, genetic algorithms models were used [8]. In a recent research, Zhang ([9], [10]) have introduced the neural networks technique as a powerful classification tool. In these studies, he showed that neural network is a promising alternative tool compared to various conventional classification techniques. In a more recent literature, a specific structure of neural network has been employed for classification task of data mining as the functional link artificial neural network (FLANN). In fact, there are a few studies ([11], [12], [13]) used this tool for solving classification problems.

In this present research, we propose a hybrid FLANN (HFLANN) model based on three metaheuristics population based optimization tools such: genetic algorithms (GAs), particle swarm optimization (PSO) and differential evolution. This model will be compared to the trained FLANN based backpropagation and multilayer perceptron (MLP) as the most famous model in the area.

## II. CONCEPTS AND DEFINITION

### A. Population based algorithms

Population based algorithms are classed as a computational intelligence techniques representing a class of robust optimization ones. These population based ones make use of a population of solution in the same time based on natural evolution.

Many population based algorithms are presented in the literature such evolutionary programming [14], evolution strategy [15], genetic algorithms [16], genetic programming [17], Ant Colony [18], particle swarm [19] and differential evolution [20]. These algorithms differ in selection, offspring generation and replacement mechanisms. Genetic algorithms, particle swarm and differential evolutions represent the most popular ones.

#### 1)Genetic algorithms

Genetic algorithms (GAs) are defined as a search technique that was inspired from Darwinian Theory. The idea is based on the theory of natural selection. We assume that there is a population composed with different characteristics. The stronger will be able to survive and they pass their characteristics to their offsprings.

The total process is described as follows:

1- Generate randomly an initial population;

2- Evaluate this population using the fitness function;

3- Apply genetic operators such selection, crossover, and mutation;

4- Turn the process "Evaluation Crossover mutation" until reaching the stopped criteria fixed in prior.

*2)Particle swarm*

Presented in 1995 by L. Kennedy and R. Eberhart [19], particle swarm optimization (PSO) represents one of the most known population-based approaches, where particles change their positions with time. These particles fly around in a multidimensional search space, and each particle adjusts its position according to its own experience and the experience of their neighboring, making use of the best position encountered by itself and its neighbors. The direction of a particle is defined by the set of neighboring and its correspondent history of experience.

An individual particle i is composed of three vectors:

- Its position in the V-dimensional search space

$\overrightarrow{X_i} = (X_{i1}, X_{i2}, \dots, X_{iV})$

- The best position that it has individually found $\overrightarrow{P_i} = (P_{i1}, P_{i2}, \dots, P_{iV})$

- Its velocity $\overrightarrow{V_i} = (V_{i1}, V_{i2}, \dots, V_{iV})$

Particles were originally initialized in a uniform random manner throughout the search space; velocity is also randomly initialized.

These particles then move throughout the search space by a fairly simple set of update equations. The algorithm updates the entire swarm at each time step by updating the velocity and position of each particle in every dimension by the following rules:

$$. V_{iD} = \chi * (W * V_{iN} + C * \varepsilon_1 (P_{nV} - X_{iV}) + C * \varepsilon_2 (P_{gV} - X_{iV})) \quad (1)$$

$$X_{iN} = X_{iV} + V_{iV} \quad (2)$$

Where in the original equations:

C is a constant with the value of 2.0

$\varepsilon_1$ and $\varepsilon_2$ are independent random numbers uniquely generated at every update for each individual dimension (n = 1 to V).

$P_{gV}$ is the best position found by the global population of particle.

$P_{nV}$ is the best position found by any neighbor of the particle.

W: the weight

$\chi$ : the constriction factor.

*3)Differential evolution*

Proposed by Storn and Price in 1995 [20], differential evolution represents a new floating evolutionary algorithm using a special kind of differential operator. Easy implementation and negligible parameter tuning makes this algorithm quite popular.

Like any evolutionary algorithm, differential evolution starts with a population. Differential evolution is a small and simple mathematical model of a big and naturally complex process of evolution. So, it is easy and efficient.

Firstly, there are five DE strategies (or schemes) that were proposed by R. Storn and K. Price [20]:

• **Scheme DE/rand/1 :**

$$\omega = x_1 + F^* (x_2 - x_3) \quad (3)$$

• **Scheme DE/rand/2 :**

$$\omega = x_5 + F * (x_1 + x_2 - x_3 - x_4) \quad (4)$$

• **Scheme DE/best/1:**

$$\omega = x_{best} + F * (x_1 - x_2) \quad (5)$$

• **Scheme DE/best/2:**

$$\omega = x_{best} + F * (x_1 + x_2 - x_3 - x_4) \quad (6)$$

• **Scheme DE/rand-to best/1:**

$$\omega = x + \lambda^* (x_{best} - x_1) + F * (x_2 - x_3) \quad (7)$$

Later, two more strategies were introduced [21].

We present the trigonometric scheme defined by:

$$\omega = (x_1 + x_2 + x_3)/3 + (p_2 - p_1) * (x_1 - x_2) + (p_3 - p_2) * (x_2 - x_3) + (p_1 - p_3) * (x_3 - x_1) \quad (8)$$

$$p_i = |f(x_i)/ (f(x_1) + f(x_2) + f(x_3))|, i = 1, 2, 3 ; \quad (9)$$

F define the constriction factor generally taken equal to 0.5

x define the selected element

$x_1$, $x_2$, $x_3$, $x_4$ and $x_5$ represent random generated elements from the population.

Many others schemes can be found in the literature [20].

*B. Functional Link Artificial Neural Networks*

The FLANN architecture was originally proposed by Pao et al. [22]. The basic idea of this model is to apply an expansion function which increases the input vector dimensionality. We say that the hyper-planes generated provide greater discrimination capability in the input pattern space. By applying this expansion, we needn't the use of the hidden layer, making the learning algorithm simpler. Thus, compared to the MLP structure, this model has the advantage to have faster convergence rate and lesser computational cost.

The conventional nonlinear functional expansions which can be employed are trigonometric, power series or Chebyshev type. R. Majhi et al. [23], shows that use of trigonometric expansion provides better prediction capability of the model. Hence, in the present case, trigonometric expansion is employed.

Let each element of the input pattern before expansion be represented as X(i), $1 < i < I$ where each element x(i) is functionally expanded as Zn(i) , $1 < n < N$ , where N = number of expanded points for each input element. In this study, we take N=5.

I= the total number of features

As presented in figure 1, the expansion of each input pattern is done as follows.

$Z_1(i) = X(i), Z_2(i) = f_1(X(i)),...., Z_5(i) = f_5(X(i))$ (10)

These expanded inputs are then fed to the single layer neural network and the network is trained to obtain the desired output.

## III. HYBRID FLANN DESCRIPTION

The proposed hybrid FLANN is based on evolutionary algorithms as genetic algorithms, particle swarm and differential evolution.

### A. Resampling technique:

In order to avoid overfitting, we use the (2*5) K fold cross-validation resampling technique. We proceed as follows:

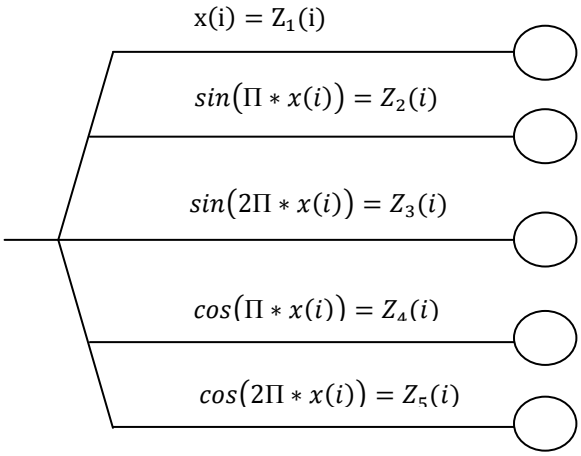



Figure 1. Functional expansion of the first element

We divide initial database into 5 folds (K=5) where each one contain the same repartition of classes. For example, if initial population contains 60% of class 1 and 40% of class 2, then all the resulted K folds must have the same repartition.

### B. Generation

We begin the process by generating randomly initial solution. We execute partial training using differential evolution in order to improve initial state.

### C. Fitness fuction and evaluation

In order to evaluate each solution, two criterions are used such the mean square error (MSE) and the misclassification error (MCE) rate. If we have to compare solutions A and B, we apply the following rules: A is preferred to B If and only if MCE(A)< MCE(B) Or MCE(A)= MCE(B) and MSE(A)< MSE(B).

### D. Selection

Many selections are defined in the literature such the Roulette wheel method, the N/2 elitist method and the tournament selection method. The last method will be used here. The principle is to compare two solutions, and the best one will be selected.

N/2 elitist is used at the beginning of the process in order to select 50% of generated solution.

### E. Crossover

Two parents are selected randomly in order to exchange their information. Two crossovers are applied and described as follows:

#### 1) Crossover 1 (over input feature):

An input feature is chosen randomly to exchange his correspondent weight between the selected two parents.

#### 2) Crossover 2 (over output nodes):

An output is chosen randomly to exchange his correspondent weight.

#### 3) Crossover 3 (Crossover over connection):

A connection position is chosen randomly and his correspondent weight is exchanged between the two parents.

### F. Mutation

#### 1) Mutation 1(over connection)

A connection position is chosen randomly and his correspondent weight has been controlled. If this connection is connected, his correspondent weight is disconnected by setting his value equal to zero. Else, this connection is connected.

#### 2) Mutation 2 (over one input feature)

An input feature is chosen randomly and his correspondent weights have been controlled. If this input feature is connected (there is at least one weights of his correspondent ones is different from zero), it will be disconnected by putting all his entire weight equal to zero. Else if this input feature is totally disconnected, it will be connected there by generating weights different from zero.

#### 3) Mutation 3 (over two input feature)

We do the same like mutation 2 but here simultaneously for the two selected features.

#### 4) Mutation 4 ( over three input feature)

In this mutation, the same principle is used for three input features.

We note that many input features connection and disconnection can be executed in the same time when having a large number of features. This crossover helps to remove undesirable features from our classification process and can improve the final performance process.

### G. Particle swarm optimization (PSO)

In the presented paper, we define three PSO model based on the notion of neighbor.

#### 1) PSO based on resulted genetic offsprings

First, we apply genetic operators. Each offspring that improve our fitness function define a neighbor, and used in equation (1).

#### 2) PSO based on Euclidian distance:

For each particle, we compute the Euclidian distance between this particle and the rest of the population. Next we choose the five nearest particles based on this distance.

From the selected subset of neighbors, we choose the best one which has the best fitness value. This selected one defines our neighbor to be replaced in equation (1).

*3) PSO based on the last best visited solution:*

In this case, each particle flies and memorizes his best reached solution. This memory defines the neighbor to be used in equation (1).

*H. Differential evolution*

In this work, we proceed as follows:

- First, for each candidate x, we generate five random solution $x_1$, $x_2$, $x_3$, $x_4$ and $x_5$.

- Next we apply seven chosen schemes as follows:

DE1: Scheme DE/direct/1 :

$$\omega = x + F^* (x_2 - x_1) \quad (11)$$

DE2: Scheme DE/best/1 :

$$\omega = x_{best} + F^* (x_2 - x_1) \quad (12)$$

DE3: Scheme DE/best/1 :

$$\omega = x_{best} + F^* (x_3 - x_2) \quad (13)$$

DE4: Scheme DE/best/1 :

$$\omega = x_{best} + F^* (x_3 - x_1) \quad (14)$$

DE5: Scheme DE/best/2 :

$$\omega = x_{best} + F * (x1 + x2 - x3 - x4) \quad (15)$$

DE6: Scheme DE/rand/2 :

$$\omega = x5 + F * (x1 + x2 - x3 - x4) \quad (16)$$

DE7: with Trigonometric Mutation:

$$\omega = (x1 + x2 + x3)/3 + (p2 - p1) * (x1 - x2) + (p3 - p2) * (x2 - x3) + (p1 - p3) * (x3 - x1) \quad (17)$$

$$p_i = |f(x_i)/ (f(x1) + f(x2) + f(x3)) |, i= 1, 2, 3 ; \quad (18)$$

*I. Stopping criterion:*

The process turns in a cycle until reaching a maximum number of epochs without any improvement. We fix the maximum number of epochs equal to 30 epochs.

## IV. EXPERIMENTAL STUDIES:

11 real-world databases were selected there to be used in simulation works. They are chosen from the UCI repository machine learning, which is commonly used to benchmark learning algorithms [24].

We compare the results of the proposed hybrid FLANN (HFLANN) with FLANN based on the gradient descent algorithm. Next, Comparison with other classifiers will be done.

*A. Description of the databases*

A brief description of used databases for experimental setup is presented in table I. Num. is the numeric features, Bin. is the binary ones, and Nom. is the nominal inputs that mean discrete with three or more distinct labels.

TABLE I. SUMMARY OF THE DATASET USED IN SIMULATION STUDIES

| Dataset | Inputs | | | | Ex. | Cls |
|---|---|---|---|---|---|---|
| | Num. | Bin. | Nom. | Total | | |
| **IRIS** | 4 | 0 | 0 | 4 | 150 | 3 |
| **VOTING** | 0 | 16 | 0 | 16 | 435 | 2 |
| **BREAST** | 0 | 0 | 9 | 9 | 699 | 2 |
| **PRIMA** | 8 | 0 | 0 | 8 | 768 | 2 |
| **CREDIT** | 6 | 4 | 4 | 14 | 690 | 2 |
| **BALANCE** | 4 | 0 | 0 | 4 | 625 | 3 |
| **WINE** | 13 | 0 | 0 | 13 | 178 | 3 |
| **BUPA** | 6 | 0 | 0 | 6 | 345 | 2 |
| **ECOLI** | 7 | 0 | 0 | 7 | 336 | 8 |
| **GLASS** | 10 | 0 | 0 | 10 | 214 | 6 |
| **ZOO** | 1 | 15 | 0 | 16 | 101 | 7 |

*B. Convergence test:*

In order to test the convergence of the proposed hybrid FLANN, a comparison will be done with trained FLANN using the back-propagation algorithm. Results are presented in figure 2 and figure 3. Comparison is done based on the required time and number of epochs for convergence.

From figure 2, we find that our process needs less than 200 seconds 20 epochs to converge. Figure 3 present results for FLANN based on back-propagation. This model requires less than 150 seconds and 15 epochs to converge.

The proposed hybrid FLANN has a strong ability to converge fast and requires approximately the same time and epochs than FLANN based back-propagation.

*C. Comparative results:*

The classification accuracy of the proposed hybrid FLANN are compared with the results obtained from FLANN trained using the back-propagation algorithm. Results are presented in table II. Bold typeface is used to highlight the results that are significantly better.

We find that the proposed model gives better results for all used databases. So, our proposed evolutionary process trains better than the back-propagation algorithm.

Next, a basic comparison between the HFLANN and five others classifiers is performed using fourth databases. These classifiers are:

- The decision tree based C.45,

- The multilayer perceptron (MLP) based back-propagation algorithms,

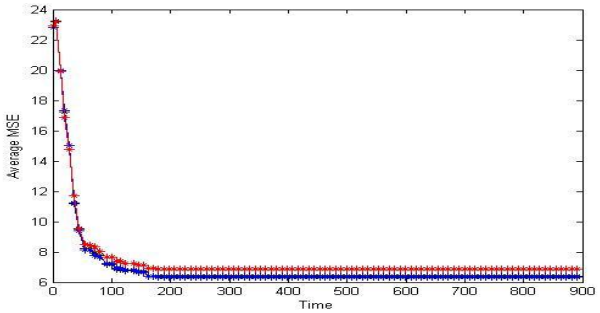
- The radical basic function (RBF),

- The support vector machine (SVM),

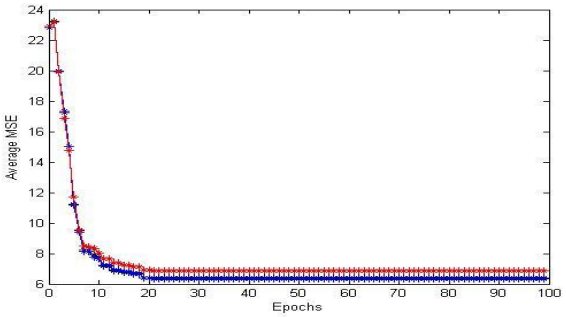
- The K-nearest Neighbor (KNN),

Results are presented in table II. From this table, we find that HFLANN is better for 3 databases and SVM is better for the forth. We say that HFLANN is a good classifier giving better results in the majority of used databases.

Table IV presents the number of different local and global improvement of used population based algorithms. We find that PSO represents the best local population based technique with 104476 improvements, and differential evolution is the best global one with 1854 of improvement.

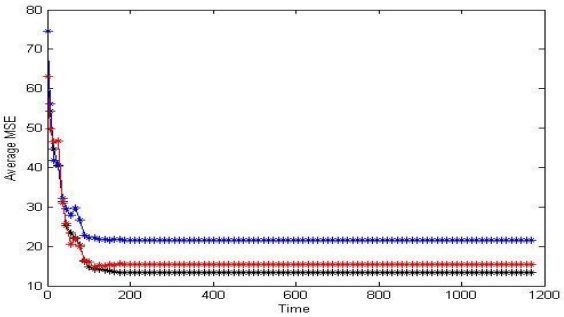



*a.*MSE vs Time

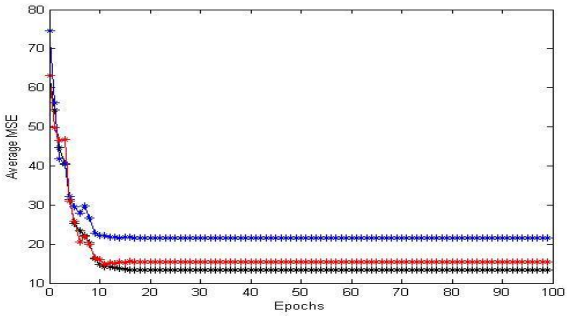



b.MSE vs epochs

Figure 2. The MSE Hybrid FLANN results vs. time and epochs applied to the iris database




a. MSE vs Time




b. MSE vs epochs

Figure 3. The MSE FLANN based back-propagation results vs. time and epochs applied to the iris database

TABLE II. AVERAGE COMPARATIVE PERFORMANCE OF HFLANN AND FLANN WITH A CONFIDENCE INTERVAL LEVEL OF 5%

| Data bases | Folds | HFLANN | | FLANN BASED BP | |
|---|---|---|---|---|---|
| **IRIS** | ***training*** | **0,9833** | ±0,0054 | 0,8958 | ±0,0351 |
| | ***validation*** | **0,9600** | ±0,0303 | 0,9000 | ±0,0228 |
| | ***test*** | **0,9600** | ±0,0288 | 0,8933 | ±0,0522 |
| | ***Time*** | 184,3925 | | 142,6253 | |
| **VOTING** | ***training*** | **0,9787** | ±0,0057 | 0,7796 | ±0,0695 |
| | ***validation*** | **0,9654** | ±0,0227 | 0,7990 | ±0,0765 |
| | ***test*** | **0,9469** | ±0,0190 | 0,7829 | ±0,0781 |
| | ***Time*** | 292,5714 | | 78,4182 | |
| **BREAST** | ***training*** | **0,9787** | ±0,0022 | 0,9277 | ±0,0192 |
| | ***validation*** | **0,9699** | ±0,0101 | 0,9342 | ±0,0192 |
| | ***test*** | **0,9527** | ±0,0139 | 0,9298 | ±0,0179 |
| | ***Time*** | 211,6272 | | 156,4693 | |
| **PRIMA** | ***training*** | **0,7906** | ±0,0061 | 0,6865 | ±0,0130 |
| | ***validation*** | **0,7773** | ±0,0170 | 0,7161 | ±0,0153 |
| | ***test*** | **0,7501** | ±0,0255 | 0,6536 | ±0,0213 |
| | ***Time*** | 161,3451 | | 133,8419 | |
| **CREDIT** | ***training*** | **0,8926** | ±0,0056 | 0,5830 | ±0,0540 |
| | ***validation*** | **0,8771** | ±0,0264 | 0,6021 | ±0,0444 |
| | ***test*** | **0,8615** | ±0,0284 | 0,5935 | ±0,0817 |
| | ***Time*** | 355,2951 | | 100,7891 | |
| **BALANCE** | ***training*** | **0,9212** | ±0,0036 | 0,6123 | ±0,0157 |
| | ***validation*** | **90,7400** | ±0,0137 | 0,6454 | ±0,0383 |
| | ***test*** | **0,9101** | ±0,0104 | 0,6036 | ±0,0367 |
| | ***Time*** | 314,7218 | | 168,9399 | |
| **WINE** | ***training*** | **0,9972** | ±0,0036 | 0,9244 | ±0,0549 |
| | ***validation*** | **0,9833** | ±0,0174 | 0,9379 | ±0,0719 |
| | ***test*** | **0,9508** | ±0,0329 | 0,9035 | ±0,0399 |
| | ***Time*** | 173,4038 | | 61,0904 | |
| **BUPA** | ***training*** | **0,7666** | ±0,0175 | 0,5630 | ±0,0249 |
| | ***validation*** | **0,7147** | ±0,0395 | 0,6202 | ±0,0276 |
| | ***test*** | **0,7027** | ±0,0199 | 0,5392 | ±0,0506 |
| | ***Time*** | 232,4010 | | 42,2137 | |
| **ECOLI** | ***training*** | **0,8077** | ±0,0157 | 0,6197 | ±0,0559 |

| | | | | | |
|---|---|---|---|---|---|
| | ***validation*** | **0,7703** | ±0,0342 | 0,6282 | ±0,0630 |
| | ***test*** | **0,7889** | ±0,0221 | 0,6279 | ±0,0811 |
| | ***Time*** | 315,8144 | | 255,8217 | |
| **GLASS** | ***training*** | **0,7103** | ±0,0181 | 0,3792 | ±0,0419 |
| | ***validation*** | **0,6901** | ±0,0342 | 0,4328 | ±0,0450 |
| | ***Test*** | **0,6054** | ±0,0464 | 0,3463 | ±0,0577 |
| | ***Time*** | 508,9287 | | 174,2889 | |
| ZOO | ***training*** | **0,8935** | ±0,0310 | 0,4683 | ±0,0789 |
| | ***validation*** | **0,8977** | ±0,0406 | 0,5005 | ±0,1055 |
| | ***test*** | **0,8322** | ±0,0422 | 0,4163 | ±0,0757 |
| | ***Time*** | 193,7053 | | 53,7702 | |

TABLE III. SUMMARY OF THE RESULTS COMPARING THE PERFORMANCE THE HFLANN AGAINST EXISTING WORKS. THE TABLE SHOWS THE REPORTED MEAN CLASSIFICATION ACCURACY OF THE VARIOUS WORKS AND THE BEST RESULTS ARE HIGHLIGHTED IN BOLD

| Evolutionary algorithm | Best Local improvement | Best Global improvement |
|---|---|---|
| **GA** | Crossover | Crossover |
| **DE** | trigonometric mutation scheme | DE/best/1 |
| **PSO** | PSO based genetic algorithms | PSO based Euclidian distance |

TABLE IV. NUMBER OF LOCAL IMPROVEMENT AND GLOBAL IMPROVEMENT OF GENETIC ALGORITHMS, DIFFERENTIAL EVOLUTION AND PARTICLE SWARM

| Methods/ references | Data set | | | |
|---|---|---|---|---|
| | **Cancer** | **Pima** | **Iris** | **Balance** |
| HFLANN | 0,9527 | **0,7501** | **0,960**0 | **0,9101** |
| C4.5 | 0,947 | 0,7313 | 0,9400 | 0,783 |
| MLP | 0,92 | 0,7330 | 0,9453 | 0,853 |
| RBF | - | 076 | 0,3850 | - |
| SVM | **95,43** | 0,6780 | - | 0,876 |
| K-NN | - | 0,6760 | - | 0,876 |

TABLE V. BEST LOCAL IMPROVEMENT AND BEST GLOBAL IMPROVEMENT OF GENETIC ALGORITHMS, DIFFERENTIAL EVOLUTION AND PSO

| Evolutionary algorithm | Local improvement | Global improvement |
|---|---|---|
| **GA** | 28423 | 662 |
| **DE** | 26148 | 1854 |
| **PSO1** | 104476 | 1246 |

Table V presents best local improvement and best global improvement of genetic algorithms, differential evolution and particle swarm. For differential evolution, trigonometric mutation scheme represent the best local search strategy and DE/best/1 scheme is the best global one. For PSO, basic model based on genetic algorithms represent the best local search strategy, and the PSO based Euclidian distance is the best global one. Comparing genetic operators, we find that crossovers improve results more than mutation.

## V. CONCLUSION

In this paper, a HFLANN was proposed based on three populations based algorithms such genetic algorithms, differential evolution and particle swarm. This classifier shows his ability to converge faster and gives better performance than FLANN based on back-propagation.

When comparing different population based algorithms, we find that PSO is the best local technique for improvement and differential evolution is the best global one. For differential evolution, trigonometric mutation scheme represent the best local search strategy and DE/best/1 scheme is the best global.

For PSO, basic model based on genetic algorithms represent the best local search strategy, and the PSO based Euclidian distance is the best global. Comparing genetic operators, we find that crossovers improve results more than mutation. Following this comparison, we are able to identify best local strategy and best global strategy. Compared to the MLP, FLANN has the advantage to optimize the process without using hidden nodes.

Future works can be addressed to compare other classifiers and others evolutionary algorithms. Others comparison criteria can be used such the needed speed and the robustness of the algorithm. A wrapper approach can be included in the proposed process in order to delete simultaneously irrelevant features over the optimization process.

## REFERENCES

[1] R .Agrawal, T. Imielinski and A. Swami, "Database mining: A performance perspective". IEEE Trans. Knowledge Data Eng., 5, pp. 914-925 (1993).

[2] U. M. Fayyad, Piatetsky-Shapiro, G., and Smyth, P. ,"From data mining to knowledge discovery: An overview". In U. M. Fayyad, G. Piatetsky-Shapiro, & P. Smyth (Eds.), Advances in knowledge discovery and data mining ,pp. 1–34,. Menlo Park, CA: AAAI Press ,1996.

[3] H.P. Kriegel, et al., "Future trends in data mining". Data Mining and Knowledge Discovery, 15(1), 87–97. Netherlands: Springer (2007).

[4] M., James, "Classification Algorithms". Wiley,1985.

[5] L. Breiman, , J.H. Friedman, R.A. Olshen and C.J. Stone, "Classification and Regression Learning". Morgan Kaufman, 1984.

[6] H.P. Kriegel, et al., "Future trends in data mining". Data Mining and Knowledge Discovery, 15(1), 87–97. Netherlands: Springer (2007).

[7] R.O. Duda, Hart PE and Stork D.G., "Pattern classification". Wiley, New York ,2001.

[8] D.E. Goldberg, "Genetic algorithms in search, optimization and machine learning". Morgan Kaufmann ,1989.

[9] G. P., Zhang, "Neural networks for classification", A survey. IEEE Transactions on Systems, Man, Cybernetics-Part C: Application and Reviews, 30(4), pp. 451–461, 2000.

[10] G. P Zhang, "Avoiding pitfalls in neural network research". IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, 37(1), pp. 3–16 , 2007. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[11] R. Majhi, G. Panda and G. Sahoo, "Development and performance evaluation of FLANN based model for forecasting of stock markets", Expert Systems with Applications 36, pp. 6800–6808,2009.

[12] S. Dehuri and Cho S.B. ,"A hybrid genetic based functional link artificial neural network with a statistical comparison of classifiers over multiple datasets". Neural Comput & Applic (19), pp.317-328. ,2010a.

[13] S. Dehuri and Cho S.B.,"Evolutionarily optimized features in functional link neural network for classification". Expert Systems with Applications (37), pp.4379–4391, 2010b .

[14] L., Fogel, J. Owens and J. Walsh, "Artificial Intelligence through Simulated Evolution". John Wiley, Chichester,1966.

[15] D.E. Goldberg, "Genetic algorithms in search, optimization and machine learning". Morgan Kaufmann ,1989.

[16] J., Holland, "Adaptation in natural and artificial systems". Univ. of Michigan Press, Ann Arbor ,1975.

[17] J. Koza, , "Genetic programming on the programming of computers by means of natural selection", Cambridge MA: MIT Press, 1992.

[18] M. Dorigo, , Stutzle, T.,"Ant Colony Optimization". MIT Press, Cambridge. ,2004.

[19] J., Kennedy Eberhart R., "Particle Swarm Optimization", In Proceedings of IEEE International Conference on Neural Networks, pp. 1942–1948,1995.

[20] R. Storn Kenneth P. "Differential evolution A simple and efficient adaptive scheme for global optimization over continuous spaces". Technical Report TR, pp. 95-012, International Computer Science Institute, Berkeley, CA, 1995.

[21] F. Hui-Y. and J. Lampinen,"A trigonometric mutation approach to differential evolution". In K. C. Giannakoglou, D. T. Tsahalis, J. Periaux, K. D. Papailiou, & T. Fogarty, editors, Evolutionary Methods for Design Optimization and Control with Applications to Industrial Problems, pp. 65–70, Athens, Greece. International Center for Numerical Methods in Engineering (Cmine) , 2001.

[22] Y.-H. Pao, S.M. Phillips and D.J. Sobajic, "Neural-net computing and intelligent control systems". Int. J. Contr., 56, pp. 263-289 , 1992.

[23] R. Majhi, G. Panda and G. Sahoo, "Development and performance evaluation of FLANN based model for forecasting of stock markets", Expert Systems with Applications 36, pp. 6800–6808,2009.

[24] C.L. Blake,. and Merz, C.J., 'UCI Repository of machine learning databases', em Irvine, CA: University of California, department of information and Computer Science. Available on-line at : http//www.ics.uci.edu/~mlearn/ML.Repository.html. , 1998.

## AUTHORS PROFILE

**Faissal MILI ,** Ph. Doctor in quantitative methods, A member of Applied Economics and Simulation laboratory, Faculty of Management and Economic Sciences, 5100 Mahdia, Tunisia**.** E-Mail: milisoft@yahoo.fr**.** Phone: 0021620545728

**Manel HAMDI,** Doctoral student in finance, Member of International Finance Group Tunisia, Faculty of Management and Economic Sciences of Tunis, El Manar University, Tunis cedex, C. P. 2092, El Manar Tunisia. E-Mail: mannelhamdi@yahoo.fr. Phone: 0021627220884

# Automatic Aircraft Target Recognition by ISAR Image Processing based on Neural Classifier

F. Benedetto, IEEE, Member, F. Riganti Fulginei, A. Laudani, IEEE, Member, G. Albanese
Dept. of Applied Electronics
University of ROMA TRE, via della Vasca Navale 84, 00146 Rome, Italy
Corresponding author's

***Abstract*— This work proposes a new automatic target classifier, based on a combined neural networks' system, by ISAR image processing. The novelty introduced in our work is twofold. We first present a novel automatic classification procedure, and then we discuss an improved multimedia processing of ISAR images for automatic object detection. The classifier, composed by a combination of 20 feed-forward artificial neural networks, is used to recognize aircraft targets extracted from ISAR images. A multimedia processing by two recently introduced image processing techniques is exploited to improve the shape and features extraction process. Performance analysis is carried out in comparison with conventional multimedia techniques and standard detectors. Numerical results obtained from wide simulation trials evidence the efficiency of the proposed method for the application to automatic aircraft target recognition.**

***Keywords- Automatic target recognition; artificial intelligence; neural classifiers; ISAR image processing; shape extraction.***

## I. INTRODUCTION

Recently, there has been an explosive growth in the research area related to inverse synthetic aperture radar (ISAR) imaging of moving targets [1]. High resolution images of targets of interest at long range can be acquired from ISAR images. Moreover, ISAR imaging is becoming an irreplaceable tool in the task of non-cooperative automatic target recognition (ATR). There are a lot of different applications, including detection, imaging, and classification of ships and aircraft with airborne, maritime, and land-based radar systems [2], [3]. In the last years, many methods of ISAR ATR techniques have been developed in the open literature. Before detecting an object, the image is first segmented and then the object is recognized [4]-[5]. Image segmentation is the process of partitioning/subdividing a digital image into multiple meaningful regions. The segmentation is usually based on measurements taken from the image and might be gray level, color, texture, depth or motion. The result of image segmentation is a set of segments that collectively cover the entire image. All the pixels of the same ensemble or region are similar with respect to some characteristic or computed property, such as color, intensity, or texture. Edge detection is one of the frequently used techniques in digital image processing. Object recognition is the task of finding a given object in an image or in a video sequence. For any object in an image, there are many features characterizing the object that can be extracted to provide a feature description of the object. This description, extracted from a training image, can then be used to identify the object when attempting to locate it in a test image containing many other objects [6]. Image segmentation is usually done using various edge detection techniques such as Sobel, Prewitt, Roberts, Canny, and other methods [7]. Then, only some features characterizing the ISAR images are tested, to identify what kind of target has been detected [8]-[11]. In fact, the typical algorithms first detect the edge of an ISAR image, and then adopt different 1-D descriptors such as Fourier descriptors (FD) [12] or 2-D descriptors such as Fourier-wavelet descriptors [13] for feature extraction. This is computationally more efficient than evaluating the whole target. Other methods exploit optimal classifiers to determine the specific kind of target, [14]-[16]. However, in all these techniques, each target profile is presented as an input feature vector to the classifier. Since providing real-time performance in radar target recognition is a crucial issue to be satisfied, usually neural networks with massive parallel structure and capacity of learning are used in the classifier [17].

The recognition process must be invariant with respect to the target position. At least three different techniques for invariant neural network recognition have been recently proposed. The first approach, namely the invariance by training, compensates for the pattern shift taking into account different targets for different pattern shifts during the training phase. The main drawback of such an approach is that it is inapplicable in many operating situations. In fact, the number of possible variations of patterns makes the training set too large, increasing at the same time the computational complexity of the learning system. The second technique, namely invariance by structure, uses neural networks whose outputs are always invariant to certain transformations. The disadvantage of such an approach is that high-order neural networks are required. Their implementation is too complicated and their application is limited. Recently, the combined neural network method is approaching as the most suitable scheme, to lower the system computational complexity (see [16] and references therein). Finally, the third technique uses feature vectors as inputs for the neural networks. These feature vectors are invariant to the required transformations and, hence, this method is called invariant feature space. This kind of recognition system usually uses the magnitude of the Fourier transform, which is invariant to linear shifts of its input vector [23]. The advantages of using

an invariant feature space are as follows: the number of features can be reduced to realistic levels, the requirements on the classifier are relaxed, and the invariance for all input objects is ensured. On the other hand, the main drawback of using invariant feature spaces is the processing time (it can be very long) needed to extract the features before the classifier can be employed.

The novelty introduced in this work is twofold. We first present a novel automatic classification procedure, based on combined neural networks' signal processing. Then, we discuss an improved multimedia processing of ISAR images for automatic object detection. In particular, the neural classifier (NC) is developed as an alternative approach to those existing in the literature (e.g. the Support Vector Machine based algorithms are widely used for the patterns recognition and classification). Designers and users will be then able to choose the different approaches depending both on the nature of the problem to be solved and on the used technology. In our case the NC is composed by combining 20 feed-forward artificial Neural Networks (NNs). Nevertheless, the number of NNs can be changed to obtain several different performances depending on the difficulty of classification problem. Moreover, it is well known that the structure of a neural network is fixed on the base of the problem to be solved and the available data. Furthermore, it's clear that a deterministic way to define the number of hidden layers and the number of neurons does not exist. In our case, after performing of several experimental results, the NNs have been all made by one input layer and two hidden layers made of 168 and 8 neurons, respectively. Then, the output of each NN consists of one neuron that returns a value characterizing the class of the related input pattern (Fourier descriptors of the ISAR image to classify). The ISAR images are first pre-processed with conventional filters, in order to reduce the speckle noise. Then, the combination of two image processing techniques, recently introduced in literature, is exploited to improve the shape and features extraction process. First, the Smallest Univalue Segment Assimilating Nucleus (SUSAN) algorithm [18] is applied to the ISAR image. Then, the output of the SUSAN method is processed by a modified level set evolution method, namely distance regularized level set evolution (DRLSE) [19]. We use the first method (i.e. SUSAN) as a pre-processing step, in order to segment the input image into two regions of pixels containing the ensemble of the target pixels and the ensemble of the background pixels (i.e. pixels not belonging to the target). Then, the DRLSE algorithm is applied to the first ensemble (i.e. the target pixels region) as a linking edge method, whose output is the target contour. Once the aircraft shape is extracted, the invariant Fourier descriptors (FD) are computed and used as the input of the neural classifier.

The remainder of this paper is organized as follows. In Section II, the proposed neural networks classifier is described, while the conventional multimedia processing is illustrated in Section III. Section IV presents the proposed ISAR images segmentation and shape extraction techniques. Numerical results and comparisons are outlined in Section V, and finally, our conclusions are depicted in Section VI

## II. Neural Classifier for Object Detection

In this Section, we discuss the structure of the proposed ATR scheme composed by a system of 20 feed-forward artificial Neural Networks (NNs) [20]. The recognition process must be invariant with respect to the target position. At least three different techniques for invariant neural network recognition have been recently proposed. The first approach, namely the invariance by training, compensates for the pattern shift taking into account different targets for different pattern shifts during the training phase [21], [22]. The main drawback of such an approach is that it is inapplicable in many operating situations. In fact, the number of possible variations of patterns makes the training set too large, increasing at the same time the computational complexity of the learning system. The second technique, namely invariance by structure, uses neural networks whose outputs are always invariant to certain transformations. The disadvantage of such an approach is that high-order neural networks are required. Their implementation is too complicated and their application is limited. Recently, the combined neural network method is approaching as the most suitable scheme, to lower the system computational complexity (see [16] and references therein). Finally, the third technique uses feature vectors as inputs for the neural networks. These feature vectors are invariant to the required transformations and, hence, this method is called invariant feature space. This kind of recognition system usually uses the magnitude of the Fourier transform, which is invariant to linear shifts of its input vector [23]. The advantages of using an invariant feature space are as follows: the number of features can be reduced to realistic levels, the requirements on the classifier are relaxed, and the invariance for all input objects is ensured. On the other hand, the main drawback of using invariant feature spaces is the processing time (it can be very long) needed to extract the features before the classifier can be employed.

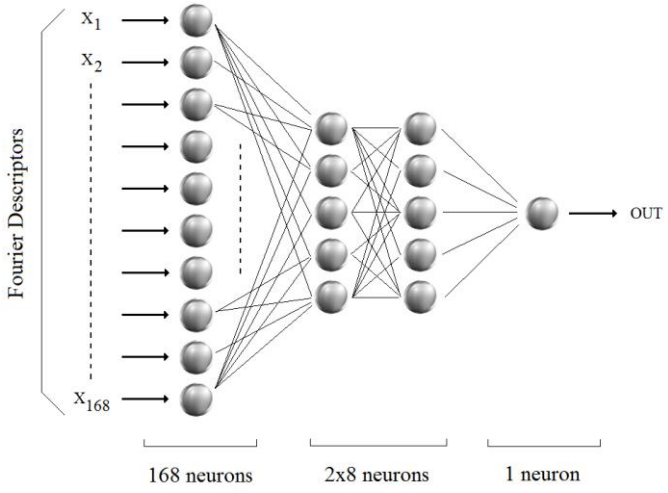



Fig. 1. Example of one Neural Network for object detection

In this work, we focus on a neural classifier that uses Fourier descriptors as inputs for the neural networks. As said before, a deterministic way to define the number of hidden layers and the number of neurons does not exist. Hence in our case, referring to the block scheme of Fig. 1, each NN is made of 4 layers: (*i*) one input-layer composed by 168 neurons and

equal to the size of the input patterns; (*ii*) two hidden-layers composed by 8 neurons; (*iii*) one output-layer characterized by only one neuron. Linear activation functions have been applied to both the input and the output-layer, while non-linear activation functions (in particular, sigmoid functions) have been chosen for the hidden-layer. Then, the overall scheme of the Neural Classifier (NC), obtained by a combined system of these NNs, is depicted in Fig. 2.

The output value of each NN can range between -1 and 1 depending of the input pattern and each input pattern contains 168 Fourier descriptors referred to a specific ISAR image (i.e. to a specific target). For example, let us now consider a NN trained for recognizing the target "TG1": to a more and more positive value of the network output corresponds a higher and higher probability that the input pattern belongs to the (correct) class TG1. Negative values of the output mean that the input pattern is not an element of the considered class. Therefore, the proposed NC takes a pattern made of 168 elements as input pattern (i.e. the Fourier descriptors of the ISAR image) and recognizes the correct class of the target. In particular, four different classes of targets, named TG1, TG2, TG3, and TG4, have been used in this paper as a *proof of concept* of the proposed combined classifier. When the NC receives a pattern belonging to one of these four classes, it returns an output value which is referred to the selected class. As shown in Fig. 2, the NC is composed by two main boxes: *a)* the inner classifier, CL*i; b)* the Final Decision-Maker, FDM.

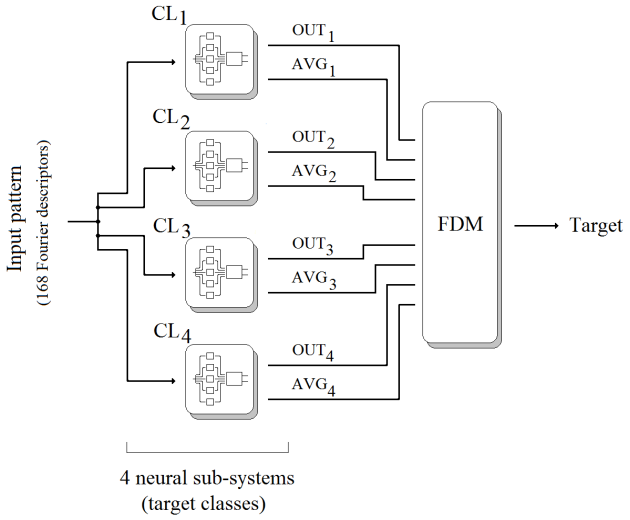



Fig. 2. Neural Classifier

Each CL*i* is referred to the related class (1, 2, 3 or 4) and consists of a neural sub-system, composed by five NNs, able to decrease the error probability of trained NNs. Indeed, each CL is composed by five different-trained NNs used to classify the same target class (see Fig. 3). The *determining boxes*, DET*i*, with *i* = 1, 2, 3, 4, perform a very important task, which is described in the hereinafter text. The rationale of our NC is as follows. The neural sub-system CL*i* contains five NNs which are separately trained, each having the aim to classify the target TG*i*. At this point, for a fixed input pattern belonging to the class TG*i* and if at least three NNs return the correct output, the CL*i* makes a correct classification of the input pattern as the TG*i* pattern. Obviously, the ideal operating case is that all the five NNs perform the correct target recognition but, to mitigate the possible output errors made by one or two NNs, this majority rule is here applied. Therefore, each CL exploits the majority rule to classify the input patterns. When the NC receives a pattern to classify, only one sub-system should be active at a time (OUT*i* is equal to 1 when CL*i* is active, -1 otherwise). The AVG*i* output returns the average value of the five NNs outputs. The box FDM simply selects the active input and shows it as the output of the whole NC system. Nevertheless, it could happen that more than one CL*i* is active at the same time anyway. In these cases, the boxes DET*i* and FDM play a very important role, exploiting the average values of the related neural sub-systems. For example, given a generic input pattern TG*x*, if the CL1 and CL2 outputs are both active, DET1 and DET2 look at the average values (AVG1 and AVG2) of the five NNs both for the CL1 and CL2 neural sub-systems, respectively. At this point, the box FDM operates as follows. If the average value AVG1 is bigger than AVG2, the input pattern belongs to the CL1 class (TG1). In the opposite case, the input pattern belongs to the CL2 class (TG2). Obviously, for the worst (limit) case in which all the CL*i* are active, the box FDM selects the output referred to the CL that performs the biggest average value.

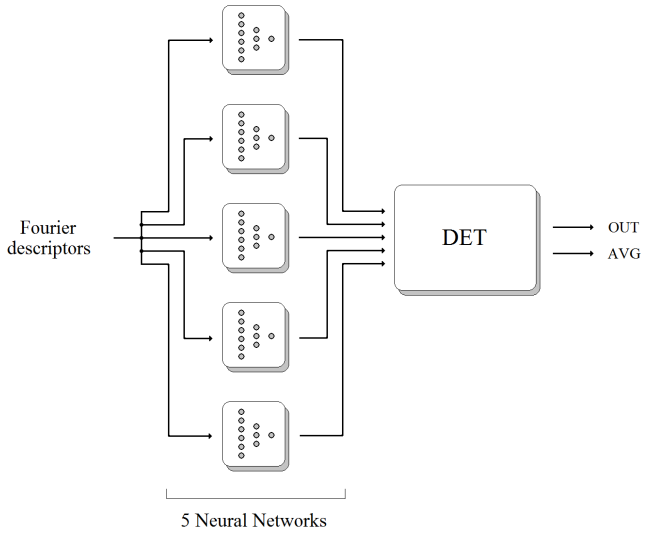



Fig. 3 - Block scheme of the generic *i*-th CL block.

## III. Conventional Multimedia Processing

The conventional multimedia processing methods for edge detection usually exploit the Sobel, Prewitt, Roberts, and Canny detectors [24]. In particular, the Sobel operator performs 2-D spatial gradient measurement on an image and so emphasizes regions of high spatial frequency that correspond to edges. The Prewitt operator is an approximate way to estimate the magnitude and orientation of the edge. Then, the Roberts operator performs 2-D spatial gradient measurement on an image and highlights regions of high spatial frequency which often correspond to edges. Finally, the Canny detector is a method to find edges by isolating noise from the image without affecting the features of the edges in

the image and then applying the tendency to find the edges and the critical value for threshold.

The classical edge detector proposed almost 20 years ago by Canny [25] performs remarkably well with its simplicity and elegance. Canny's edge detector attempts to maximize simultaneously localization and signal-to-noise ratio. A typical implementation of the Canny edge detector is as follows [26]: (*i*) first, smooth the image with an appropriate Gaussian filter to reduce desired image details; (*ii*) determine gradient magnitude and gradient direction at each pixel; (*iii*) if the gradient magnitude at a pixel is larger than those at its two neighbors in the gradient direction, mark the pixel as an edge. Otherwise, mark the pixel as the background; (*iv*) remove the weak edges by hysteresis thresholding. Indeed, in recent comparisons of edge detector performances (see for example [24] and references therein), Canny detector was the best or one of the best. This is the reason why in the following of this paper we have decided to compare the results obtained with the new multimedia processing, described in the next Section, with the multimedia processing obtained by the Canny operator.

## IV. Proposed Multimedia Processing for Target Recognition

### A. Database creation

We have used 4 targets, corresponding to 4 different military aircrafts: one MIG-29, one F-104, one F-22, and one Eurofighter-Typhoon. The ISAR images of these targets, provided by the multinational firm MBDA (Rome, Italy) are represented in Fig. 4. Then, we have created a database of ISAR images composed of more than 500 images, each one representing a target with a different azimuth angle, as shown for example in Fig. 5 for 15 different angles for the Eurofighter Typhoon target. The training and validation tests for each target class are then made of 30 and 120 ISAR images, respectively. All the NNs described in the previous section have been trained by using the well-know Levenberg-Marquardt back-propagation algorithm [27].

### B. Pre-Processing

ISAR images are usually affected by a multiplicative noise known as speckle noise. This is due to the interferences produced by radar waves and results in light and dark pixels in the ISAR image that drastically reduce the image quality. Automatic interpretation of the image as well as performing shape and features extraction become cumbersome issues to be implemented. Therefore, image pre-processing is the first and crucial phase to be addressed in order to reduce the speckle effects, and improve the image quality. A great number of different filters have been proposed in the open literature, such as the Frost [28], Lee [29], and median [18] filters. Since our ISAR images are affected by low speckle noise values, following the same approach of [18], we have used a linear filter followed by a median filter to improve the image quality. It has to be noted that, in case of images highly corrupted by noise, the median filter has been replaced by a Lee filter [28] to facilitate the automatic segmentation process.

### C. Object Detection

The shape extraction process, i.e. the process by which the contour plots are extracted, is here performed using the cascade of two different methods. First, we apply to the ISAR image the SUSAN algorithm [18], and then the output of the SUSAN method is processed by a recently introduced level set evolution (LSE) method, called distance regularized level set evolution (DRLSE) [19].

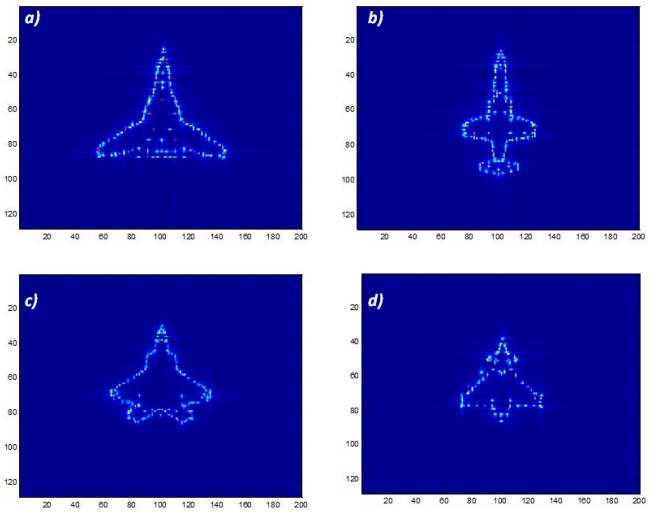



Fig. 4 - ISAR images of the target: ***a)*** MIG-29; ***b)*** Eurofighter-Typhoon; ***c)*** F-104; ***d)*** F-22

In particular, the SUSAN method is here used to extract pixels from the ISAR image belonging to two different regions: target pixels and background pixels. Then the DRLSE algorithm is used as an edge linking method, to extract the target contour.

More in details, each pixel in the input ISAR image is processed with a circular mask (named also window or kernel), and the sum of grayscale comparison between the mask center (Nucleus) and a local mask area (known as the USAN, Univalue Segment Assimilating Nucleus) is calculated. The mask is placed at each point in the input image, and then the brightness of each pixel within the USAN area is compared with the nucleus (center point), as follows:

$$C(P_i, P_0) = \begin{cases} 1 & if \left| im(x_i, y_i) - im(x_0, y_0) \right| < t \\ 0 & if \left| im(x_i, y_i) - im(x_0, y_0) \right| > t \end{cases} \quad (1)$$

where $P_0$ and $P_i$ correspond to pixel of the nucleus and any pixel of USAN area, respectively.

Then, $im(x_i, y_i)$ is the gray level of the pixel that have coordinate $(x_i, y_i)$, while $t$ stands for the brightness difference threshold. The comparison expressed by eq. (1) is performed for each pixel within the mask and the sum $S$ of all these comparisons is evaluated. Finally, the sum $S$ is compared with a fixed threshold $G$ (namely the *Geometric threshold* [28]) which is set to ¾ of the maximum value which can take $S$.

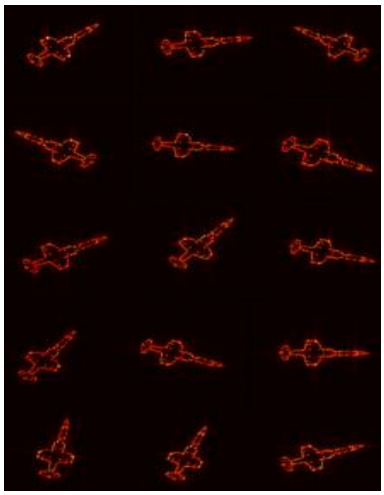


Fig. 5 - ISAR images of the target Eurofighter-Typhoon with 15 different azimuth angles.

The value of the treated pixel is then replaced by the following:

$$R(P_0) = \begin{cases} G - S & if \quad S < G \\ 0 & otherwise \end{cases} \qquad (2)$$

Like in [18], we consider the application of the SUSAN algorithm as a pre-processing step, in order to segment the input image into two regions of pixels, the ensemble composed by the target pixels and the other one full of background pixels. Hence, we can we add another condition in the standard SUSAN algorithm, modifying eq. (2) as follows:

$$R(P_0) = \begin{cases} G - S & if \quad S < G \\ im(x_0, y_0) & if \quad im(x_0, y_0) > t \\ 0 & otherwise \end{cases} \qquad (3)$$

where the threshold $t$ has been chosen according to the following [1]:

$$t = \frac{k}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} im(x_i, y_j) \qquad (4)$$

and $k$ is a constant that depends on the image size-to-target ratio. Finally, the DRLSE method of [18] is applied to extract the target shape from the results of the modified SUSAN algorithm.

### D. Features Extraction

The Fourier descriptors (FD) have been frequently used as features for image processing, remote sensing, shape recognition and classification [30]. They are chosen accordingly to their good performance in recognition systems and their implementation simplicity and efficiency. In fact, they are invariant to geometrical transformations, such as translation, scaling and rotation. The authors in [18] have used the *centroid* distance as shape signature. This distance is expressed by the distance between boundaries. Conversely, here we have applied a simple discrete Fourier transform (DFT) on the shape boundaries, extracted by the previously described methods. In particular, we have identified the boundary of a target by means of a (complex) vector, whose elements are the coordinates of the contour points. Then, we have applied the 1D-DFT to this (complex) vector, obtaining the FDs of the target. The obtained FDs are invariant to geometrical transformations.

Finally, in order to decrease the computational complexity of the entire system, we have constrained the vector length to 168 elements, corresponding to 168 FDs of the target's contour. These FDs have been used as the inputs of the new ATR technique, detailed in the following Section.

## V. Performance Analysis

In this Section, we first discuss our results obtained through the proposed multimedia processing technique, in terms of shape and edge extraction.

Then we present the results obtained through our neural classifier in terms of mean detection probability, comparing our results with state of the art detectors.

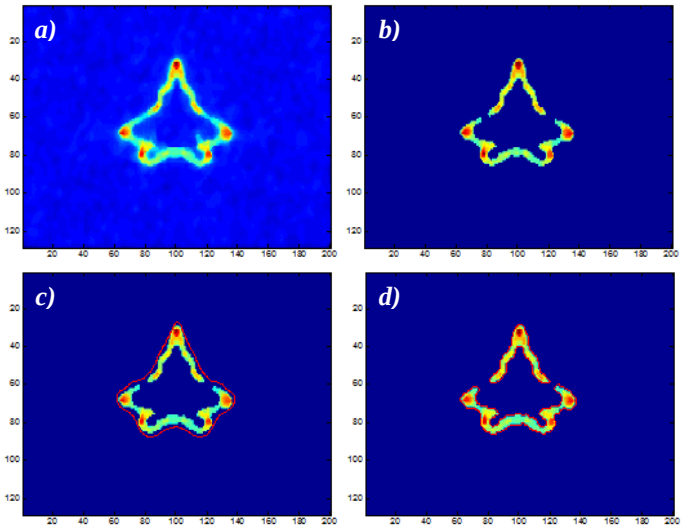



Fig. 6 - Target F22: ***a)*** noisy ISAR image; ***b)*** filtered ISAR image; ***c)*** new edge-linking method; ***d)*** edge-linking method (Canny).

### A. Results about shape and edge extraction

Here, we discuss the results of the proposed multimedia processing versus the conventional Canny detector. In the pre-processing steps, we have used a SUSAN circular mask composed by 37 pixels with a radius of 3 pixels. Referring to Fig. 6, the noisy ISAR images are first pre-processed to reduce the speckle noise (Fig. 6a), then the SUSAN method (Fig. 6b) is applied in order to extract the edge of the target (image segmentation).

Finally, the DRLSE method (Fig. 6c) and the Canny detector (Fig. 6d) are used as linking-edge techniques and the FDs are computed form both the images, as explained in Section IV. It is clearly evident from the figures that the edges obtained with the Canny method (see also Fig. 7 and Fig. 8) are characterized by a poorer quality in respect to the ones obtained with the new multimedia processing.

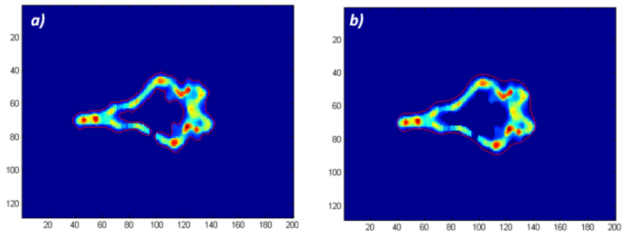



Fig. 7 - Target Eurofighter Typhoon, edges obtained by means of the: ***a)*** Canny detector; ***b)*** new multimedia processing.

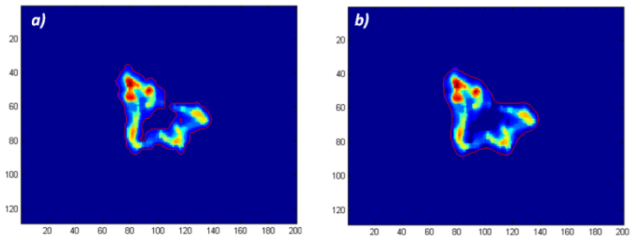



Fig. 8 - Target F-104, edges obtained by means of the: ***a)*** Canny detector; ***b)*** new multimedia processing.

### *B. Results about target classification*

The 168 FDs, describing the specific target under investigation, are then passed to the classification step, for the training and validation phase. As previously detailed, each ISAR image is characterized by a pattern of 168 Fourier descriptors.

The confusion matrix for the validation test, for the four targets examined in this paper, is shown in Tab. 1 and in Tab. 2 for the Canny detector and the new multimedia processing, respectively.

In particular, we have reported the percentage of correct detection, indicated by bold numbers, and the percentage of errors (false recognition) in the tables. For example, the target TG1 is recognized with a detection probability of more than 93% with the new multimedia processing, while the Canny method reaches only a percentage of 90.83%.

Then, the target TG4 is detected with lower probability in both cases and it is automatically identified as TG1 with a percentage of 5.0% or 33.33%, for the new and Canny processing respectively. Notwithstanding the last two targets are characterized by lower detection probabilities, the obtained results, by the new multimedia processing, seem really promising since the NC is able to achieve quite large percentage values of correct object detection. In particular, this is due to the bad performance of the Canny detector for different azimuth angles.

See for example Fig. 7 (a, b) and Fig. 8 (a, b) where the edges extracted with the Canny detector and the new processing are compared. In particular, two targets of interest are considered: the Eurofighter Typhoon and the F-104, respectively. It is clearly visible from the figures that, in the case of the Canny detector, the smoothing effect due to the segmentation process do not allow to correctly extract the edges of the detected object.

Tab. 1. Object detection performance of our neural classifier with the Canny multimedia processing

| | TG1 | TG2 | TG3 | TG4 |
|---|---|---|---|---|
| TG1 | **90.83%** | 5.81% | 0.00% | 3.34% |
| TG2 | 60.84% | **25.82%** | 0.81% | 12.50% |
| TG3 | 73.33% | 15.01% | **0.81%** | 10.83% |
| TG4 | 33.33% | 12.51% | 0.00% | **54.17%** |

**Tab. 2**. Object detection performance of our neural classifier with the new multimedia processing

| | TG1 | TG2 | TG3 | TG4 |
|---|---|---|---|---|
| TG1 | **93.33%** | 0.00% | 6.66% | 0.00% |
| TG2 | 2.55% | **96.66%** | 0.83% | 0.00% |
| TG3 | 22.50% | 0.00% | **70%** | 7.50% |
| TG4 | 5.00% | 0.00% | 28.33% | **66.66%** |

Another advantage of our processing lies in the combined structure of the proposed classifier. In fact, we always exploit the most suitable neural sub-system for each target class, i.e. the inner classifier CL composed by 5 NNs. The CL sub-system aims at decreasing the error probability with respect to the conventional case in which only one NN decides about the class of the target. Moreover, the further use of average values (performed by the DET*i* sub-systems) improves the performances of the proposed classifier when ambiguities exist at the outputs of the neural sub-systems. Finally, in order to prove the efficiency of our NC with respect to state of the art detectors, a comparative analysis is shown in Tab. 3. In particular, the mean values of the correct classification probability are reported in tab. 2, for our proposed NC (mean recognition of 81.6%) and for two classifiers proposed in [18]. In particular, the authors in [18] obtain a mean recognition percentage of 75.98%, using a K Nearest-Neighbor classifier (K-NN) and then, they improve the system performances exploiting the Support Vector Machine (SVM) algorithm (reaching a mean detection value of 80.37%). However, this further approach appears less effective than the one here presented.

Tab. 3. Comparison between our method and the classifiers of [18]

| | Proposed NC | K-NN | SVM |
|---|---|---|---|
| Mean Detection | **81.60%** | 75.98% | 80.37% |

## VI. CONCLUSIONS AND FUTURE WORKS

This work has proposed a new automatic target classifier, based on a combined neural networks' system, by ISAR image processing. The novelty introduced in our work is twofold. We have first introduced a novel automatic classification procedure, and then we have discussed about an improved multimedia processing of ISAR images for automatic object detection.

We have exploited a neural classifier, composed by a combination of 20 feed-forward artificial neural networks. The classifier is used to recognize aircraft targets extracted from ISAR images. The combination of two image processing techniques, recently introduced in literature, is exploited to improve the shape and features extraction process. Then, Invariant Fourier descriptors are computed and used as input features to our combined system. Performance analysis is carried out in comparison with conventional multimedia processing techniques as well as with classical automatic target recognition systems. Numerical results, obtained from wide simulation trials, evidence the efficiency of the proposed approach for the application to automatic aircraft target recognition. Future works will regard the improvement of the performances of the single NNs by applying suitable optimization algorithms to the NNs learning process. Indeed, it is possible to operate a multivariate function decomposition with the aim to perform the learning optimization of Multi-Input-Single-Output (MISO) feed-forward Neural Networks [31]. Furthermore, applying new powerful search algorithms (e.g. meta-heuristic algorithms such as those shown in [32]-[35]) can increase the generalization feature of the Neural Networks in particular after they are built by using a partitioning of the domain (see [36]). Finally, other more elaborate algorithms could be applied to the multimedia processing by starting from novel concepts already existing in literature (e.g. [37][38]).

## ACKNOWLEDGMENT

The authors acknowledge the support of MBDA, Rome (Italy), for providing the ISAR images.

## REFERENCES

[1] F. L. Lewis, *Wireless Sensor Networks*, in Smart Environments: Technologies, Protocols, and Applications, ed. D.J. Cook and S.K. Das, John Wiley, New York, 2004.

[2] I. F. Akyildiz , T. Melodia, K. R. Chowdhury, "A survey on wireless multimedia sensor networks", *Elsevier Computer Networks*, vol. 51, pp. 921-960, 2007.

[3] Editorial: "Signal processing techniques for ISAR and feature extraction", *IET Sign. Proc*., vol. 2, no. 3, pp. 189 – 191, 2008.

[4] V. Zeljkovic, Q. Li, R. Vincelette, C. Tameze, F. Liu, F., "Automatic algorithm for inverse synthetic aperture radar images recognition and classification", *IET Radar, Sonar & Navig.*, vol. 4, no. 1, pp. 96 – 109, 2010.

[5] D. Pastina, C. Spina, "Multi-feature based automatic recognition of ship targets in ISAR", *IET Radar, Sonar & Navig.*, vol. 3, no. 4, pp. 406 – 423, 2009.

[6] B. Leibe, A. Leonardis, and B. Schiele(2004), "Combined Object Categorization andSegmentation with an Implicit Shape Model," *Proc. ECCV Workshop Statistical Learning in Computer Vision*, 2004.

[7] I. Kokkinos, P. Maragos, "Synergy between Object Recognition and image segmentation using Expectation and Maximization Algorithm", *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 31(8), pp. 1486-1501, 2009.

[8] M. N. Saidi, B. Hoeltzener, A. Toumi, A. Khecnhaf, D. Aboutajdine, "Recognition of ISAR images: target shapes features extraction", *IEEE 3rd Int. Conf. on Inf. and Commun. Techn.: from Theory to Appl. (ICTTA)*, pp. 1–6, Apr. 2008.

[9] G. T. Maskall, "An application of nonlinear feature extraction to the classification of ISAR images", *Proc. of RADAR 2002*, pp. 405–408, 2002.

[10] M. Vespe, C. J. Baker, H. D. Griffiths, "Outline structural representation for radar target classification based on non-radar templates", *IEEE Int. Conf. on Radar (CIE '06)*, pp. 1–4, 2006.

[11] J. Manikandan, B. Venkataramani, M. Jayachandran, "Evaluation of edge detection techniques towards implementation of automatic target recognition", *Int. Conf. on Computational Intelligence and Multimedia Appl.*, vol. 2, pp. 441–445, Dec. 2007.

[12] A. Toumi, B. Hoelzener, A. Khenchaf, "Using Watersheds segmentation on ISAR image for automatic target recognition", *Int. Conf. on Digital Information Management*, pp. 285-290, 2007.

[13] F. Wang, W. Sheng, X. Ma, H. Wang, "Target automatic recognition based on ISAR image with wavelet transform and MBLBP", *Int. Symp. on Sign. Syst. and Electr. (ISSSE)*, vol. 2, pp. 1 – 4, 2010.

[14] K. T. Kim, D. K. Seo, H. T. Kim, "Efficient classification of ISAR images", *IEEE Trans. Antennas Propag.*, vol. 53, no. 5, pp. 1611–1621, 2005.

[15] M. Martorella, E. Giusti, L. Demi, Z. Zhou, A. Cacciamano, F. Berizzi, B. Bates, "Automatic target recognition by means of polarimetric ISAR images: a model matching based algorithm", *Int. Conf. on Radar*, pp. 27-31, Sept. 2008.

[16] H. Yuankui, Y. Yiming, "Automatic target recognition of ISAR images based on Hausdorff distance", *1st Asian and Pacific Conf. APSAR*, pp. 477-479, Nov. 2007.

[17] P. T. Dung, "Combined neural networks for radar target recognition from radar range profiles", *IEEE Int. Conf. on Adv. Techn. for Commun.*, 2008.

[18] M. N. Saidi, A. Toumi, B. Hoeltzener, A. Khenchaf, D. Aboutajdine, "Aircraft Target Recognition: A novel approach for features extraction from ISAR images", *IEEE Int. Radar Conf. - Surveillance for a Safer World*, pp. 1 – 5, 2009.

[19] C. Li, C. Xu, C. Gui, M. D. Fox, "Distance Regularized Level Set Evolution and Its Application to Image Segmentation", *IEEE Trans. on Image Proc*., vol. 19, no 12, pp. 3243 – 3254, 2010.

[20] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995

[21] V. M. Orlenko, A. S. Bityutskiy, "Simulation of aerial target radar recognition using neural networks", *Proc. of the IRS*, pp. 729-733, 2003.

[22] A. Zyweck, R. E. Bogner, "Radar target classification of commercial aircraft", *IEEE Trans. on Aerospace and Electr. Syst.*, vol. 32, no. 2, pp. 598-606, 1996.

[23] S. Hudson, D. Psaltis, "Correlation filter for aicraft identification from radar range profiles", *IEEE Trans. on Neural Networks*, vol. 29, no. 3, pp. 741-748, 1193.

[24] F. A. Pellegrino, W. Vanzella, V. Torre, "Edge Detection Revisited", *IEEE Trans. on Systems, Man, And Cybernetics—Part B: Cybernetics*, vol.. 34, no. 3, pp. 1500-1518, June 2004.

[25] J. Canny, "A Computational Approach To Edge Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp.:679–698, 1986.

[26] L. Ding, A. Goshtasby, "On the Canny edge detector", *Elsevier Pattern Recognition*, vol. 34, pp. 721-725, 2001.

[27] P. R. Gill, W. Murray, M. H. Wright, *The Levenberg-Marquardt Method*, in Practical Optimization. London: Academic Press, pp. 136-137, 1981.

[28] V. S. Frost, J. A. Stiles, A. Josephine, K. S. Shanmu-gan, J. C. Holtzman, "A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise," *IEEE Trans. on Pattern Analysis and Machine Intellig.*, vol. PAMI-4, no. 2, pp. 157-166, 1982.

[29] J. S. Lee, "Speckle Analysis and Smoothing of Synthetic Aperture Radar Images," *Computer Graphics and Image Processing*, vol. 17, pp. 24-32, 1981.

[30] M. Sarfraz, "Object Recognition using Fourier Descriptors: Some Experiments and Observations," *IEEE Int. Conf. on Computer Graphics, Imaging and Visualisation*, pp.281-286, 2006.

[31] F. Riganti Fulginei, A, Salvini and M. Parodi, "Learning optimization of neural networks used for mimo applications based on multivariate functions decomposition". Inverse Problems in Science & Engineering (IPSE). vol. 20, p. 29-39, 2012

[32] F. Riganti Fulginei, A. Salvini and G. Pulcini, "Metric-topological-evolutionary optimization". Inverse Problems in Science & Engineering (IPSE), vol. 20, p. 41-58, 2012

[33] S. Coco, A. Laudani, F. Riganti Fulginei and A. Salvini, "Shape optimization of multistage depressed collectors by parallel evolutionary algorithm". IEEE Transactions on Magnetics, vol. 48, p. 435-438, 2012

[34] F. Riganti Fulginei, A. Salvini, A. Laudani and S. Coco, "Team 22 problem approached by a hybrid artificial life method". Compel, vol. 31, p. 816-826, 2012

[35] F. Riganti Fulginei, A. Laudani, A. Salvini, S. Coco, G. Pollicino and G. Pulcini, "TWT magnetic focusing structure optimization by parallel evolutionary algorithm". Compel, vol. 31, p. 1338-1346, 2012

[36] F. Riganti Fulginei and A. Salvini, "Neural network approach for modelling hysteretic magnetic materials under distorted excitations". IEEE Transactions on Magnetics, vol. 48, p. 307 -310, 2012

[37] F. Benedetto and G. Giunta, "A Fast Time-Delay Estimator of PN Signals". IEEE Transactions on Communications, vol. 59, p. 2057 - 2062, 2011

[38] F. Benedetto, G. Giunta and S. Bucci, "A Unified Approach for Time-Delay Estimators in Spread Spectrum Communications". IEEE Transactions on Communications, vol. 59, p. 3421 - 3429, 2011

# An Effective Identification of Species from DNA Sequence: A Classification Technique by Integrating DM and ANN

Sathish Kumar S
Research Scholar
Dr MGR Educational and Research Institute University.
Chennai, Tamil Nadu, India
Tel: +919886372152

Dr.N.Duraipandian, M.E., Ph.D
Vice Principal
Velammal Engineering College
Ambattur – Redhills Road, Chennai,
Tamil Nadu, India

***Abstract*— Species classification from DNA sequences remains as an open challenge in the area of bioinformatics, which deals with the collection, processing and analysis of DNA and proteomic sequence. Though incorporation of data mining can guide the process to perform well, poor definition, and heterogeneous nature of gene sequence remains as a barrier. In this paper, an effective classification technique to identify the organism from its gene sequence is proposed. The proposed integrated technique is mainly based on pattern mining and neural network-based classification. In pattern mining, the technique mines nucleotide patterns and their support from selected DNA sequence. The high dimension of the mined dataset is reduced using Multilinear Principal Component Analysis (MPCA). In classification, a well-trained neural network classifies the selected gene sequence and so the organism is identified even from a part of the sequence. The proposed technique is evaluated by performing 10-fold cross validation, a statistical validation measure, and the obtained results prove the efficacy of the technique.**

***Keywords- Pattern Generation; DNA Sequence; Pattern Support; Mining; Neural Network.***

## I. INTRODUCTION

Bioinformatics is a rapidly growing area of computer science [19] that deals with the collection, organization, and analysis of Deoxyribonucleic acid (DNA) and protein sequence [18]. Today it addresses the formal and practical issues that occur in the management and analysis of genomic and proteomic data because it includes the formation and development of databases, algorithms, computational and statistical technique, and hypothesis [1].

Genomic signal processing (GSP) is a relatively new area in bio-informatics that uses traditional digital signal processing techniques to deal with digital signal representations and analysis of genomic data [2] [12]. GSP gains biological knowledge by the analysis, processing, and use of genomic signals and translates the gained biological knowledge into systems-based applications [3]. Integration of signal processing theories and methods with global understanding of functional genomics with significant emphasis on genomic regulation is the main objective of GSP [4].

The whole DNA of a living organism is known as its Genome [5]. Genomic signals carry genomic information to all the processes that take place in an organism [6]. Essentially DNA is a nucleic acid that has two long strands of nucleotides twisted in the form of a double helix and its external backbone is made up of alternating deoxyribose sugar and phosphate molecules. The nitrogenous bases Adenine, Guanine, Cytosine and Thymine are present in the interior portion of the DNA in pairs [13] [9]. DNA and proteins can be mathematically represented as character strings, where each character is a letter of the alphabet [6] [10] [11].

One of the vital tasks in the study of genomes is gene identification [7]. DNA analysis utilizes methods such as clustering [20], data mining [21] [22] [23], gene identification [24] and gene regulatory network modeling [25] [26]. These methods present cutting edge research topics and methodologies for the purpose of facilitating collaboration between researchers and bioinformaticians. Mining bioinformatics data is a rising field at the intersection of bioinformatics and data mining [14]. Some of them belong to the category of data mining that decides whether or not an example not yet noticed is of a predefined type. Increased availability of huge amount of biomedical data and the expectant need to turn such data into useful information and knowledge is the main reason for the recent increased attention in data mining in the biomedical industry.

Large number of research works that incorporate data mining in bioinformatics for different purposes are available in the literature [15] [16] [17]. . A few important such researches are reviewed in section 2. One important research of this type is the identification of species or name of an organism from its gene sequence. Characterization of the unknown environmental isolates with the genomic species is not easy because genomic species are especially heterogeneous and poorly defined [8]. Identifying the species or the organism from its gene sequence is a challenging task. In this paper, we propose a classification technique to effectively classify the species or name of an organism from its DNA sequence. This technique is detailed with mathematical formulations and illustrations in section 3. Section 4 discusses the implementation results and Section 5 concludes the paper.

## II. RELEATED WORKS

Plenty of research works deals with the mining knowledge from the genomic sequences. Some of the recent research works are briefly reviewed here. Riccardo Bellazzi et al. [27] have discussed that in the past years, the gene expression data analysis that are aiming at complementing microarray analysis with data and knowledge of various existing sources has grown from being purely data-centric to integrative. Focusing on the evolution of gene expression data mining techniques toward knowledge-based data analysis approaches, they have reported on the overabundance of such techniques. Particularly, latest developments in gene expression-based analysis methods utilized in association and classification studies, phenotyping and reverse engineering of gene networks have been discussed.

The gene expression data sets for ovarian, prostate, and lung cancer was examined by Shital Shah et al. [28]. For genetic expression data analysis, an integrated gene-search algorithm was presented. For making predictions and for data preprocessing (on partitioned data sets) and data mining (decision tree and support vector machines algorithms), a genetic algorithm and correlation-based heuristics was included in the their integrated algorithm. The knowledge, which was obtained by the algorithm, has high classification accuracy with the capability to recognize the most important genes. To further improve the classification accuracy, bagging and stacking algorithms were employed. The results were compared with the literary works. The cost and complexity of cancer detection and classification was eventually condensed by the mapping of genotype information to the phenotype parameters.

Locating motif in bio-sequences, which is a very significant primitive operation in computational biology, was discussed by Hemalatha et al. [29]. Computer memory space requirement and computational complexity are few of the computational requirements that are needed for a motif discovery algorithm. To overcome the intricacy of motif discovery, an alternative solution integrating genetic algorithm and Fuzzy Art machine learning approaches was proposed for eradicating multiple sequence alignment process. The results that were attained by their planned model to discover the motif in terms of speed and length were compared with the enduring technique. By their technique, the length of 11 was found in 18 sec and length of 15 in 24 sec, whereas the existing techniques found length of 11 in 34 sec. When compared to other techniques, the proposed one has outperformed the accepted existing technique. By employing MATLAB, the projected algorithm was put into practice and with large DNA sequence data sets and synthetic data sets, it was tested.

An interactive framework which is based on web for the analysis and visualization of gene expressions and protein structures was described by Ashraf S. Hussein [30]. The formulation of the projected framework encountered various confronts because of the variety of significant analysis and visualization techniques, moreover to the survival of a diversity of biological data types, on which these techniques function. Data incorporated from heterogeneous resources, for instance expert-driven data from text, public domain databases and various large scale experimental data and the lack of standard I/O that makes it difficult to integrate the most recent analysis and visualization are the two main challenges that directed the formulation of the current framework. Hence, the basic novelty in their proposed framework was the integration of the state-of -art techniques of both analysis and visualization for gene expressions and protein structures through a unified workflow. Moreover, a wide range of input data types are supported by it and three dimensional interactive outputs ready for exploration by off-the-shelf monitors and immersive, 3D, stereo display environments can be exported by it using Virtual Reality Modeling Language (VRML).

A stomach cancer detection system, which is on the basis of Artificial Neural Network (ANN) and the Discrete Cosine Transform (DCT), was developed by Ahmad M. Sarhan [31]. By employing DCT the projected system extracted the classification features from stomach microarrays. The extracted characteristics from DCT coefficients were applied to an ANN for further classification (tumor or non-tumor). The microarray images that were employed were acquired from the Stanford Medical Database (SMD). Simulation results has illustrated that a very high success rate was produced by the proposed system.

The challenging issue in microarray technique which was to analyze and interpret the large volume of data was discussed by Valarmathie et al. [32]. This can be made possible by the clustering techniques in data mining. In hierarchical and k-means clustering techniques which are hard clustering, the data is split into definite clusters, where each cluster has exactly one data element so that the result of the clustering may be wrong many times.

The problems that are addressed in hard clustering can be resolved in fuzzy clustering technique. Amid all fuzzy based clustering, fuzzy C-means (FCM) is best suited for microarray gene expression data. The problem that is related with fuzzy C-means was the amount of clusters that are to be generated for the given dataset and that needs to be notified first. By combining the technique with a popular probability related Expectation Maximization (EM) algorithm, it can be solved to model the cluster structure of gene expression data and it has offered the statistical frame work. Determining the accurate number of clusters and its efficient interpretation is the main purpose of the projected hybrid fuzzy C-means technique.

Explorative studies in support of solutions to facilitate the analysis and interpretation of mining results was described by Belmamoune et al. [33]. A solution that was located in the extension of the Gene Expression Management System (GEMS) was described, i.e. an integrative framework for spatio-temporal organization of gene expression patterns of zebra fish to a framework that supports data mining, data analysis and patterns interpretation.

As a proof of principle, the GEMS is provided with data mining functionality which is appropriate to monitor spatio-temporal, thus generating added value to the submission of data for data mining and analysis. On the basis of the availability of domain ontologies, the analysis of the genetic networks was done which vigorously offers the meaning to the

discovered patterns of gene expression data. Grouping of data mining with the already accessible potential of GEMS considerably augments the existing data processing and functional analysis strategies.

## III. The Integrated Technique for Species Classification

The proposed species classification technique classifies species based on the given DNA sequence. The DNA sequence is comprised of four basic nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). Every species has a long DNA sequence, which is formed by the four nucleotides.

The DNA sequence defines the attributes, nature and type of the species. The proposed technique is an integration of data mining and artificial intelligence. In the proposed technique, firstly, nucleotide patterns are mined from the sequence. The mined patterns form a nucleotide pattern database with higher dimension. So, secondly, the dimension of the pattern database is reduced by MPCA. Finally, the dimensionality reduced pattern database is used to train the neural network. The technique is described in the further sub sections.

The proposed species classification technique classifies a species based on its DNA sequence. The four basic nucleotides, Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) are the building blocks of the long DNA sequence in every species. A DNA sequence defines the attributes, nature and type of the species. The proposed technique is developed by integrating data mining and artificial intelligence techniques. Firstly, the proposed technique mines the nucleotide patterns from the sequence and forms a high dimensional nucleotide pattern database with the mined patterns.

Secondly, the technique uses MPCA and reduces the dimension of the pattern database. Finally, the dimensionality reduced pattern database is used to train a neural network. The following sub sections elaborately describe this technique.

### *A. Mining Nucleotide Patterns from DNA sequence*

The first and initial stage of the proposed technique mines the nucleotide pattern from the DNA sequence. At this stage, patterns formed by different combinations of nucleotides are mined using a novel mining algorithm. Let be the DNA sequence, which is a combination of four nucleotides A, G, C and T. For instance, a sample DNA sequence is given as CGTCGTGGAA.

From the sequence, the mining algorithm extracts different nucleotide patterns and their support. The algorithm is comprised of two stages, namely, pattern generation and support finding. In pattern generation, patterns with different length are generated whereas in support finding, support values for every generated pattern are determined from the DNA sequence. The basic structure of the algorithm is given as a block diagram in Fig. 1.

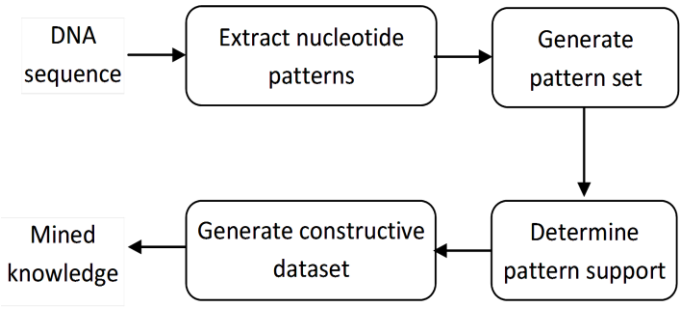



Figure 1. Block diagram of the pattern mining algorithm

#### *1) Pattern generation*

In pattern generation, different possible combinations of nucleotide base pairs are generated. As a reference, a base set $B$ is generated with cardinality $|B|=4$, which has the elements $\{A,G,C,T\}$. Let, $\{P_l\}; l=1,2,\Lambda\ ,L_{\max}$, be the pattern set to be generated, where, $L_{\max}$ is the maximum length of a pattern in a pattern set. The pattern set is generated as follows

$$\{P_l\}_{k^{(l)}} = \{P_l\}_{k^{(l)}-1} \cup \{B(a_1)B(a_2)\Lambda\ B(a_l)\} \quad (1)$$

Where, $k^{(l)}=1,2,\Lambda\ ,|B|^l$, $1\le a_1,a_2,\Lambda\ ,a_l \le |B|$ and $\{B(a_1)B(a_2)\Lambda\ B(a_l)\}$ is a set of different combinations of nucleotide bases. Eq. (1) operates with the criterions, $\{P_l\}_{k^{(l)}-1} \subseteq \{P_l\}_{k^{(l)}}$ and $\{B(a_1)B(a_2)\Lambda\ B(a_l)\} \subseteq \{P_l\}$. Eq. (1) formulated for pattern generation is analyzed using two examples.

***Example 1:*** *To generate a two length pattern set* $P_2$, *i.e.* $l=2$. Here, two indexing variables $a_1$ and $a_2$ are generated. At every $k$ i.e. $k=1\ to\ 16$ and for its corresponding $a_1$ and $a_2$, the obtained $\{B(a_1)B(a_2)\}$ and $\{P_2\}_{k^{(2)}-1}$ are tabulated in Table II.

Hence, $P_2$ is obtained as $\{P_2\}=\{AA,AG,AC,AT,GA,GG,GC,GT,CA,CG,CC,CT,TA,TG,TC,TT\}$ by integrating $P_2$ obtained from every $k^{th}$ iteration. From $P_1,P_2,\mathrm{K}\ ,P_{L_{\max}}$, a consolidated pattern set $P$, which is the required pattern to be generated, is obtained as $P=P_1\cup P_2\cup\Lambda\ P_{L_{\max}}$. The cardinality of $P$ can be determined as $|P| = \sum_{l=1}^{L_{\max}} |P_l|$.

TABLE I. DIFFERENT COMBINATIONS AND PATTERN SETS GENERATED FOR EVERY $a_1$, $a_2$ AND $k$

| k | $a_1$ | $a_2$ | {B($a_1$) B($a_2$)} | $\{P_2\}_k(2)_{-1}$ |
|---|---|---|---|---|
| 1 | 1 | 1 | AA | {} |
| 2 | 1 | 2 | AG | {AA} |
| 3 | 1 | 3 | AC | {AA, AG} |
| 4 | 1 | 4 | AT | {AA, AG, AC} |
| 5 | 2 | 1 | GA | {AA, AG, AC, AT} |
| 6 | 2 | 2 | GG | {AA, AG, AC, AT, GA} |
| 7 | 2 | 3 | GC | {AA, AG, AC, AT, GA, GG} |
| 8 | 2 | 4 | GT | {AA, AG, AC, AT, GA, GG,GC} |
| 9 | 3 | 1 | CA | {AA, AG, AC, AT, GA, GG,GC, GT} |
| 10 | 3 | 2 | CG | {AA, AG, AC, AT, GA, GG,GC, GT, CA} |
| 11 | 3 | 3 | CC | {AA, AG, AC, AT, GA, GG,GC, GT, CA, CG} |
| 12 | 3 | 4 | CT | {AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC } |
| 13 | 4 | 1 | TA | {AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT} |
| 14 | 4 | 2 | TG | {AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT,TA} |
| 15 | 4 | 3 | TC | {AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT,TA,TG} |
| 16 | 4 | 4 | TT | {AA, AG, AC, AT, GA, GG,GC, GT, CA, CG, CC,CT,TA,TG,TC} |

*2) Determination of Pattern Support*

The support, which has to be determined for every extracted pattern, describes the DNA attribute. By performing a window based operation over the sequence $g$, the support can be determined. Window of sequences are determined for different lengths as follows

$$w_l(j) = g(j, j+1, \Lambda\ , j+l-1) \qquad (2)$$

Once the window of sequences is extracted support is determined for the mined patterns. The pseudo code, which is given below, describes the procedure to determine the support for every pattern.

**Initialize** $C$ to zero
**Read** window of sequence $w$
**For** every $l$ - length pattern, $P_l$
  **For** each element $i$ in $P_l$, $P_l(i)$
    **For** each window of sequence $w_l(j)$
      **If** $P_l(i)$ and $w_l(j)$ are same
        **Increment** $C_l(i)$
      **End if**
    **End for**
  **End for**
**End for**
**Return** $C$

Figure 2. Pseudo code to determine support for every mined

Figure 3. ained $C$ for each different length pattern hapattern

The obts the support for all the elements that are present in the corresponding pattern set. From the mined pattern and its corresponding support, a constructive dataset is generated.

*3) Constructive dataset generation*

A raw dataset is generated using the aforesaid mining algorithm. But the dataset is not constructive for further operation. In this stage, a constructive dataset is generated from the mined dataset, which comprises of patterns with different lengths and their support.

To accomplish this, firstly the patterns which have length $l \geq 2$ are taken. From the pattern set, the modified and constructive dataset is generated as given in Table 3.

TABLE II. A GENERAL STRUCTURE OF THE PROPOSED CONSTRUCTIVE DATASET

| | **A** | **G** | **C** | **T** |
|---|---|---|---|---|
| A | $C_2(1)$ | $C_2(2)$ | $C_2(3)$ | $C_2(4)$ |
| G | $C_2(5)$ | $C_2(6)$ | $C_2(7)$ | $C_2(8)$ |
| C | $C_2(9)$ | $C_2(10)$ | $C_2(11)$ | $C_2(12)$ |
| T | $C_2(13)$ | $C_2(14)$ | $C_2(15)$ | $C_2(16)$ |
| AA | $C_3(1)$ | $C_3(2)$ | $C_3(3)$ | $C_3(4)$ |
| AG | $C_3(5)$ | $C_3(6)$ | $C_3(7)$ | $C_3(8)$ |
| AC | $C_3(9)$ | $C_3(10)$ | $C_3(11)$ | $C_3(12)$ |
| AT | $C_3(13)$ | $\Lambda\ \Lambda$ | $\Lambda\ \Lambda$ | $\Lambda\ \Lambda$ |
| M<br>M | | | | |

In the constructive dataset, all the patterns except single length pattern are considered. Hence, the dataset is of size $4^{L_{\max}-1} \times 4$. The generated constructive dataset belongs to a particular gene sequence. Similarly, the constructive dataset for different sequences are generated. Hence, the final dataset $G_{xy}^{(z)}$ ; $x = 1,2,\Lambda\ ,4^{L_{\max}-1}$ , $y = 1,2,3\ and\ 4$ and $z = 1,2,\Lambda\ , N_g$ is obtained, which is subjected to further processing.

*B. MPCA-based Dimensionality reduction*

In all tensor modes, the multilinear algorithm MPCA captures most of the variation present in the original tensors by seeking those bases in each mode that allow projected tensors and performs dimensionality reduction [35]. Initially, in the process of dimensionality reduction, the distance matrix for every $z^{th}$ matrix is determined as follows, $D^{(z)} = G^{(z)} - \mu$ (3)

Where,

$$\mu_{xy} = \frac{1}{N_G} \sum_{z=0}^{N_G - 1} G_{xy}^{(z)} \quad (4)$$

Using Eq. (3) and by determining the mean matrix $\mu$ for $G^{(z)}$ using Eq. (4), the distance matrix can be calculated. Then with mode 2, tensor representations [34] $T_1^{(z)}$ and $T_2^{(z)}$ are given to the obtained distance matrix. A projection matrix $\Psi$ is determined as follows,

$$\Psi = \sum_{z=0}^{N_G - 1} T^{(z)} \left( T^{(z)} \right)^T \quad (5)$$

For both $T_1^{(z)}$ and $T_2^{(z)}$, the projection matrix ( $\Psi_1$ and $\Psi_2$ ) are determined using the generalized form of calculation given in Eq. (5). For $\Psi_1$ and $\Psi_2$ , the corresponding eigenvectors $E_1$ and $E_2$ and the corresponding Eigen values $\lambda_1$ and $\lambda_2$ are determined by subjecting the projection matrix to a generalized eigenvector problem. The rows of the eigenvector are arranged based on the index of the eigenvalues sorted in the descending order. The modified eigenvector $E_1^{'}$ and $E_2^{'}$ are obtained by transposing the arranged eigenvector. The cumulatively distributed Eigen values for the sorted eigenvalues are generally determined using the following equation.

$$\lambda_x^{'} = \frac{\lambda_x^{cdf}}{\sum_{x=0}^{|\lambda|-1} \lambda_x^{sort}} \quad (6)$$

The sorted Eigen values $\lambda_x^{sort}$ and the cumulatively distributed Eigen values $\lambda_x^{cdf}$ of Eq. (6), can be determined as

$$\lambda_x^{cdf} = \lambda_x^{sort} + \lambda_{x-1}^{cdf} \quad (7)$$

Where, $\lambda_0^{cdf} = \lambda_0^{sort}$ at $x = 0$ . The new dimension $\lambda_T$ is calculated from the obtained $\lambda_x^{'}$, using a dimensional threshold $D_T$ . To accomplish this, the indices of all eigenvalues that satisfy the condition $\lambda_x^{'} \geq D_T$ are identified. Then, by extracting the first $\lambda_T$ rows of $E_1^{'}$ and $E_2^{'}$, the corresponding dimensionality reduced eigenvectors $E_1^{''}$ and $E_2^{''}$ are determined. For the $E_1^{''}$ and $E_2^{''}$ , again tensor matrices but $T_1^{(z)}$ and $T_2^{(z)}$ times are determined [35]. The

process followed for projection matrix is repeated for the tensor matrices to obtain $\lambda_{x_1}^{new}$ and $\lambda_{x_2}^{new}$ and $E_1^{new}$ and $E_2^{new}$. The weight of both the tensor eigen values are determined as $\lambda_x^w = \sqrt{\lambda_{x_1}^{new}\lambda_{x_2}^{new}}$. Then, the dimensionality reduced matrix $G^{(z)}$ of size $N_R \times N_T$, is obtained by using the MPCA projections [35], where, $N_R$ can be determined as $N_R = \lambda_{T_1}.\lambda_{T_2}$.

### C. Classification using ANN

For $N_G$ gene sequences, the dimensionality reduced gene patterns and their support are provided by the MPCA. Using ANN, the class of the original sequence can be identified using the dataset. Two classical operations, training and testing are involved in the classification. The neural network is trained using the $N_G$ pattern dataset. Here, the process is performed using multilayer feed forward neural network, depicted in Fig. 3. $N_R$ Input nodes, $N_H$ hidden nodes and an output node are present in the network.

Before performing any task, the ANN must be trained. Once trained, the ANN capably identifies the species by finding the class of the gene sequence. The training phase and classification phase of the ANN are described below.

#### 1) Training Phase

Back Propagation (BP) algorithm is used to train the constructed feed forward network. The step-by-step procedure utilized in the training process is given below.

1. Assign arbitrary weights generated within the interval to links between the input layer and hidden layer as well as hidden layer and output layer.
2. Using Eq. (8), (9) and (10), determine the output of input layer, hidden layer and output layer respectively by inputting constructive dataset to the network.

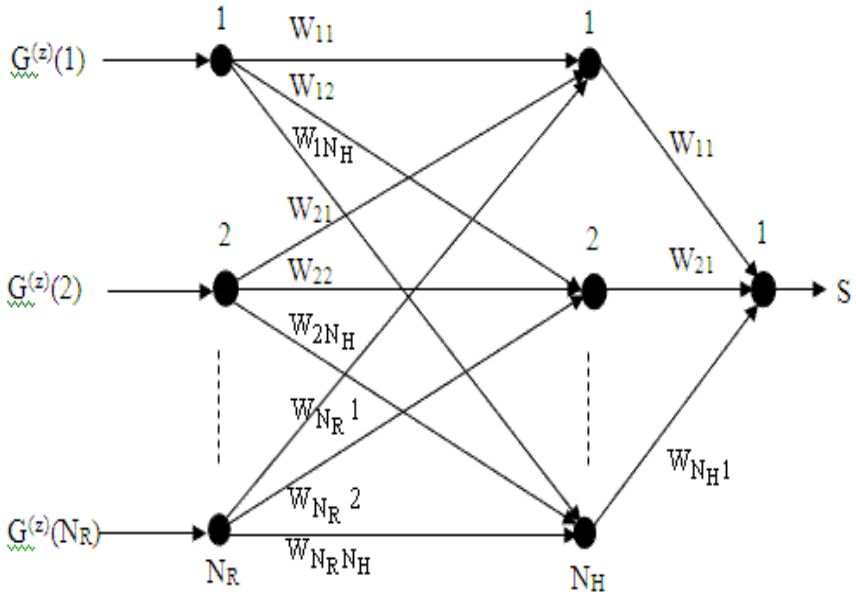



Figure 4. The multilayer feed forward neural network used in the proposed technique

$$h_q^{(1)} = \alpha + \sum_{r=1}^{N_R} W_{rq} G(r);$$
$$r = 1,2,\Lambda\ , N_R,\ q = 1,2,\Lambda\ , N_H \qquad (8)$$

$$h^{(2)} = \frac{1-e^{-h^{(1)}}}{1-e^{-2h^{(1)}}} \qquad (9)$$

$$S = h^{(2)} \qquad (10)$$

where, Eq. (8) is the basis function for the input layer and Eq. (9) and (10) are the activation functions for hidden and output layer, respectively.

1. Determine BP error using

$$e = \frac{1}{N_G} \sum_{p=0}^{N_G - 1} \left(S_T - S_p\right) \qquad (11)$$

where, $e$ is the BP error,
$S_T$ is the target output

2. By adjusting the weights of all the neurons based on the determined BP error, obtain new weights using

$$W^{new} = W^{old} + \Delta W \quad (12)$$

In Eq. (12), the weight to be changed $\Delta W$ depends on the rate of network learning $\gamma$ and the obtained network output $S_p$ for the $p^{th}$ gene sequence and it is determined using the formula $\Delta W = \gamma . S_p . e$.

3. Until the BP error gets minimized to a minimum extent, repeat the process from step 2. The termination criterion for practical cases, is $e < 0.1$.

*2) Classification Phase*

In the classification phase, the network finds the class of a given or test gene sequence and determines the species to which it belongs. The same processes performed on the training sequence are repeated for the test sequence . Using the mined patterns and their support, the constructive nucleotide dataset is generated. Subsequent to dimensionality reduction of the generated dataset they are tested in a neural network. The neural network decides the class of the species to which the gene sequence belongs.

## IV. Implementation Results

The proposed technique is implemented in the working platform of MATLAB (version 7.10) and the technique is evaluated using the DNA sequence of two different organisms, Brucella Suis and Caenorhabditis Elegans (C. Elegans). The evaluation process is performed using 10-fold cross validation test. Here, nucleotide patterns are mined with $L_{\max} = 5$. The nucleotide patterns for $l = 2 \; and \; 3$ and their corresponding support are given in Table III. In Fig. 5, different length patterns and their support are depicted and the constructive dataset that is generated from the pattern set is given in Table IV.

TABLE III. Mined nucleotide patterns from the DNA sequence of Brucella Suis and C.elegans (a) $l = 2$ and (b) $l = 3$ (a part of the pattern is given)

(a)

| Species | | | |
|---|---|---|---|
| S. No | Pattern | *Support* | |
| | | *Brucella suis* | *C-elegans* |
| 1 | aa | 169042 | 168149 |
| 2 | ag | 56284 | 59645 |
| 3 | ac | 53509 | 54824 |
| 4 | at | 100894 | 101778 |
| 5 | ga | 72354 | 72651 |
| 6 | gg | 45341 | 45368 |
| 7 | gc | 46001 | 43023 |
| 8 | gt | 53423 | 56002 |
| 9 | ca | 67662 | 69882 |
| 10 | cg | 47630 | 40344 |
| 11 | cc | 47205 | 44205 |
| 12 | ct | 57377 | 58316 |
| 13 | ta | 70670 | 73713 |
| 14 | tg | 67864 | 71687 |
| 15 | tc | 73159 | 70695 |
| 16 | tt | 171584 | 169717 |

(b)

| Species | | | |
|---|---|---|---|
| S. No | Pattern | *Support* | |
| | | *Brucella suis* | *C-elegans* |
| 1 | aaa | 86090 | 83349 |
| 2 | aag | 17627 | 18623 |
| 3 | aac | 18374 | 19422 |
| 4 | aat | 46951 | 46755 |
| 5 | aga | 19089 | 19871 |
| 6 | agg | 11100 | 10442 |
| 7 | agc | 11318 | 12264 |
| 8 | agt | 14777 | 17068 |
| 9 | aca | 17239 | 17901 |
| 10 | acg | 10577 | 9453 |
| 11 | acc | 10491 | 10676 |
| 12 | act | 15202 | 16794 |
| 13 | ata | 20142 | 21368 |
| 14 | atg | 15783 | 16752 |
| 15 | atc | 17406 | 16451 |
| 16 | att | 47563 | 47207 |

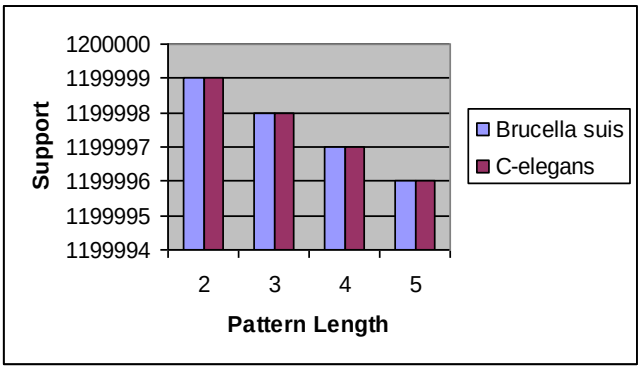



Figure 5. Support obtained for different length patterns

TABLE IV. CONSTRUCTIVE DATASET GENERATED FROM THE MINED NUCLEOTIDE PATTERNS (A) $l = 2$ AND (B) $l = 3$ (A PART OF THE PATTERN IS GIVEN).

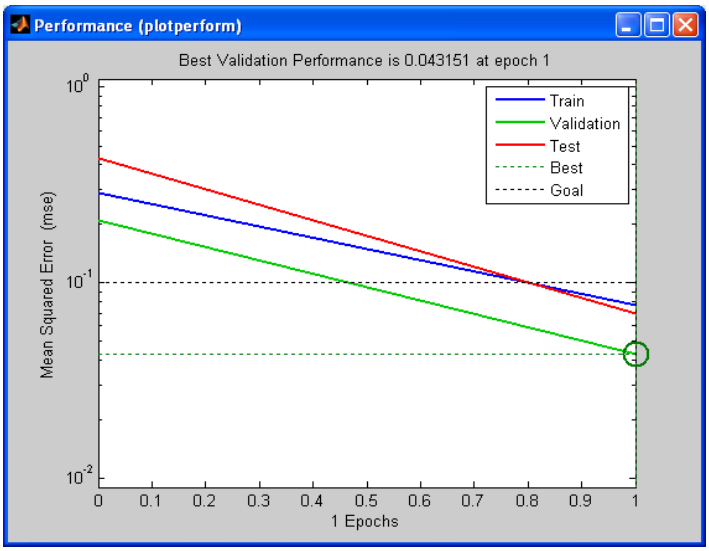
(A)

| S.No | | Species | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Brucella suis* | | | | *C-elegans* | | | |
| | | ***A*** | ***G*** | ***C*** | ***T*** | ***A*** | ***G*** | ***C*** | ***T*** |
| 1 | **A** | 169042 | 56284 | 53509 | 100894 | 168149 | 59645 | 54824 | 101778 |
| 2 | **G** | 72354 | 45341 | 46001 | 53423 | 72651 | 45368 | 43023 | 56002 |
| 3 | **C** | 67662 | 47630 | 47205 | 57377 | 69882 | 40344 | 44205 | 58316 |
| 4 | **T** | 70670 | 67864 | 73159 | 171584 | 73713 | 71687 | 70695 | 169717 |

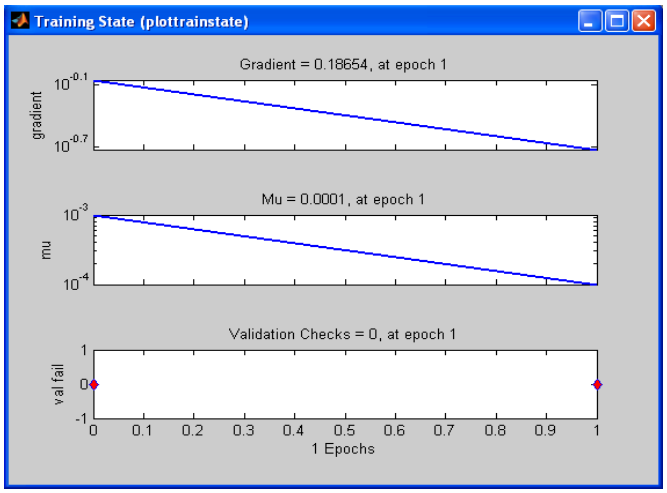
(B)

| S. No | | Species | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Brucella suis* | | | | *C-elegans* | | | |
| | | ***A*** | ***G*** | ***C*** | ***T*** | ***A*** | ***G*** | ***C*** | ***T*** |
| 1 | **AA** | 86090 | 17627 | 18374 | 46951 | 83349 | 18623 | 19422 | 46755 |
| 2 | **AG** | 19089 | 11100 | 11318 | 14777 | 19871 | 10442 | 12264 | 17068 |
| 3 | **AC** | 17239 | 10577 | 10491 | 15202 | 17901 | 9453 | 10676 | 16794 |
| 4 | **AT** | 20142 | 15783 | 17406 | 47563 | 21368 | 16752 | 16451 | 47207 |
| 5 | **GA** | 31140 | 13379 | 10257 | 17578 | 31355 | 13670 | 10422 | 17204 |
| 6 | **GG** | 14729 | 8619 | 11555 | 10438 | 14982 | 9275 | 9700 | 11411 |
| 7 | **GC** | 13025 | 9601 | 12066 | 11309 | 13656 | 7494 | 9646 | 12227 |
| 8 | **GT** | 12136 | 12780 | 10458 | 18049 | 12976 | 12650 | 10000 | 20376 |
| 9 | **CA** | 26644 | 12400 | 12801 | 15817 | 27422 | 13532 | 12501 | 16427 |
| 10 | **CG** | 16283 | 10873 | 9652 | 10822 | 13747 | 9848 | 7594 | 9155 |
| 11 | **CC** | 15108 | 11450 | 9503 | 11144 | 15133 | 9631 | 9315 | 10126 |
| 12 | **CT** | 13210 | 12537 | 13724 | 17906 | 12775 | 13648 | 13226 | 18667 |
| 13 | **TA** | 25167 | 12878 | 12077 | 20548 | 26023 | 13819 | 12479 | 21392 |
| 14 | **TG** | 22253 | 14749 | 13476 | 17386 | 24051 | 15803 | 13465 | 18368 |
| 15 | **TC** | 22290 | 16002 | 15145 | 19722 | 23192 | 13766 | 14568 | 19169 |
| 16 | **TT** | 25182 | 26764 | 31571 | 88066 | 26594 | 28637 | 31018 | 83467 |

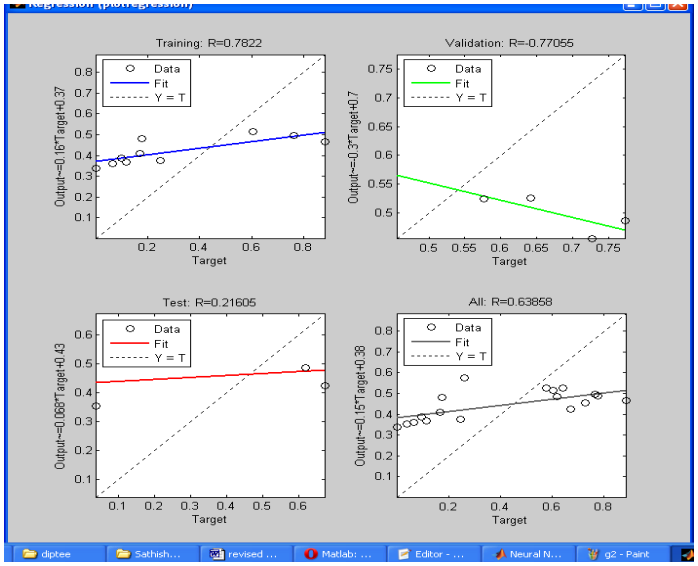
The pattern data and constructive dataset given in Tables III and IV are generated from one of the ten folds of gene sequence of Brucella Suis. Thus, from all the ten folds of gene sequence of both Brucella Suis and C. Elegans, the pattern data have been mined and constructive datasets have been generated. The generated ten folds of data are used to train the neural network. The results obtained from network training are given in Fig. 5.




(a)




(b)




(c)

Figure 6. Performance of training and test results from ANN: (a) Network performance, (b) Training evaluation and (c) Regression analysis.

Once the training process has been completed, the technique is validated using the test sequence. The results obtained from 10-fold cross validation are given in Table VI.

TABLE V. PERFORMANCE EVALUATION USING 10-FOLD CROSS VALIDATION RESULTS

| **Rounds in cross validation** | **Species** | | | |
|---|---|---|---|---|
| | ***Brucella suis*** | | ***C-elegans*** | |
| | ***ANN Output*** | ***Classification Result*** | ***ANN Output*** | ***Classification Result*** |
| 1 | 0.2421 | TP | 0. 5171 | TP |
| 2 | 0.0769 | TP | 0. 6272 | TP |
| 3 | 0.0828 | TP | 0. 6361 | TP |
| 4 | 0.2634 | TP | 0. 8974 | TP |
| 5 | 0.2493 | TP | 0. 6063 | TP |
| 6 | 0.2613 | TP | 0. 0141 | TN |
| 7 | 0.5277 | TN | 0. 9163 | TP |
| 8 | 0.3616 | TP | 0. 6714 | TP |
| 9 | 0.5849 | TN | 0. 5103 | TP |
| 10 | 0.2143 | TP | 0. 5142 | TP |
| **Mean Classification Accuracy** | 80% | | 90% | |

From the results, it can be seen that when a gene sequence is given to the proposed technique it identifies the corresponding species. Here, the technique is evaluated with the DNA sequence of only two genes. The technique is developed in such a way that it can be applied to any kind of DNA sequence. The test results claim that the performance of the technique reaches a satisfactory level.

## V. CONCLUSION

In this paper, we have proposed a species identification technique by integrating data mining technique with artificial intelligence. Initially, the nucleotide patterns have been mined effectively. The resultant has been subjected to MPCA-based dimensionality reduction and eventually classified using a well-trained neural network. The implementation results have shown that the proposed technique effectively identifies the organism from its gene sequence and so the species. Moreover, results obtained from 10-fold cross validation have proved that the organism can be identified even from a part of the DNA sequence.

Though the technique has been tested with the DNA sequence of only two organisms, the 10-fold cross validation results have reached a remarkable performance level. From the results, it can be hypothetically analyzed that a technique, which identifies the organism only with a part of gene sequence, have the ability to classify any kind of organism and so the species.

## REFERENCES

[1] P. P. Vaidyanathan and Byung-Jun Yoon, "The role of signal-processing concepts in genomics and proteomics", Journal of the Franklin Institute Genomics, Signal Processing, and Statistics Vo. 341, Issue. 1-2, pp. 111-135, January-March 2004.

[2] Achuth Sankar S. Nair, T.Mahalakshmi " Visualization Of Genomic Data Using Inter-Nucleotide Distance Signals", Silico Biology, Issue Volume 6 , 215-222, March ,2006

[3] Edward R. Dougherty, Ilya Shmulevich and Michael L. Bittner, "Genomic Signal Processing: The Salient Issues", EURASIP Journal on Applied Signal Processing, Vol. 1, pp. 146–153, 2004.

[4] Michel Tibayrenc "The species concept in parasites and other pathogens: a pragmatic approach?" TRENDS in Parasitology Vol.22 No.2 February 2006. http://www.th.ird.fr/downloads/TIP.pdf

[5] Swapnoneel Roy, Minhazur Rahman, and Ashok Kumar Thakur "Sorting Primitives and Genome Rearrangement in Bioinformatics: A Unified Perspective", World Academy of Science, Engineering and Technology, Issue 38, 2008.

[6] Mai S. Mabrouk, Nahed H. Solouma, Abou-Bakr M. Youssef, and Yasser M. Kadah "Eukaryotic Gene Prediction by an Investigation of Nonlinear Dynamical Modeling Techniques on EIIP Coded Sequences", International Journal of Biological and Life Sciences 3:4 2007

[7] Jianbo Gao, Yan Qi, Yinhe Cao, and Wen-wen Tung, "Protein Coding Sequence Identification by Simultaneously Characterizing the Periodic and Random Features of DNA Sequences", Journal of Biomedicine and Biotechnology, Vol. 2, pp. 139–146, 2005.

[8] C.L. Winder1, E. Carr2, R. Goodacre1 and R. Seviour2 "The rapid identification of Acinetobacter species using Fourier transform infrared spectroscopy" Journal of Applied Microbiology 96, 328–339,2004

[9] Francielle B. Silva, Sabrina N. Vieira, Luiz R. Goulart Filho, Julien F. C. Boodts , Ana G. Brito-Madurro and João M. Madurro "Electrochemical Investigation of Oligonucleotide-DNA Hybridization on Poly(4-Methoxyphenethylamine)",International Journal of Molecular Sciences, 9, 1173-1188,8-july,2008.

[10] Anne Jensen,Guillaume Calvayrac, Bensu Karahalil, Vilhelm A. Bohr and Tinna Stevnsner "Mammalian 8-Oxoguanine DNA Glycosylase 1 Incises 8-Oxoadenine Opposite Cytosine in Nuclei and Mitochondria, while a Different Glycosylase Incises 8-Oxoadenine Opposite Guanine in Nuclei", The journal of biological chemistry, Vol. 278, 19541-19548, 2003, DOI:10.1074/jbc.M301504200

[11] Jeremy D. Volkening, Stephen J. Spatz "Purification of DNA from the cell-associated herpesvirus Marek's disease virus for 454 pyrosequencing using micrococcal nuclease digestion and polyethylene glycol precipitation", Journal of Virological Methods, Vol. 157, p.p. 55–61. 2009.

[12] Edward R. Dougherty and Aniruddha Datta "Genomic Signal Processing: Diagnosis and Therapy", IEEE Signal Processing Magazine, Vol. 22, No. 1, p.p. 107-112, 2005, DOI: 10.1109/MSP.2005.1407722

[13] TrevorW. Fox,Alex Carreira "A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression" ,EURASIP Journal on Applied Signal Processing, Vol.1, p.p. 108–114, 2004

[14] A. Bharathi, Dr.A.M.Natarajan, "Cancer Classification of Bioinformatics data using ANOVA" International Journal of Computer Theory and Engineering, Vol. 2, No. 3, June, 2010

[15] Jayanthi Ranjan "Applications of Data Mining Techniques In Pharmaceutical Industry", Journal of Theoretical and Applied Information Technology, 2005 – 2007.

[16] Paul Hooley, Ian J. Chilton, Daron A.Fincham, Alan T.Burns and Michael P.Whitehead "Assigning Level in Data-mining Exercises", Bioscience Education Journal, Volume 9:June 2007

[17] Antony Brownea , Brian D. Hudsonb , David C. Whitleyb ,Martyn G. Fordb , Philip Pictonc "Biological data mining with neural networks:implementation and application of a flexible decision tree extraction algorithm to genomic problem domains" Neurocomputing, Volume 57, March 2004, Pages 275-293

[18] Gerd Pfeiffer, Stefan Baumgart, Jan Schr¨oder, and Manfred Schimmler "A Massively Parallel Architecture for Bioinformatics", Lecture Notes in Computer Science, 2009, Volume 5544/2009, p.p. 994-1003, DOI: 10.1007/978-3-642-01970-8_100

[19] Simon Miles, "Agent-Oriented Data Curation in Bioinformatics", In Workshop on Multi-Agent Systems in Medicine, Computational Biology, and Bioinformatics (MAS*BioMed'05), July 2005, Utrecht, Netherlands
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.124.2202&rep=rep1&type=pdf

[20] Dhaeseleer P, Liang S, Somogyi R, "Genetic network inference: from co-expression clustering to reverse engineering", Bioinformatics, Vol. 16, No. 8, pp.707—726, 2000.

[21] Kirschner M, Pujol G, Radu A, "Oligonucleotide microarray data mining: search for age-dependent gene expression", Biochemical and Biophysical Research Communications, Vol. 298, No. 5, pp. 772—778, 2002.

[22] Ponomarenko J, Merkulova T, Orlova G, Fokin O, Gorshkov E, Ponomarenko M, "Mining DNA sequences to predict sites which mutations cause genetic diseases", Knowl-based Syst, Vol. 15, No. 4, pp.225—233, 2002.

[23] Oliveira and Johnston, "Mining the schistosome DNA sequence database", Trends Parasitol, Vol. 17, No. 10, pp.501—503, 2001.

[24] Fuhrman, Cunningham, Wen, Zweiger, Seilhamer and Somogyi, "The application of Shannon entropy in the identification of putative drug targets", Biosystems, Vol. 55, pp.5—14, 2000.

[25] Arkin, Shen and Ross, "A test case of correlation metric construction of a reaction pathway from measurements", Science, Vol. 277, pp. 1275-1279, 1997.

[26] Cho and Won, "Machine learning in DNA Microarray analysis for cancer classification. In: Yi-Ping Phoebe Chen, in proceedings of the First Asia-Pacific Bioinformatics Conference. Australian Computer Society, pp. 189-198, 2003

[27] Riccardo Bellazzi and Blaz Zupan, "Towards knowledge-based gene expression data mining", Journal of Biomedical Informatics, Vol.40, pp.787-802, 2007

[28] Shital Shah and Andrew Kusiak, "Cancer gene search with data-mining and genetic algorithms", Computers in Biology and Medicine, Vol.37, pp.251-261, 2007

[29] Hemalatha and Vivekanandan, "Genetic Algorithm Based Probabilistic Motif Discovery in Unaligned Biological Sequences", Journal of Computer science, Vol.4, No.8, pp.625-630, 2008

[30] Ashraf S. Hussein, "Analysis and Visualization of Gene Expressions and Protein Structures", Journal of software, Vol.3, No.7, pp.2-11, October 2008

[31] Ahmad M. Sarhan, "Cancer Classification Based on Microarray Gene Expression Data Using DCT and Ann", Journal of Theoretical and Applied Information Technology, Vol.6, No.2, pp.208-216, 2009

[32] Valarmathie, Srinath, Ravichandran and Dinakaran, "Hybrid Fuzzy C-Means Clustering Technique for Gene Expression Data", International Journal of Research and Reviews in Applied Sciences, Vol.1, No.1, pp.33-37, October 2009

[33] Belmamoune, Potikanond and Fons J.Verbeek, "Mining and analysing spatio-temporal patterns of gene expression in an integrative database framework", Journal of Integrative Bioinformatics, Vol.7, No.3, pp.1-10, 2010

[34] Hans knutsson, "A Tensor representation of 3-D structure", 5th IEEE-ASSP and EURASIP Workshop on Multidimensional Signal Processing, The Netherlands, September, 1987.

[35] Haiping Lu Plataniotis, K.N. Venetsanopoulos, A.N., MPCA: Multilinear Principal Component Analysis of Tensor Objects, IEEE Transactions on Neural Networks, Vol.19 No.1, p.p. 18 – 39, 2008, ISSN: 1045-9227, DOI: 10.1109/TNN.2007.901277

[36] Emilio Corchado, Álvaro Herrero, " Neural Visualization of network traffic data for intrusion detection", Applied Soft Computing, Volume 11, Issue 2, March 2011, Pages 2042‑2056

[37] E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun**,** "The application of datamining techniques in financial fraud detection: A classification framework and an academic review of literature" Decision Support Systems, Volume 50, Issue 3, February 2011, Pages 559‑569

[38] Arzu Şencan Şahin, İsmail İlke Köse & Reşat Selba, "Comparative analysis of neural network and neuro – fuzzy system for thermodynamic properties of refrigerants" Applied Artificial Intelligence: An International Journal, Volume 26, Issue 7, 2012, DOI: 10.1080/08839514.2012.701427

# Brainstorming 2.0: Toward collaborative tool based on social networks

MohamedChrayah [1],Kamal Eddine El Kadiri[1], Boubker Sbihi[1],Noura Aknin[2]
[1]Laboratory LIROSA, [2]Laboratory LaSIT
Faculty of sciences
Tétouan, Morocco

***Abstract*— Social networks are part of Web 2.0 collaborative tools that have a major impact in enriching the sharing and communication enabling a maximum of collaboration and innovation globally between web users. It is in this context that this article is positioned to be part of a series of scientific research conducted by our research team and that mixes social networks and collaborative decision making on the net. It aims to provide a new tool open source for solving various social problems posed by users in a collaborative 2.0 based on the technique for generating ideas, brainstorming method and social networks together for the maximum possible adequate profiles to the virtual brainstorming session. A tool is run by a user called expert accompanied by a number of users called validators to drive the process of extracting ideas to the loan of various users of the net. It offers then the solution to the problem of sending a satisfaction questionnaire administered by an expert ready for the affected user to measure the level of his satisfaction and also the success of the process launched. For its implementation, we propose a unified modeling using UML language, followed by a realization using the JAVA language.**

***Keywords- component: Web2.0, brainstorming, social networks, UML.***

## I. INTRODUCTION

Over time, the web has seen many changes starting with the static web which allowed only to display static pages made by the directors of the net and whose content was not always updated followed by collaborative Web 2.0 who proposed the involvement of users in content creation. Web 2.0 was proposed in August 2001 by Dale Dougherty of O'Reilly Media, but the real release of concept 2.0 was published in an international conference in 2005 by Tim O'Reilly [10]. He proposed a new vision of the web which consists on a higher participation of Internet users as producers of information thus forming communities participating in the communication, sharing and dissemination of information.

With this concept a lot of software and services are freely available on the web and therefore the amount of information has increased which encouraged users to participate and inter exchange. Social networks have existed since 2003, where they have grown exponentially up to date [7]. They collect data on members, and then store this information as data profiles, these sites represent an appropriate database to search for suitable profiles to any operation or survey in the web. Moreover, decision-making has changed a lot with the emergence of information and communication technologies (ICT) [6]. Makers become less statically located; on the contrary, they act in a distributed manner. This change creates a new set of requirements: collaborative decision-making based on collaboration using Web 2.0 tools.

In the next section, we present the web 2.0; its principles, its most used tools and especially social networking the tool used in our article, and then we'll present the notion of collaborative decision based on the method of generating ideas (brainstorming) and social networks to achieve in the end a collaborative decision as a result of a series of proposals and virtual meeting by web users 2.0. Finally, we'll propose a design and implementation of the proposed tool using the UML.

## II. WEB 2.0

### A. *Web 2.0 and its dimensions*

Web 2.0 is social, is open, it lets you control your data, mixing the global with the local. Web 2.0 is new interfaces - new ways to search and access content. Web 2.0 is a platform ready to receive the educators, the media, politics, and communities. Thus, users who contribute to information exchange can simply interact (share, exchange, etc.) with both the content and page structure, but also between them, including creating the social Web. The user becomes, using the tools at its disposal, an active person of the cloth [9].

Web 2.0 can be viewed in three dimensions as shown in the following figure:

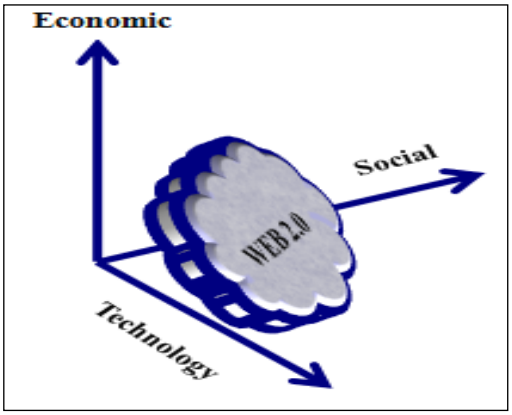



Figure 1: Dimensions Of Web 2.0.

Social dimension: Web 2.0 is a real network of social interaction based on the participation of Internet users. User communities are created in this context based on areas of mutual benefit. Anyone can easily create an information space accessible by anyone and anywhere in the world where he can put anything (anything is the one of the boundaries of Web 2.0).

Technical dimension: Web 2.0 is an advanced technique that makes it simple to access the production and use of information through the tilting of the software installation to use online services. Thanks to the use of multiple technologies (XHTML, CSS and JavaScript for the presentation of the sites, DOM, Document Object Model, for dynamic and interactive signage, XML and XSLT for data manipulation).

Economic dimension: Funding Web 2.0 sites is done mainly through advertisements, commercial offers and trafficking networks instead of gifts or payments for licenses to use proprietary software. A project based on an economic model of Web 2.0 is based on the large mass of users who consume information mixed with advertising or commercial content which finances containers [2].

### *B. The Tools Of Web 2.0*

Web 2.0 consists of a set of Internet technologies that facilitate the open and participatory work. Its main feature is that Web 2.0 tools allow users to control the network and interact proactively to improve or transform situations that affect them.

#### *1) Blog:*

The term "blog" is short for weblog, which can be translated as "Internet newspaper". Frequently defined as a personal site, this is an individual space of expression, created to give voice to all Internet users (individuals, businesses, artists, politicians, associations ...). Blogs are extremely simplified sites and dedicated to writing, where "the entries appear in ante-chronological order." The animation of blogs is initially limited to technophiles capable of creating the structure of their blog and have it hosted on a server.

The Pew Research Center's Internet and American Life Project has conducted a survey in 2008 which has resulted in startling statistics that 40% of adult Internet users in the United States have blogs.

Blogs have been discussed recently as a innovative knowledge of sharing technology, knowledge and management [12].

#### *2) Wiki:*

The term is derived from the expression Hawaiian wiki-wiki, meaning quick. A wiki is a collaborative website whose content can be edited by visitors on the site, allowing users to easily create and edit collaborative web pages [8]. In essence, a wiki is a simplification of the process of creating HTML web pages combined with a system that records each individual change that occurs over time, so that at any time a page can be forced to return to the one of its previous states. A wiki can also provide tools that allow the users community to monitor the changing state of the wiki and discuss issues that arise. Some wikis restrict access to a group of members, allowing only members to change the page content but everyone can see it. Others allow unrestricted access, allowing anyone to both modify the content and display.

#### *3) Social Networks:*

A social network is a set of social entities such as individuals or social organizations linked together by bonds created during social interactions.

It is represented by a structure or a dynamic form of a social group it's a web space to:

- Express them selves
- To promote itself
- Exchange
- Get back in touch

Social networks have as common basis the sharing:

- Sharing knowledge
- Sharing of professional contacts
- Content Sharing

Social networks are social websites that enable people to form online communities and share content created by these users. People can be users of the open Internet or restricted to those who belong to a particular organization (eg company, university, etc.). [14]

Table 1 gives an idea of the number of users of social networks, and the classification of these social sites depending on the number of participants.

TABLE I. SOCIAL NETWORKS USED MOST [15]

| *Site Name* | *Users (in Million)* |
|---|---|
| Facebook | 309 |
| MySpace | 253 |
| WindowsLiveSpaces | 120 |
| Habbo | 117 |
| Friendster | 90 |
| Hi5 | 80 |
| Tagged | 70 |
| Orkut | 67 |
| Flixter | 63 |

As the table shows, Facebook is positioning itself in first place with 309 million users. The implication of this number in a decision-making would be a dream for designers of collaborative Web 2.0

#### *4) RSS Feeds:*

RSS (Really Simple Syndication) is a simple XML syntax to describe the recent additions of content to a website. These additions can include elements of news, blog updates, library acquisitions or any other information. it just facilitates dynamic sharing of content between a publisher (website) and a reader (the Internet) by allowing authors and editors of a website to make available to the community some content that can be reused for integration into another site [13]. Since RSS uses XML to disseminate information relevant to user needs, RSS could well become the universal method for extracting information from the Internet.

## III. BRAINSTORMING 2.0:COLLABORATIVE DECISION MAKING

### A. Brainstorming Method:

Brainstorming is a technique for generating ideas that stimulates creative thinking in finding solutions to a given problem. This is to produce as many ideas as possible in the shortest time on a given topic without criticism, without judgment. This searching method favors quantity, spontaneity and imagination [11].

Table II shows the essential steps for a successful brainstorming session:

TABLE II. STEPS BRAINSTORMING

| *Steps* | *Sequence* |
|---|---|
| Step 1 | Presentation of the problem |
| Step 2 | warming period |
| Step 3 | Brainstorming |
| Step 4 | classification / grouping of ideas |
| Step 5 | Final Decision |

As shown in Figure 3, the Collaborative Decision Making Brainstorming consists of four iterative steps starting with the presentation of the problem and ending with the classification of ideas.

### B. Brainstorming2.0:

#### 1) Towards a tool for collaborative decision making:

The collaborative decision making is based on user's participation as actors for the production and wide dissemination of the decision subsequently forming communities. The size and mass of decisions will increase the quantitative level but still the qualitative decisions suggested by users.

The users of the system must be ordered according to their importance and give more privileges. It is not enough to give any decision, but the right decision, for this we proposed Brainstorming 2.0 tool to overcome the problems already discussed above.

Brainstorming 2.0 is an Open Source, free dedicated to all users of Web 2.0; their goal is to found a topic concerning decisions. It is not as tools publisher's social owner as Google Plus (G+), which is a social network where there are paid services, and their only purpose the meeting between friend or the professional. Brainstorming 2.0 is a social tool that organizes virtual brainstorming sessions between users of the web communities.

#### 2) Classification Decisions:

The decision generated in the brainstorming 2.0 system can be classified into four classes [4][3][5][1] according to their quality:

TABLE III. CALASSIFICATION DECISIONS

| *Code* | *libel* | *Weighting* |
|---|---|---|
| G | Good | 10 |
| M | Meduim | 5 |
| L | Low | 1 |
| E | Error | -10 |

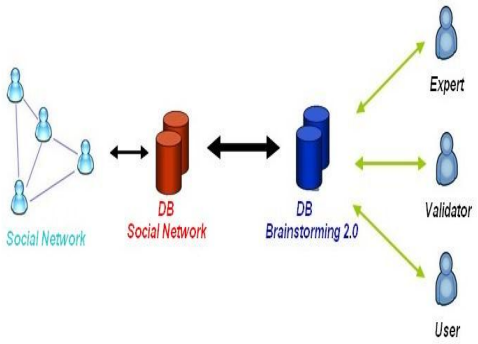



Figure2. Brainstorming 2.0 Architecture

According to the above table the user who produces the right decisions and averages has the chance to become a validator in a short time; it also proposes to create a virtual currency that will increase every time someone publishes validated decisions. Users will then have access to some opportunity not given to all others. This value will depend on the turnout of participation in the virtual brainstorming sessions and also the notation affected by the validators.

#### 3) Classification Of Users:

It is proposed to decompose the system users into three groups: simple users who consume and produce decisions, the validators who validate decisions and finally the experts who make tracking validations, pointing validators for each problem and its publication when validated by the validators.

A simple user can become a validator if its weight exceeds 1000 pt and is recommended by an expert, a validator must communicate with other validators and expert in the validation process. A validator can become an expert if its weight exceeds 10000 pt. Brainstorming 2.0 users are represented in the following table [4][3][5][1]:

TABLE IV. CLASSEMENT DES UTILISATEURS

| *Code* | *Libel* | *Weighting* |
|---|---|---|
| E | EXPERT | >=10000 |
| V | VALIDATOR | >10000And >=1000 |
| U | USER | <=1000 |

Regarding the weighting at each decision by a user there is an increase in the value using the following formula (1):

$$P(User)=\Sigma P(Decisions) \quad (1)$$

The following figure shows the use case diagram tool braistorming 2.0

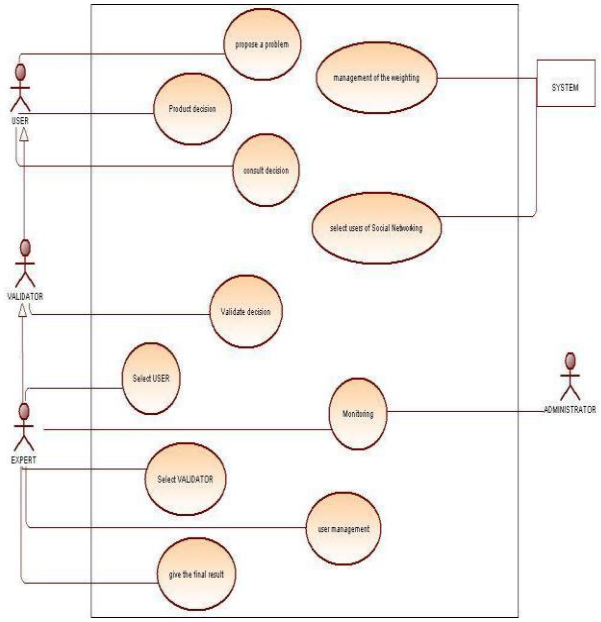



Figure3 Uses Cases Brainstorming 2.0.

The use case diagram shown in Fig.3 clarified the interaction between the system and all actors (users), it is clear from the diagram that the expert is the main user of the system, he inherits all the actions of other users in addition to managing the flow of information he can administrate the platform. The actor system is who is responsible to the updating of weights and linking with social networking to select the appropriate profiles.

*4) Decision Making Process In The 2.0 Brainstorming:*

The process of decision making in the Brainstorming 2.0 runs as follows, figure 4.

A user poses a problem, a system expert considers the issue and distributes it to all users with an adequate profile to this subject by searching social networks related to the system, and in parallel the expert selects validators to validate decisions proposed by them. After treatment and decisions classification by the validators, the expert groups the decisions that have obtained Class B (Good) by the validators to generate a final decision of the issue to be released in the portal. Finally a satisfaction questionnaire is sent to the user who submitted the problem to get an idea about his satisfaction with the solution proposed.

The figure.5 gives us an idea about the chronological interactions between all major and minor players in the system, from the diagram it is the simple user who initiates the transaction, the one who started the virtual brainstorming session, then after the analysis of the issue the expert send it to other users to give these decisions and in the end transfer them to the validators.

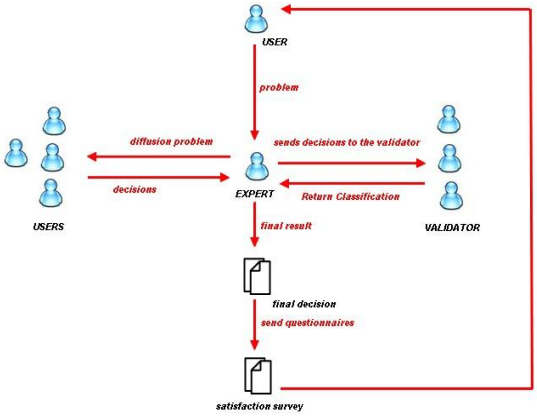



Figure4 Process Brainstorming 2.0.

But the sequence diagram is insufficient to give us an overview of the system, that is why we use the class diagram to understand the relationship between the classes of the system, the Fig6 represents the class diagram of the brainstorming 2.0 system; According to the diagram the system is composed of 9 classes, Expert class inherits the methods of the Validator class that also inherits from the User class, the class Profil_Network_Social contains information on existing profiles in social networks interacting with the system.

It is clear that the class Decision_final is related directly to the Expert class as the one who distributes the final decision, Decision_Validate class with the class validator and finally the classes Decision and Problem with the USER class, because he's the one who launches the problem and also the one who gives decisions.

## IV. EXAMPLE : A CASE STUDY.

An expert receives a user's question "What is your opinion on Web 2.0?" And then he launches it into the system, after consultation with user's six users answered the question with judgment:

- User1 : Favourable opinion
- User2 : Unfavourable opinion
- User3 : Favourable opinion
- User4 : Favourable opinion
- User5 : Unfavourable opinion
- User6 : Favourable opinion

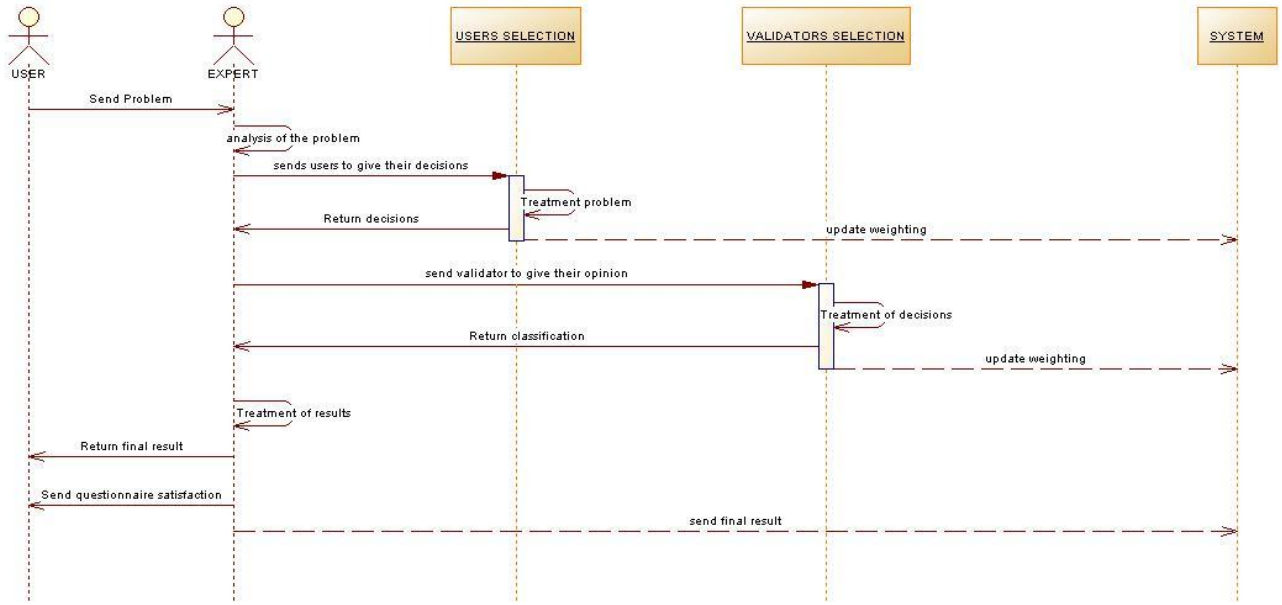



Figure5 Sequence Brainstorming 2.0 Diagram

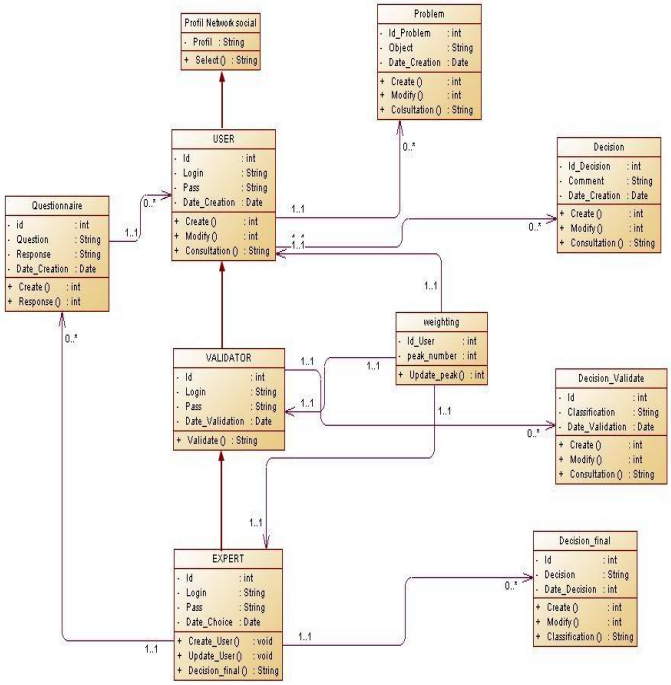



Figure6 Class Brainstorming 2.0 Diagram

The expert selects two system validators to review and validate user responses by giving them a ranking:

TABLE V. RANKING ANSWERS

| | ***Validator1*** | ***Validator2*** |
|---|---|---|
| User1 | G | G |
| User2 | L | L |
| User3 | G | G |
| User4 | G | G |
| User5 | L | L |
| User6 | G | G |

After ranking the responses, the expert selects the response that received a Class B by both validators and diffuses it on the portal (Favourable Opinion) with the judgment of users who responded favorably, at the end the expert sends a questionnaire to the user that sent the question to get an idea of his level of satisfaction with the answer.If the user is satisfied, it will make him seek help and become part of the community of this tool and if it is not satisfied, he can restart another process or ask for help from People experts in the wanted domain.What is clear is that decisions vary from personal context to professional staff. To improve performance in a professional context, it might be thought to pay employees and to create training sessions to achieve meaningful results at the expense of a sum of money to take advantage of collaborative network intelligence.

## CONCLUSION

The aim of our proposal is to find a mechanism mixing between the benefits of Web 2.0 tools and the technique for generating ideas brainstorming in order to achieve a system of collaborative decision making. This new system will help find lot of solutions through their connection with social networks, which contains adequate profiles and also good decisions because the raw information has no value in the new system and as the one who adopts the problem is an expert. The limitations of this tool are that it contradicts the general concept of Web 2.0 (the participation of everyone in the decision) since the expert takes some decision. There is also the responsibility of the validators in the selection and classification decisions. In addition, it is necessary to test and

measure the level of satisfaction undefined users to meet the needs of each context.

This article is a beginning of a series of articles that will follow and that will be the implementation of this tool in java. A presentation detailed the tool will be made in future publications

## PERSPECTIVE

As prospects we propose to generalize the use of this tool and measure the rate of satisfaction of its users. We propose also to design tools for collaborative semantic decision support which understands the sense of the decision. Adapt this tool to other collaborative tools are blogs, wikis and RSS feeds. Impose a single and secure identification with fingerprints to ensure good use away from hackers and malicious people.

## REFERENCES

[1] Boubker Sbihi, K.E. El Kadiri, and N. Aknin, "The Vblogs: Towards a New Generation of Blogs", International Journal of Computer Science Issues, Vol. 7, issue 3, No. 2, May 2010, pp. 9-14.
[2] Boubker Sbihi, TOWARDS A NEW VISION OF WEB 2.0, Computer Sciences and Telecommunications, 2009
[3] Boubker Sbihi, and Kamal Eddine El Kadiri, "Web 2.2: Toward classified information on the Web", International Journal of Web Applications, Vol. 1, No. 2, June 2009, pp. 102- 109.
[4] Chrayah Mohamed, Kamal Eddine El Kadiri, Boubker Sbihi and Noura Aknin, Brainstorming +: Towards a tool for decision-making In the Web 2.0, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 2, March 2012
[5] Chrayah Mohamed, Boubker Sbihi and Kamal Eddine El Kadiri :Towards an instrument for a new generation of blog. EMSCE 2011 May 19_20 2011 ISBN:978-84-694-4025-4.
[6] Claus Rinnera,Carsten Keßlerb,Stephen Andrulisa, The use of Web2.0 concepts to support deliberation in spatial decision-making, Computers, Environment and Urban SystemsVolume 32, Issue 5, September 2008, Pages 386–395.
[7] Danah boyd, Nicole Ellison, Social Network Sites: Definition, History, and ScholarshipJournal of Computer-Mediated Communication Volume 13, Issue 1, pages 210–230, October 2007.
[8] Howard T. Welser ,Patrick Underwood, Wiki Networks: Connections of Creativity and Collaboration 2011, Pages 247–271.
[9] Musser J, O'Reilly T. O'Reilly Radar Report: Web 2.0 Principles and Best Practices: O'Reilly Media 2006.
[10] O'Reilly T. What Is Web 2.0? Design Patterns and Business Models
[11] for the Next Generation of Software; 2005.
[12] Patrick C.Shih, Gina Venolia, Gary M. Olson Brainstorming under constraints: Why software developers brainstorm in groups 2011 Pages 74-83
[13] Sangmi Chaia, Minkyun Kimb, What makes bloggers share knowledge? An investigation on the role of trust, International Journal of Information Management 30 (2010) 408–415
[14] Yu-Feng Lan,Yang-Siang Sie, Using RSS to support mobile learning based on media richness theory Volume 55, Issue 2, September 2010, Pages 723–732
[15] WonKim a, Ok-RanJeong a, Sang-WonLee On social Web sites Information Systems 35 (2010) 215–236

## AUTHORS PROFILE

**Chrayah Mohamed**is computer engineer, pHD student and themember of LIROSA laboratory.
**Kamal Eddine El Kadiri**is pHD doctor and professor of computerscience at Faculty of Sciences of Tétouan in Morocco. He is theDirector of the ENSA School of engineers of Tetouan and theDirector of LIROSA laboratory. He has published several articleson E-learning and Web 2.0. He is part of many boards ofinternational journals and international conferences.
**BoubkerSbihi**is pHD doctor and professor of computer scienceat the School of Information Science in morocco. He is theresponsible of Department of Information Management. He haspublished many articles on E-learning and Web 2.0. He is part ofmany boards of international journals and international conferences.
**NouraAknin**is pHD doctor and professor of computer science atFaculty of Sciences of Tétouan in Morocco. She has publishedmany articles on E-learning and Web 2.0. She is part of manyboards of international journals and international conferences. Shehas member of the IEEE and the IEEE Computer Society

# A Review On Cognitive Mismatch Between Computer and Information Technology And Physicians

Fozia Anwar
CIS Department
Universiti Teknologi Petronas
Perak, Malaysia

Dr. Suziah Sulaiman
CIS Department
Universiti Teknologi Petronas
Perak, Malaysia

Dr. P.D.D.Dominic
CIS Department
Universiti Teknologi Petronas
Perak, Malaysia

***Abstract*— Health Information Technology has a great potential to transform the existing health care systems by making them safe, effective and efficient. Multi-functionality and interoperability of health information systems are very important functions. Hence these features cannot be achieved without addressing the knowledge and skills of the health care personnel. There is a great mismatch between Information Technology knowledge and skills of physicians as this discipline is completely missing in their educational tenure. So usability of health information technologies and system as well as evidence based practice in the future can be improved by addressing this cognitive mismatch. This will result in persistent partnership in HIS design between physician and IT personnel to get maximum usibility of the systems,**

***Keywords- cognitive mismatch; HIT; usibility.***

## I. INTRODUCTION

Information technology in health sector is spreading globally [1]. There are major campaigns in many countries which are involved not only large expenditures but public resources as well to boost the use and adoption rate of health information systems (HIS) and technologies. Therefore for the successful usability of HIS it is critical to address physician's views and skills on the use of these technologies [2].

Use of health information technology is offering evidence-based practice to endorse health and human prosperity. Health information technology (HIT) consists of a set of technologies with a great diversion for transmitting and managing health care data for the use of all stakeholders of health care systems [3]. Therefore the development of user-centered information systems is important to get best usability and advantages of the developed systems. Development of the user centered information system is like a pendulum that has to be swung between users and developer. The success of the ICT system depends on the balance of the pendulum in the form of input and feedback from the both ends. In most of the cases of the HIS the balance of the pendulum is disturbed due to lack of information technology cognitive skills of physicians. Hence the usability of the particular healthcare information system could not be achieved at the extent of the expectations [4].

Health care software developers often neglect significant characteristics, tasks, user preferences, knowledge skills and usability issues from the physician point of view resulting in systems that decreases productivity or simply remain unusable for the end-users [4]. Users of healthcare information system are dynamic and hence their needs are, so, development of any solution, system, or technology should be accounted the needs and requirements of the end users [6] [7] [8].

Definition of the term Cognitive skills varies but American Psychological Association's (2007) define as "all forms of knowing and awareness, such as perceiving, conceiving, remembering, reasoning, judging, imagining and problem solving."[9] Simple definition of cognitive skills could be that these are the skills that one has learnt and ability to learn further new knowledge. Cognitive skills and cognitive ability of people had a profound effect in making a successful adoption and implementation of an information system with great usability.

The information technologies have greatly influenced on medical research, education, and healthcare delivery [10]. Therefore, medical schools should fit into place specialists to educate both students and tutors simultaneously; hence it is necessary to make additions and alterations in medical education accordingly [11].

## II. LITERATURE REVIEW

### *A. Role of Health Informatics*

A great philosopher Immanuel Kant said: "We see things not as they are but as we are". Perception and interpretation are having the same meaning. Anxiety of present situation having past experience is the perception. Both terms processing and perception are frequently used interchangeably. The etiology of misinterpretation is lack of experience or knowledge about the information technology.

Information plays a crucial role in medicine. All the time physicians are playing with the information they create, collect, search, adapt; in fact they drown in information. Medical or health informatics is all about this. The exact position of medical informatics is at the intersection of information technology, cognitive science, artificial intelligence, and medicine. So this is not a simple field involving only one aspect like medical computing, telecommunications, or information engineering, rather it is a dialogue between

physicians, patients, and medical informaticians in a medical information system. Rather it explores and develops new knowledge, builds new theories, and organizes principles and solutions. Today's challenge is not to have an access of hardware rather it is the ability to use the information system. HIT technologies and information systems develop the way to positively increase the outcome of clinical care. Computers and evidence decision makings are two pillars of health informatics. To take full advantage of the HIT we have to learn the skills for framing and analyzing and integrating the healthcare information [12].

It is needed that end-users especially physicians should participate in healthcare IT development process [13]. Different models are suggested for better collaboration between users and IT experts. Parmit K Chilana suggested a model which emphasizes the persistent partnership of healthcare domain experts and IT experts throughout the development, implementation, evaluation and deployment processes. It gives emphasis on improving interdisciplinary skills on both ends of the pendulum to achieve a good balance. As a result of which an ICT system can be made which is not only be able to support the goals of the complex domain but can clearly outline the challenges of clinical ICT system [14].

It is essential in the research field of health informatics that a good fit and faith must persist in an ICT system and clinical practices [13]. Nevertheless, few numbers of studies have been done to answer the question of the cognitive aspects and need of interdisciplinary skills of the users and IT experts of different domains and the impact of the trainings given to users of HIS. Terms HIT and HIS are used in the research literature to cover a wide range of information technology applications in health care.

Trend of IT applications use is increasing by the Health Professionals (physicians, nurses and allied Staff), as, some health professional also develops, deploy or research health care IT [15]. Consequently, Health Professionals need to educate themselves for their respective roles along with their temporal requirements in health informatics [16] as their cognitive skills about IT and IS are not addressed throughout their educational tenure. Some examples of IT use in healthcare sector are use of physician digital assistance (PDA's), computerized physician order entry (CPOE), electronic health record (EHR), clinical decision support system (CDSS), picture archiving and communication system (PACS), radiology information system (RIS), pharmacology information system (PIS), hospital information system (HIS), disease early warning system (DEW), Telemedicine and health management information system (HMIS) [18].

Shah and Robinson found some key impediments to user involvement in HIT which are lack of resources, attitudes of technical developers and healthcare personnel, and lack of understanding the appropriate interdisciplinary knowledge and skills. However, a physician's involvement is very crucial for the success of HIT [19]. A proactive use of information technology in the health sector empowers consumers of health services to have an easy access health information and decision making tool and by the employment of HIT. Health professionals can collaborate more easily when distance is a major factor in health care delivery [20]. In past HIT applications were used for administrative activities and financial activities rather assisting and delivering health care [21].

Majumder identifies different trends in medical education. There is a noticeable change in healthcare environment due to advancement of ICT and the pervasion of World Wide Web in medicine. The health sector is investing a lot in the information systems and health technologies, so , physician have to depend on ICT and informatics skills for providing their contributions in research, education and health care organization. Therefore, knowledge and skill to use ICT application for self-directed learning is important for medical professionals. Some international organizations like the World Bank and WHO have taken initiatives to orient interdisciplinary skills in the health sector to prepare physicians to cope with the ever changing dynamic health environment. Such pedagogical environments require a lot of cognitive skills and interdisciplinary knowledge [22].

By addressing the IT knowledge cognitive barrier medical students not only be better exposed to evidence based decision making but it will foster their capacity towards a new world of learning. By exposing medical professionals to the information technology knowledge will prepare them for a better orientation of electronic health information systems. Koschmann assessed three different approaches to address cognitive mismatch, which are learning about, through and with computers. He concluded that none of the above mentioned approaches are sufficient alone in fact combination is required [23].

HMSO Report showed that the failure of The National Program for IT in the NHS is the lack of IT skills within National Health Services [24]. Still culture of use of IT in the health care system is not established yet. Bond, 2006 in his doctoral dissertation explored the ways of developing informatics skill and knowledge to fill the knowledge gap .He used a questionnaire tool to collect data from 132 nurses about their usability experience and practicing IT skills behavior in their routine practices [25].

In another study Steven R. Simon and Rainu Kaushal randomly took 1884 physicians as their sample size with the response rate of 71.4%. They explored the level of physicians' use of electronic health record among two groups and also investigated the factors which correlate this use. One group is that which used EHR's functions and thought that they are a useful tool for their practice. The other group is low users of the information system and they even don't aware of the total available and useful functionalities of the system. Therefore the adoption of HIT and HIS is very critical and is a key issue in the health care system. In the past more attention was given on adoption but little literature is available on physician's capabilities to use the system [26].

On CDCP (the Centers for Disease Control and Prevention) panel a report named "Electronic Medical Record/Electronic Health Record Systems of Office-based Physicians: United States" states that physicians are not using the core functionalities of electronic health record because of two possible reasons. One is that their system may lack in having

these functionalities or the other possibility is that the physician may not aware or have not had knowledge of the presence of those functionalities in their system. When it is analyzed that what is the reason that physicians don't know about the availability of the system results the largest gap in the knowledge and cognition mismatch was there [27]. Adherence to data definitions with the proper involvement of end users in the development process of HIS are important factors in gaining the interoperability of the system [28].

### *B. Knowledge And Skills Of Health Professionals*

The health care system is an information intensive domain in which timely access to quality information with accuracy is very critical. Medical professionals are one of the major stakeholders of health information technology. There are lots of barriers to acceptance of information technology by this group of stakeholders. Barriers of this knowledge gap are sociological, cultural, and organizational and technological.

Some researchers reported that physicians are reluctant to introduce the systems into their practice because of time constraints [29] [30] [31]. At present healthcare professionals still seem to be lagging behind in participation in IT Development [32]. One important factor is knowledge and cognitive barrier as their level of information technology literacy is low to use applications of IT [33]. This low level even indefensibly low level of IT literacy or basic knowledge and skills is the reason argued by the health professionals. Due to this reason understanding the concepts and importance of health information systems and information technology cannot be justified by medical professionals. Nykanen and Karimaa in their research stated that the starting point for development of health information system should be to acquire an insight into the healthcare domain where the information system is going to be used [34].

### *C. Adoption And Use Of Hit*

Health information technology (HIT) is a vital element which can address inefficiencies and discrepancies in healthcare in an efficient way. So it is necessary to understand the challenges hurdles and barriers which can limit meaningful use of HIT. Poorly developed user interface and system design is a big hurdle in the clinical workflow and can result in wasted time, poor data collection, misleading data analysis, and potentially negative clinical outcomes. Decisions on technology acquisitions and implementations are often made by individuals or groups that lack clinical informatics expertise [35]. Hence it results in poor usability of the system. Developing such a system is a waste of human resources as well as economics. Proper introductory education of IT is one of the key antecedents of attitude toward computer use behavior by humans [36].

A "design–reality gap" concept is introduced by Heeks arguing the misunderstandings and mismatches between the current realities and design of healthcare information systems [37]

## III. CONCLUSION

Information technology cognitive skills for physicians are very critical to address in gaining persistent partnership of physicians in the HIT Application development and in achieving the maximum usability of health information systems and technologies. It will be very helpful in successful implementation of HIS but also foster the evidence based practice and increase usability of HIS. Interdisciplinary skills in medical education will be of great value in future research as well.

## REFERENCES

[1] Fozia Anwar, Azra Shamim" Barriers in Adoption of Health Information Technology in Developing Societies" International Journal of Advanced Computer Science and Applications.Vol 2 Issue 8 August 2011

[2] Norm Archer, Mihail Cocosila, A Comparison of Physician Pre-Adoption and Adoption Views on Electronic Health Records in Canadian Medical Practices, J Med Internet Res. 2011 Jul-Sep; 13(3): e57

[3] Viitanen J, Hyppönen H, Lääveri T, Vänskä J, Reponen J, Winblad I. National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int J Med Inform. 2011 Oct;80(10):708-25.

[4] Marc Berg, Patients and professionals in the information society: what might keep us awake in 2013, International Journal of Medical Informatics 66 (2002) 31_37.

[5] Susanna Martikainen, Johanna Viitanen, Mikko Korpela, Tinja Lääveri. Physicians' experiences of participation in healthcare IT development in Finland: Willing but not able. International Journal of Medical Informatics 8 1 (2 0 12) 98–113 Häyrinen, K., Saranto, K., Nykänen, P. (2008) Definition, Structure, Content, Use and Impacts of Electronic Health Records: A Review of the Research Literature. International Journal of Medical Informatics 77, 291-304.

[6] J. Bardram, A. Mihailidis, W. Dadong (Eds.), Pervasive Computing in Healthcare, CRC Press, Taylor & Francis Group, USA (2006)

[7] R. Lenz, T. Elstner, H. Siegele, K.A. Kuhn A practical approach to process support in health information systemsJ. Am. Med. Inform. Assoc., 9 (6) (2002), pp. 571–585

[8] American Psychological Association. 2007. APA Dictionary of Psychology. Washington, D.C.:American Psychological Association.

[9] Masys DR. Advances in information technology. Implications for medical education. West J Med. 1998; 168: 341 –347.

[10] Ward JPT, Gordon J, Field MJ, Lehmann HP. Communication and information technology in medical education. Lancet. 2001; 357: 792 – 796.

[11] Paperny DM. Computers and information technology: implications for the 21st century. Adolesc Med. 2000; 11: 183– 202.

[12] Pauker SG, Stahl JE. Medical informatics: where the action is [editorial].West J Med 1997 Feb; 166:148-150

[13] Reuss, E., Rochus, K., Naef, R., Hunziker, S., Furler, L. (2007a) Nurses' Working Practices: What can We Learn from Designing Computerized Patient Record Systems. In A. Holzinger (eds.) USAB2007, Graz, Austria, Springer-Verlag, Berlin, Heidelberg, 55-68.

[14] Parmit K. Chilana, Andrew J. Ko, Jacob O. Wobbrock, Tovi Grossman, George Fitzmaurice. Post-Deployment Usability: A Survey of Current Practices, CHI 2011

[15] Sebastian Garde, David Harrison, Evelyn Hovenga Australian Skill Needs Analysis of Health Informatics Professionals Volume 1: Rationale & Methods, Key Findings & Conclusions, Research Report 1/2005

[16] Shah, S. G. S., Robinson, I. (2007) Benefits of and Barriers to involving Users in Medical Device Technology Development and Evaluation. International Journal of Technology Assessment in Health Care 23, 131-137.

[17] Viitanen J, Hyppönen H, Lääveri T, Vänskä J, Reponen J, Winblad I. National questionnaire study on clinical ICT systems proofs: physicians suffer from poor usability. Int J Med Inform. 2011 Oct; 80(10):708-25.

[18] Innovators and Visionaries: Strategies for Creating a Person-centered Health System. FACCT: Foundation for Accountability; September 2003.

[19] Shah, S. G. S., Robinson, I. (2006) User Involvement in Healthcare Technology Development and Assessment: Structured Literature Review. International Journal of Health Care Quality Assurance 19, 500-515.

[20] Innovation of health technology, Report by Centre of AHIP center of policy and research, Effective new solution for Americans health insurance plans

[21] Audet AM, Doty MM, Peugh J, Shamasdin J, Zapert K, Schoenbaum S, Information Technologies: when will they make it into physicians' black bags? MedGenMed. 2004; 6:2. [PMID: 15775829](13)

[22] Majumder A, D'Souza U, Rahman S. Trends in medical education: Challenges and directions for need-based reforms of medical training in South-East Asia. Indian J Med Sci 2004; 58:369-80

[23] Koschmann T. Medical education and computer literacy: learning about, through, and with computers. Acad Med. 1995 Sep; 70(9):818-21

[24] Guest Editorial Nurses' requirements for information technology: A challenge for educators International Journal of Nursing Studies 44 (2007) 1075–1078.

[25] National Audit Office, 2006. Department of Health. The National Program for IT in the NHS.

[26] Steven R. Simon; Rainu Kaushal; Paul D. Cleary; Chelsea A. Jenter; Lynn A. Volk; E. John Orav; Elisabeth Burdick; Eric G. Poon; David W. Bates. Physicians and Electronic Health Records: A Statewide Survey Arch Intern Med. 2007; 167(5):507-512.

[27] http://www.cdc.gov/nchs/products/pubs/pubd/hestats/electronic/electronic.htm. Accessed on February 6, 2012.

[28] Khadzir bin Sheikh Haji Ahmad, Health Data Integration, Malaysian Journal of Public Health Medicine, Vol. 11(Suppl 1) 2011

[29] Anogianakis, G., S. Maglavera, and A. Pomportsis, ATTRACT--applications in telemedicine taking rapid advantage of cable television network evolution. Studies in Health Technology & Informatics, 1998. 50: p. 60-6.

[30] Bielli, E., A Wireless Health Outcomes Monitoring System (WHOMS): development and field testing with cancer patients using mobile phones. BMC Medical Informatics & Decision Making, 2004. 4: p. 7.

[31] Muuronen, A. (2008) Lääkäreillä on liikaa paperitöitä (in Finnish). (The doctors have to do too much paper work). Helsingin Sanomat newspaper, an article published in discussion column,published in December 6th, 2008.

[32] S. De Rouck, A. Jacobs, M. Leys A methodology for shifting the focus of e-Health support design onto user needs: a case in the homecare field Int. J. Med. Inform., 77 (9) (2008), pp. 589–601

[33] David A. Dzewaltowski, Paul A. Estabrooks, Lisa M. Klesges, Sheana Bull and Russell E. Glasgow . Behavior change intervention research in community settings: how generalizable are the results? Health Promotion International Oxford Journal, Volume19, Issue2 Pp. 235-245.

[34] P. Nykanen, E. Karimaa, Success and failure factors in the regional health information system design process—result from a constructive evaluation study, Methods Inf. Med. 4 (2006) 85–89

[35] Kadry, Bassama; Sanderson, Iain Ca; Macario, Alexb, Challenges that limit meaningful use of health information technology, Current Opinion in Anaesthesiology: April 2010 - Volume 23 - Issue 2 - p 184–192

[36] Chuang, Y.H., Chuang, Y.W., 2002. "Attitudes of two-year RNBSN nursing students towards computers", The Journal of Health Science 5 (1), 71–84

[37] R. Heeks, Health information systems: failure success and improvisation, Int. J. Med. Inform. 75 (2006) 125–137

AUTHOR'S PROFILE

Fozia Anwar is a research scholar in CIS department of University Technology Petronas and after completing her Bachelor degree in dentistry she did MS in Health Informatics from COMSATS Institute of Information and Technology.

# Techniques to improve the GPS precision

Nelson Acosta
Institute of Advanced Informatic Technology Research
University Center of the Province of Buenos Aires
Tandil, Argentina

Juan Toloza
Institute of Advanced Informatic Technology Research
University Center of the Province of Buenos Aires
Tandil, Argentina

***Abstract*— The accuracy of a standard market receiver GPS (Global Positioning System) is near 10-15 meters the 95% of the times. To reach a sub-metric level of accuracy some techniques must be used [1]. This article describes some of these procedures to improve the positioning accuracy by using a low-cost GPS in a differential relative positioning way. The proposed techniques are some variations of Kalman, fuzzy logic and information selection.**

***Keywords- GPS accuracy; relative positioning; DGPS; precision farming GPS.***

## I. INTRODUCTION

The GPS operation principle is based on measuring ranges of distances between the receiver and the satellites [2] [3]. The GPS has architecture of three segments: spatial, control and users. The spatial includes 24 satellites over 20 thousand km away from the Earth, with six orbital levels and a 12-hour period. The second segment includes the Earth stations to control the satellites trajectories. Finally, the user's segment includes GPS receivers using two frequencies: L1 at 1575.42 Mhz for civil use, and L2 at 1227.60 Mhz reserved to military use [4].

The accuracy in longitude and latitude coordinates is of 10-15 meters 95% of the readings [5]. Sometimes, it is more precise, but it depends on a variety of factors that include from the deviation or the delay of the signal when cross the atmosphere, the bouncing of the signal in buildings or its concealment due to the presence of trees [6], low accuracy of clocks and noise in the receiver. In altitude the accuracy is reduced to 50% regarding the obtained in terms of longitude and latitude (15-23 meters 95%) [7].

Systems that enhance positional accuracy are: the DGPS (Differential GPS), AGPS (Assisted GPS), RTK (Real-Time Kinematic), e-Dif (extended Differential), amongst others.

The DGPS corrections service has two hard restrictions: it must be afforded and the receiver must be close to a DGPS station (less than 1000 km). The achieved accuracy can be of a few meters [7, 8, 9, 10]. The correction signal cannot be received if it is a mountainous zone.

In the case of AGPS, it is necessary to have mobile devices with active data connection or cell phone like GPRS, Ethernet or WiFi [11]. It is used in the cases where there is a weak signal due to a surrounding of buildings or trees; this implies having a not much precise position. Standard GPS receivers, in order to triangulate and position, need a certain time of cold start [12][13].

*a)* For use the RTK system, it is paid for the service and, besides, it is very expensive to acquire the infrastructure. This is a technique used in topography, marine navigation and in agricultural automatic guidance in the use of measurements of signals carrying navigators with GPS, GLONASS (Globalnaya Navigatsionnaya Sputnikovaya Sistema) and/or Galileo's signals, where only one reference station provides correction in real time, obtaining a sub-metric accuracy [14].

*b)* The last case, e-Dif system, is autonomous and it process files with RINEX (Receiver Independent Exchange) format, which was created to unify data of different receivers manufacturers [15]. It generates autonomous corrections regarding a coordinate of arbitrary reference and it extrapolates them in time [16]. It is a very consistent relative positioning and its accuracy is of about 1 meter. The system's objective is to study waste from the initializing process to isolate the most important systematic errors that introduce the corresponding equations to each satellite. It is applicable for a reduced time of 40 minutes approximately; since later the systematic error changes, in this case a new error must be calculated again. In regions where differential corrections aren't available and it is paid for the service, like in South American, African and Australian, this system become more interesting.

*c)* Besides, there are systems that increase accuracy to sub-metric levels. Those based on satellite SBAS (Satellite-Based Augmentation System), based on ground GBAS (Ground-Based Augmentation System) and based on aircraft ABAS (Aircraft-Based Augmentation System). Most of these implementations are used in different applications and some of them are available for users without special permissions. Even then, costs are high due to the need of certain devices with special characteristics or some infrastructure in agreement with the accuracy level desired.

Errors produced by the GPS system affect in the same way the receivers located near each other in a limited radius. This implies that errors are strongly correlated among near receivers. Thus, if the error produced in one receiver is known, it can be spread towards the rest in order to make them correct their position. This principle is only applicable to receivers that are exactly the same, the same methodology of [17]; since, if different, their specifications change so the signal processed by one individual is not the same to that processed by another one.

All GPS differential methods use the same concept [18]. DGPS requires a base station with a GPS receiver in a precise known position. The base compares its known position with that calculated by the satellite signal. The estimated difference in the base is applicable then to the mobile GPS receiver as a differential correction with the premise that any two receivers relatively near experiment similar errors [5].

In this article it is emphasized in the behavior of GPS errors in time, after the techniques developed here to determine magnitude and direction of error are applied. In Section 2 techniques to calculate errors are analyzed. In Section 3 algorithms used are described and finally the error behavior is presented in figures. In Section 4, the last one, conclusions and future work are presented.

## II. Techniques For Error Analysis

The experiment carried out is based on the principle of the adopted methodology by the DGPS but with a low cost standard GPS receiver. In order to get measurements, three Garmin 18X USB GPS receivers are used connected to two notebooks. The base station is composed of a notebook and two of the three GPS receivers; the mobile for the other notebook and the remaining GPS receiver. The link between the base station and the mobile one is a point to point wireless connection.

In this context, in the base system, measurements from the GPS receivers are obtained and after a certain period of time, which is necessary for the system stabilization, two positions are estimated. The positions' estimation are carried out with a Kalman filter, since an estimation problem with so many noisy redundant data is a natural application for the Kalman filter; this allows using some of the redundant information to remove the effects from the error sources. The Kalman filter is used to eliminate the white Gaussian noise [19].

Receivers are placed at a known distance between themselves (relative positioning) [20]. At the end of this stage, a cloud of points from standard GPS system is obtained with the positions delivered by the receivers and with those values a position is estimated for each receiver. The estimated point from GPS 1 is selected as anchor point of the whole experiment. From this, all necessary calculations are carried out with the objective of finding the GPS system error.

With both estimated positions, it proceeds to calculate the distance between them. If the estimated distance is different from the actual one (more/less a threshold) it is detected that there is a positioning error. Besides, a circumference with a radius equal to the actual distance measured with a tape measure from GPS 1 to GPS 2 with center in the estimated point for the GPS 1 is drawn. A circumference is chosen, as GPS 2 can be at that distance but in any point of the circumference. This is the working principle that the GPS system uses to get the receiver's position.

After calculating the distance of the estimated points and contrasting it to the actual one, a positioning error is deduced. Once it is known that there is an error, it must compute its magnitude and direction. On the one hand, the two estimated points are learnt with which the straight line is drawn and which bonds them. Equation 1 belongs to the straight line that crosses these two points.

$$\frac{(y-y1)}{(y2-y1)} = \frac{(x-x1)}{(x2-x1)} \tag{1}$$

where the *x* represents the component of Longitude and the *y* that of Latitude.

On the other hand, it is known that the GPS 2 is in some point of the circumference with center in the GPS 1 and of radius the distance that was defined at the moment of positioning the two receivers. Equation 2 belongs to the circumference with center in *(x1,y1)* of radius r.

$$(x-x1)^2 + (y-y1)^2 = r^2 \tag{2}$$

Knowing about the equations that define the straight line crossing both estimated points and the radius circumference equation equal to the GPS 1-GPS 2 distance with center in GPS 1, it is proceeded to approximate, by means of the intersection of the straight line and the circumference.

This intersection is presented as a polynomial of second degree. It is mathematically solved and *a*, *b* and *c* coefficients are obtained (equations 4, 5 and 6). Equation 3 only presents an auxiliary estimate in order not to repeat it in the other operations and to increase legibility in the rest of the equations. With these coefficients cleared by means of Bascara (equation 7) the roots are found (two because of being of second degree) from the polynomial. From the two roots found, one is chosen and the intersection points are calculated.

$$divisor = y_2^2 - 2y_2y_1 + y_1^2 \tag{3}$$

$$a = 1 + \frac{\left(x_2^2 - 2x_2x_1 + x_1^2\right)}{divisor} \tag{4}$$

$$b = -2y_1 + \frac{\left(-2x_2^2y_1 + 4x_2x_1y_1 - 2x_1^2y_1\right)}{divisor} \tag{5}$$

$$c = y_1^2 - r^2 + \frac{\left(x_2^2y_1^2 - 2x_2x_1y_1^2 + x_1^2y_1^2\right)}{divisor} \tag{6}$$

$$x = \frac{-b \pm \sqrt{\left(b^2 - 4ac\right)}}{2a} \tag{7}$$

After having obtained the two roots, only one is taken into account. The nearest to the estimated point of GPS 2 in some of its components is chosen, in this case in Latitude, since the other one is meaningless due to being too far away (on the other side of the circumference).

$$result_{latitude} = rootCloser\left(y_2, roots.root_1, roots.root_2\right)$$

Finally, the other component (Longitude) is obtained depending on the Latitude found in the previous point as seen in equation 8.

$$result_{longitude} = x_1 + \left( (x_2 - x_1) * \frac{(result_{latitude} - y_1)}{y_2 - y_1} \right) \quad (8)$$

Once the error magnitude is known, is necessary to get the direction in order to have the final correction vector. There are nine possible cases, that are the combination of two variables (latitude and longitude) and three values that are: greater, lesser and equal. In Table I these cases are presented.

TABLE I. DETERMINATION OF ERROR DIRECTION.

| **Estimated** | |
|---|---|
| **Latitude** | **Longitude** |
| Greater | Greater |
| Greater | Lesser |
| Greater | Equal |
| Lesser | Greater |
| Lesser | Lesser |
| Lesser | Equal |
| Equal | Greater |
| Equal | Lesser |
| Equal | Equal |
| **Latitude** | **Longitude** |
| **Actual** | |

The way to read the table is top-down. Estimated latitude greater than actual latitude and an equal longitude to actual is presented in the third row. In the case where actual and estimated components are equal, like in ninth row of the Table I, there is no correction.

Therefore, in order to get the error direction it is proceeded to verify which side of the corrected point the estimated point is. If the straight line that joins both estimated points is horizontal or vertical, some of the components are null, in the horizontal case, Latitude is eliminated and in the vertical the Longitude.

## III. ALGORITHMS AND USED TECHNIQUES

For the analysis of data, combinations of different techniques and algorithms are used in order to find a better result. In a first processing stage, applied mathematics covers:

- Static Kalman: is a set of mathematical equations that provide an efficient recursive solution of the method of least squares. This solution allows calculating a linear, unbiased and optimum estimator of the state of a process in each moment of time (*t*) with base on the available information at the moment *t-1*, and update, with the available information at the moment *t*, the estimator value.
- Dynamic Kalman: is the system in which the value of variable *x* to be estimated has a value that changes throughout the time ( $x_{i+1} \neq x_i$ ), but these states have some known relationship with the instant *i* and *i+1*. For example, if an object position is measured, it can be predicted that the position will be:

$$x_{i+1} = x_i + \Delta t * v_i$$

where $\Delta t$ is the passed time and $v_i$ the speed at instant *i*. Position can be obtained by a GPS, for instance, and speed with an additional measurement element such as an accelerometer.
- Kalman with adjustment of error standard deviation: the deviation is modified and checked in order to see which adjust better. This measure is calculated as the square root of variance, which is at the same time the sum of the squares of each error (Table II) as shown in equation 9. It is worth mentioning that from Table II the only error that is not taken into account is that of signal P(Y) arrival; since work is carried out without the precision code.

$$\sigma_R = \sqrt{3^2 + 5^2 + 2.5^2 + 2^2 + 1^2 + 0.5^2}\, m = 6{,}7\, m \quad (9)$$

TABLE II. GPS ERROR SOURCE.

| **Source** | **Effect (Meters)** |
|---|---|
| Arrival of signal C/A | ± 3 |
| Arrival of signal P(Y) | ± 0.3 |
| Ionosphere | ± 5 |
| Ephemeris | ± 2.5 |
| Satellite clock error | ± 2 |
| Multipath | ± 1 |
| Troposphere | ± 0.5 |
| Numerical errors | ± 1 |

Now, error standard deviation ( $\sigma_{rc}$ ) in the receiver's position is estimated, but having into account additionally the PDOP (Position Dilution of Precision) and the numerical error; therefore, the PDOP is added to the calculated deviation from typical errors, since for each measurement taken, this varies according to the instant geometry of satellites. The result of standard deviation used for the Kalman filter is equation 10.

$$\sigma_{rc} = \sqrt{PDOP^2 * \sigma_R^2 + \sigma_{num}^2} = \sqrt{PDOP^2 * 6.7^2 + 1^2}\, m \quad (10)$$

The fact of applying Kalman with adjustment of standard deviation, since it fluctuates for each piece of information coming from the receivers in each moment as geometry of satellites varies.

- Points average: one of the media limitations is that it is affected by extreme values; very high values tent to increase it while very low values tend to reduce it; this implies that it may stop being representative of the population. It is analyzed but not implemented in the solution. This solution was used in [21].
- Fuzzy logic: to determine the position error degree. Rules that determine the position error are related to analyzing some parameters like: PDOP, SNR (Signal-to-Noise Ratio) and difference of tracked

satellites. The fuzzy system output weights the Kalman filter gain, providing more weight to more precise positions and the other way round.

- Filters allow discarding measurements with much noise or error that influence over the final result of an estimation of a position. Thus, measurements having many errors do not slant the final estimation towards a position far away from the actual one. The application of these filters can be made as measurements are not very far away in time and it is supposed that the Vehicle in which the mobile receiver is placed does not move at high speed; this implies that values do not change radically. High/low step filters are used in an analogical way to the electronic filter.

After apply the current techniques, in Figs. 1, 2 and 3 it is observed the original and estimated errors. The absolute errors are calculated by means of relative positioning determined by sets of two receivers. The distance between the pairs of receivers is known at starting the experiment. In Fig. 1, the estimated error tends to zero. By the other hand, in Fig. 2, estimated error fluctuates around the average of the original error. Finally, in Fig. 3, the estimated error is zero in a moment, but after a period, begins to oscillate. With this set of graphics, it is observed an oscillation in the standard system that doesn't allows having a known error to correct a position. When the presented techniques are applied it is possible to obtain an smoothing error, a value that fluctuates less in time, which allows to correct positions in other receivers in order to improve the accuracy.

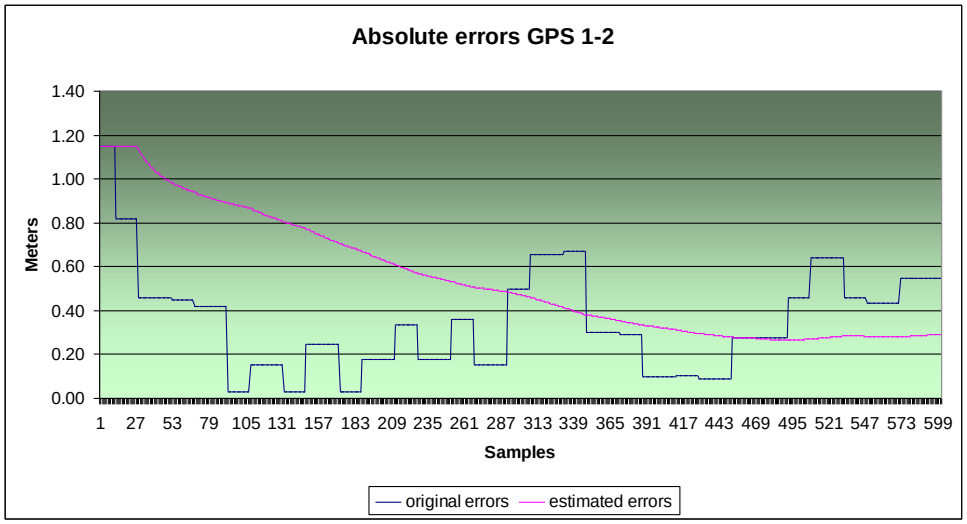



Figure 1. Behavior in time of estimated and original errors for GPS 1 regarding to GPS 2.

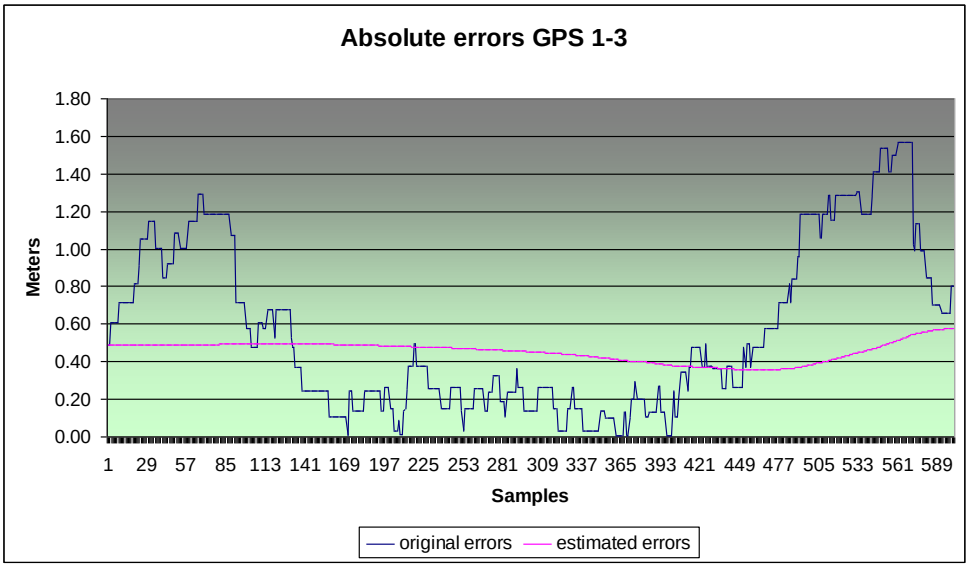



Figure 2. Behavior in time of estimated and original errors for GPS 1 regarding to GPS 3.

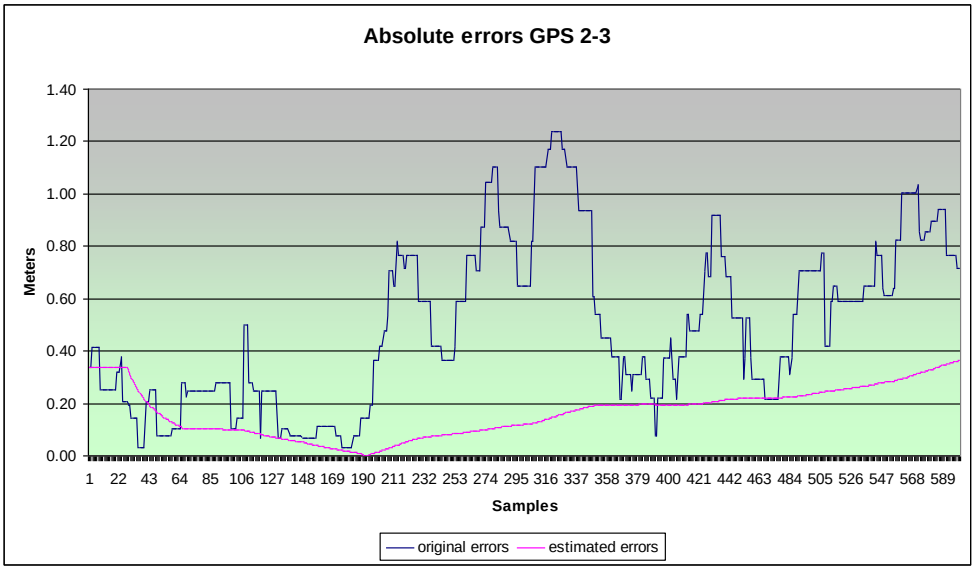



Figure 3. Behavior in time of estimated and original errors for GPS 2 regarding to GPS 3.

Since the tool that implements these techniques is thought to operate in different places of heterogeneous characteristics, relations and configurations are used in order to be able to customize the use according to needs. Relations and configurations used are the following:

- Rate Degree/Meters in Latitude: given the asymmetry, in different places on Earth, the distance that a Latitude degree measures varies.
- Rate Degree/Meters in Longitude: ditto to Latitude.
- Cold start time: a start time is considered so the system can be stabilized. In this time, samples of the device are taken and, only at its end, estimation is carried out. The objective is to reduce or soften systematic and random errors of the GPS system.
- Receivers' distance threshold: it can be determined, in an accurate way, the whole interval of the distance measurement between the base devices. As positioning is relative and its distance is known, it can be added a ± value, since there exists a possibility that an element of distance measurement be not accurate enough. Besides, it reduces the computational load because of not having to process data if distance is within the allowed threshold.

## IV. Conclusion And Future Works

The techniques developed allow obtaining the magnitude and direction of error provoked by the GPS system as presented in [21]. This correction is used by another receiver to correct its own position and thus increase the positional accuracy with the aim of measuring the most precise distances.

The experiments carried out with different sets of data provide positions that are used to measure distances and error fluctuates in ± 1 meter the 95% of measurements and in some cases in ± 0.2 meters.

The principle is based in mathematical, geometrical functions and filters. With the techniques presented in this article, it is possible to obtain a smoothing error and a value constant in time, which allows correcting positions in other receivers in order to improve their accuracy.

As future work, the aim is to increase the accuracy until reaching a maximum error of positioning of ± 0.1 meters. This increase can include the use of another additional signal. Besides, further measurements will be carried out in order to analyze data and to deduce its behavior. It is intended to extend the use in faster vehicles in order to widen the application field of the DGPS system introduced here.

## Acknowledgment

Agencia Nacional de Promoción Científica y Tecnológica for supporting since 2009.

## References

[1] M. S. Grewal, L. R. Weill and A. P. Andrews, Global Positioning Systems, Inertial Navigation, and Integration. 2nd Edition, Wiley, 2007.

[2] G. Xu, GPS: Theory, Algorithms and Applications. 2nd Edition, Springer-Verlag Berlin Heidelberg, 2007.

[3] P. Misra and P. Enge, Global Positioning System: Signals, Measurements, and Performance. New York, Ganhga-Jamuna Press, 2010.

[4] T. Feldmann, A. Bauch, D. Piester, H. Esteban, J. Palacio, F. J. Galindo, T. Gotoh, H. Maeno, U. Weinbach and S. Schon, "GPS carrier phase and precise point positioning time scale comparisons using different software packages," Frequency Control Symposium, 2009 Joint with the 22nd European Frequency and Time Forum, IEEE, pp. 120-125, 2009.

[5] P. A. Zandbergen and L. L. Arnold, "Positional accuracy of the wide area augmentation system in consumer-grade GPS units," Computers and Geosciences Volume 37 Issue 7, Elsevier, pp. 883-892, 2011.

[6] C. Ordónez Galán, J. R. Rodríguez-Pérez, J. Martínez Torres and P. J. García Nieto, "Analysis of the influence of forest environments on the accuracy of GPS measurements by using genetic algorithms," Mathematical and Computer Modelling Volume 54 Issue 7-8, Elsevier, pp. 1829-1834, 2011.

[7] S. Featherstone, Outdoor Guide to Using Your GPS. Creative Publishing International, Inc., 2004.

[8] G. Satheesh, Global Positioning Systems: Principles and Applications. Mc-Graw Hill, 2005.

[9] S. D. Ilčev, Global Mobile Satellite Communications for Maritime, Land and Aeronautical Applications. Springer, 2005.

[10] M. Ghavami, L. B. Michael and R. Kohno, Ultra Wideband Signals and Systems in Communication Engineering. 2nd Edition, John Wiley and Sons, Ltd., 2007.

[11] C. Ho, "An effective approach in improving A-GPS accuracy to enhance hybrid positioning computation," 17th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), Toyama, Japan, IEEE, pp. 126-130., 2011.

[12] J. Li and M. Wu, "A positioning algorithm of AGPS," International Conference on Signal Processing Systems, Singapore, IEEE, pp. 385-388, 2009.

[13] F. Van Diggelen, A-GPS, Assisted GPS, GNSS, and SBAS. Artech House, 2009.

[14] D. Dardari, E. Falletti and M. Luise, Sattellite and Terrestrial Radio Positioning Techniques: A Signal Processing Perspective. 1st Edition, Elsevier, 2012.

[15] N. H. M. Hanif, M. A. Haron, M. H. Jusoh, S . A. M. Al Junid, M. F. M. Idros, F. N. Osman and Z. Othman, "Implementation of real-time kinematic data to determine the ionospheric total electron content," 3rd International Conference on Intelligent Systems Modelling and Simulation (ISMS), Kota Kinabalu, Malaysia, IEEE, pp. 238-243, 2012.

[16] GPS World (serial online), "CSI wireless differential software patented," Volume 13 Issue 8, EDS Foundation Index, Ipswich, MA. pp. 48, 2002. Accessed June 18, 2012.

[17] M. G. Wing and J. Frank, "Vertical measurement accuracy and reliability of mapping-grade GPS receivers," Computers and Electronics in Agriculture Volume 78 Issue 2, Elsevier, pp. 188-194, 2011.

[18] V. Di Lecce, A. Amato and V. Piuri, "Neural technologies for increasing the GPS position accuracy," International Conference on Computational Intelligence for Measurement Systems And Applications (CIMSA), Istanbul, Turkey, IEEE, pp. 4-8, 2008.

[19] H. Eom and M. Lee, "Position error correction for DGPS based localization using LSM and Kalman filter," International Conference on Control, Automation and Systems (ICCAS), Gyeonggi-do, Korea, IEEE, pp. 1576-1579, 2010.

[20] Y. He, H. Yu and H. Fang, "Study on improving GPS measurement accuracy," Instrumentation and Measurement Technology Conference (IMTC), Otawa, Canada, IEEE, pp. 1476-1479, 2005.

[21] J. Toloza, N. Acosta and A. De Giusti, "An approach to determine the magnitude and direction error in GPS system," Asian Journal of Computer Science and Informatin Technology, in press, 2012.

AUTHORS PROFILE

Ph.D. Héctor Nelson Acosta was graduated at the Universidad Nacional del Centro de la Provincia de Buenos Aires (Argentina), and got the PhD degree at the Autonomous University of Madrid (Madrid, Spain). He is the director of the Instituto de Investigación en Tecnología Informática Avanzada (INTIA) at the Universidad Nacional del Centro de la Provincia de Buenos Aires since the 1998. He is working as a proffesor since 1993 at the Computer and Systems Department, at the Universidad Nacional del Centro de la Provincia de Buenos Aires, Tandil (Argentina). He has been a visiting proffesor of several Universities in Argentine. He has suppervised 6 Ph.D. students in Computer Sciences and 2 MSc in Informatics. His research insterest includes real-time systems, signal procesing, pattern recognition, custom architectures, custom processors, robotic navigation and accuracy positioning systems.

Eng. Juan Manuel Toloza was graduated in march 2009 at the Universidad Nacional del Centro de la Provincia de Buenos Aires (Argentina). He is in the last year of Ph. D. (Computer Science) at the Universidad Nacional de La Plata (Argentina). He have a scholarship of the Agencia Nacional de Promoción Científica y Tecnológica since 2009. He works in the Instituto de Investigación en Tecnología Informática Avanzada at the Universidad Nacional del Centro de la Provincia de Buenos Aires. His research insterest includes signal procesing, data analysis from diferents sensors, robotics and accuracy GPS. He is teacher assistant in Computer`s Architecture and Digital Techniques.

# M-Commerce service systems implementation

Dr.Asmahan Altaher
Applied Science University
Amman - Jordan

***Abstract*— Mobile commerce supports automated banking services. However, the implementation of m-commerce services systems has become increasingly important in today's dynamic banking environment. This research studied the relationships between technology acceptance model and m- commerce services. The results of the survey on 249 respondents in several Jordan banks revealed that technology acceptance model had a significant impact on m-commerce services. The results led to the recommendation that the technology acceptance model is a success model for support using new services for electronic commerce. In addition, managers play a significant role in influencing the mobile services in banks through social interaction. Managers should focus on relative advantage, usefulness, and ease of use, in order to develop the mobile commerce services implementation.**

***Keywords- M-commerce services; usefulness; and ease of use; social interaction.***

## I. INTRODUCTION

The growth in the mobile phone industry has accelerated in few years due to constant technological development. Recently, new forms of mobile services have made possible text messaging, web surfing, digital imaging, payments, banking, financial instrument trading and shopping [6]. With the rapid advance of telecommunication technologies, Mobile Services (MDS) is defined here as wireless access to the internet through a mobile communication network. However, mobile services are becoming increasingly important for companies and consumers because of ubiquitous, universal, easy-to-access information, and personalized exchange of information [21]. Hence, it is very important to understand how individual differences influence the use of mobile services and the behavioral adoption requirements of these services, marketers and service providers [4]. Employees may resist the change and may be a major source of concern for them. To avoid negative consequences of resistance to change, system implementers and managers must actively manage the change process and gain acceptance for new IS. The Technology Acceptance Model (TAM) suggests that employee attitudes may change if they think the new system will help them do more or better work for the same effort, and that it is easy to use [15]. TAM suggests that managers cannot get employees to use a system until they want to use it. To convince employees to use the systems, managers may need to change employees' attitudes about the system [10]. Employees attitudes may change if employees believe that the system will allow them do better work for the same amount of effort (perceived usefulness), and that it is easy to use. Training, documentation, and user support consultants are external variables that may help explain the usefulness of the system and make it easier to use; in addition, mobility model is very important to improve the efficiency of secure mobility and routing discovery [20]. TAM has many variants, for example, one variant considers subjective norms, whereas another adds attitudes toward behaviours like social influence (subjective norms), and facilitating condition (top management support). Although social influences are important, they are likely to be important only for yang workers when they are likely to start using the system. TAM assumes that technology will be accepted if people's attitudes and beliefs support its use. One way to make sure that employees' attitudes and beliefs are favourable toward the system is to have them participate in its design and implementation. When future users of the system participate in its design and implementation, they may be more willing to accept the consequences of the trade-off [10] [18]. The objective of this study is to investigate how the technology acceptance model can involve employees and allow them to understand the m- commerce services in Jordan banks. In order to address these questions, the researcher drew post-adoption components from previous studies in related fields, and conducted a large-scale survey on mature TAM in terms of how each component affected the use of m-commerce services systems. Subsequently, the theoretical background and hypotheses of this study is presented, after which a description of the research methods and data collected was presented. This was followed by an analysis of the research results. Thereafter, the conclusion was drawn.

## II. LITERATURE REVIEW

Kim et al. [4] studied and examined factors that are most important in converting services to m-commerce services and whether or not they differ from those that are effective in maintaining continuers. The researchers conducted an online survey to compare continuers and discontinuers empirically in terms of the relative importance of four post-adoption factors to the behavioural intention to use m-commerce. The results show that usefulness and social influence were more important for discontinuers; and ubiquitous connectivity for continuers [17] find that the mobile phone user types are important to the mobile phone service companies, and communicative references for understanding the mobile phones. Mobile phone user types recently have become a popular subject of discussion. The researchers classify the mobile phone user types into four types: guanxi-expanding, illness-phobia, convenience-oriented and life-interrupting. The users of these mobile service companies' references for understanding the mobile phone use were adopted. However, Weber et al. [18] argued that the use of mobile technology for carrying out

surveys is very important and are necessary actions which provide a guideline for realizing the potential of 'mobile surveys'. Essentially, the benefits are enhancements in the areas of survey quality, management and technology. The researchers recommended that the development of mobile surveys is important to create a new business model in the market. On the other hand, Xie et al. [10] present some potential security threats against m-commerce. The security attacks such as wireless-internet can seriously degrade the performance of mobile services of users; in order to promote the widespread deployment of m-commerce, the firms need to design novel and robust as well as efficient security schemes to handle these attacks. In this article, the researchers discussed some important challenges for providing ubiquitous and secure m-commerce. Finally, Lin [5] agreed that m-commerce service is being totally satisfied with the system and their users' values; mobile technology trusting expectations were very important in the continued m-commerce service usage behaviors, and the providers might not fulfills the m-commerce service need for consumers, but satisfied with the m-commerce service delivered. M-commerce services were introduced only a few years ago and are distinctly different from prior services and information systems because m-commerce services systems are a type of information system; and many adoption studies [for example, Technology Acceptance Model (TAM)] have been conducted in the area. To study acceptance of the m-commerce services systems, the researcher selected four key components that have been frequently used in previous adoption studies in these areas: social influence, usefulness, ease of use and mobility model. Definitions of each component, as well as accounts of why they are pertinent to this study are as follow.

## III. PROCEDURAL DEFINITIONS

### A. Social Influence:

Social influence refers to the effect of interaction among people in their social context. Social influence helps to determine whether technologies were adopted and whether products are purchased [14]. Fisher and Price [3] found that when people purchased new products, others' opinions influenced their purchasing decisions. Social influence may also affect the use of MDS. In fact, since M-commerce is a part of the telecommunications industry – an industry specifically designed to facilitate social interactions – social influence may be an even more important factor in service choice [4] [13][20].

### B. Usefulness:

Usefulness is how helpful a user feels a new product is to his or her work [10]. When the usefulness of a new product is high, the product is adopted rapidly in the market. [4] [16] have established that good quality and good function, as perceived by users, allow them to adopt a new system easily in an organizational environment. Usefulness appears to be a key factor in the adoption of M- commerce [11] [12] – especially when their usefulness differentiates them from traditional internet services. For example, subjects in one study reported using m–commerce services only when the usefulness of their mobility really mattered [6].

### C. Ease of Use:

Ease of use**,** another subjective measure, describes how easy and comfortable people find a system to learn and use. Kim et al. [4] suggested that a system is adopted quickly when a user can easily learn how to use it. Novak [8] proposed that a better system must be efficient to use. Ease of use will also be an important factor for the adoption of M- commerce services [1]. Providers are coming more and more to believe that ease of use is the key to retaining the greatest barrier to the adoption of M-commerce. However, compared with the resources of other systems, the resources available to make M-commerce services easier are severely constrained [4] [17].

### D. Mobile Commerce Services:

Mobile internet services numerously involve mobile commerce (m-commerce) services for continued and promoted business profits. Mobile technology trusting expectations were very important in the continued m-commerce service usage behaviors'; and the providers might not fulfills the m-commerce service need of consumers, but satisfied with the m-commerce service delivered. M-commerce provides a significant value in convenience, efficiency, entertainment, mobility and location and besides extending the benefits of the web [5] [6]. Therefore, this study proposes personal values for the prediction of continuance intention. This may offer additional information or insights beyond TAM in explaining why m-commerce services are used. M-commerce service and usage continuance intentions in terms of personal values lead to trust expectations about the m-commerce trusting and perceived performance. Most efficient factor is quick customer service. Legislation has increased customers' rights and technology and competition have increased their choice of products and providers. These changes will result in growth in users with sophisticated needs supported by [2] [9] [19].

Internet technology is rapidly changing the way personal financial services are being designed and delivered. For several years, commercial banks in Jordan have tried to introduce electronic banking (e-banking) systems to improve their operations and to reduce costs. Despite all their efforts aimed at developing better and easier e-banking systems, these systems remained largely unnoticed by the customers, and certainly were seriously underused in spite of their availability.

In this Internet age, when the customer is having access to a variety of products and services it is becoming very difficult for banks to survive. In this situation, when customer enquiries are not met easily or transactions are complicated, the customer will ask for new levels services, and only chose those institutions who are making a real effort to provide a high level of quality, fast and efficient service through all the bank's touch points, call centers, TAM, voice response systems, Internet and branches. It is considered as one of the better financial sectors in the region and generates in total close to 5% of the Gross Domestic Product (GDP). One of the weakest points in the financial sector is, with the exception of mortgage lending, the lack of long-term lending and the absence of secured loans. It is worth mentioning that the percentage of Jordanian households who own personal computer is 15.9%, Internet access is 6%, 1,000,000 regular telephone lines, around 1.6 million mobile telephony

subscribers, 21 licensed Internet service providers, and more than 500,000. The banking sector is very dynamic and liberal in Jordan. Moreover, some of the commercial banks in Jordan are offering electronic services. Samples of these services are:

(1) Internet banking: Arab Bank is the first bank to launch Internet banking service. This has been started in Jordan in May 2000.

(2) Internet Shopping Card (ISC): It provides convenient and easy access to on-line shopping transactions.

(3) WAP banking: Customers can use WAP mobile phone and access their accounts.

(4) SMS banking: Customers can use a mobile phone to access their accounts.

(5) Phone bank service: This provides access to customers' accounts.

(6) On-line stock trading: Jordan Kuwait Bank (JKB) offers this service in collaboration with its affiliate (United Financial Investment).

(7) Net banker: For performing banking transactions.

(8) Mobile Banking: This service allows the customers to perform banking transactions by using a mobile.

(9) Pre-paid mobile cards: Customers can buy the mobile prepaid cards electronically.

(10) Banking via SMS: It enables the customers to receive information on their transactions through their mobile telephones [1]. The financial services industry will be used to explore m-commerce in Jordanian banking services sector, and to investigate the effects of TAM through carrying out m-commerce in Jordanian banking sector (e-banking). The following are the research model.

## RESEARCH HYPOTHESES

$H_1$: There is a significant positive effect between technology acceptance model and m-commerce services implementation.

$H_2$: There is a significant positive effect between technology acceptance model and social influences.

$H_3$: There is a significant positive effect between technology acceptance model and technology usefulness.

$H_4$: There is a significant positive effect between technology acceptance model and technology ease of use.

## RESEARCH MODEL

The model depends in Technology acceptance model pearlson(2009). The researcher tries to investigate the model statistical in order find new way that can that supports involved employees and allow them to understand the m-commerce services in Jordan banks.

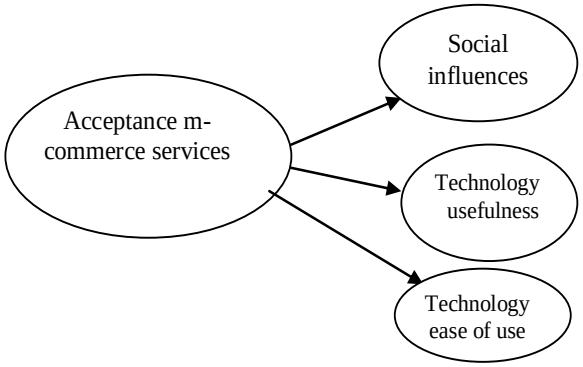



Figure.1. Research Model

Recourse: Pearlson (2009)

## IV. RESEARCH METHODOLOGY

This research is concerned with making our problems accessible by directing it in a way that generates precise answers to precise question. The research methodology can be derived from two approaches that can be classified into two main categories: quantitative and qualitative methodology [13].

### *A. Data Collection Methods*

The data and information will be gathered from two resources, namely: primary and secondary resources.

i) Primary resources: Individual focus groups, and a panel of respondents set up by the researcher whose opinions may be sought on specific issues from time to time are examples of primary data sources [1]. Data can also be culled from administering questionnaire. Questionnaires are an efficient data collection mechanism when the researcher knows exactly what is required and how to measure the variable of interest [22]. In this study, questionnaires were sent to respondents in the senior level, and top managers in management.

ii) Secondary resources: Data can also be obtained from secondary sources, as the scientific (Books, articles, etc) sources concerned with the study.

### *B. Initial Design And Development Of The Survey Instrument:*

Many criteria should be considered when designing a questionnaire survey [22]. On the choice of wording, questionnaire design and layout were adopted. Items in the questionnaire were designed to being simple, clear, short, technical, accurate, bias free, and at an appropriate reading level [1] [22] were taken into account when designing the questionnaire, such as those that started with a brief description on how to answer the questionnaire. An initial draft of questionnaire was developed based on an extensive literature review and existing measures.

### *C. Decisions Related To Population And Sample Selections:*

The banks market is huge and has been growing rapidly in recent years. Banks can use all of the e-services.

However, many large banks receive additional e-services, the banks was seen as an appropriate business environment that is particularly suitable to test for the research model and the determinations of the m-commerce implementations in e-banking. Jordan has sixteen banks, Arab banks controlling the local market and, account for more than 40 - 50% of the country's total production. Sekaran [22] defined research population as any exactly defined set of people, or collection of items, that is under investigation. In the light of this definition, the research population, and the actual sample are identified as the four banks that dominate the local market and account for more than 75% of the country's total production. These are Arab Bank, The Housing Bank for Trading Finance, Jordan Bank, and Standard Chartered Bank.

A questionnaire was sent to respondents in operation, middle and top management. A stratified random sampling method will be used, as it is the most convenient, and the most applicable in the Jordanian context.

The unit of analysis in this study is managers working in Jordan banks. 200 questionnaires were sent to 320 populations; 263 were returned, and 14 questionnaires were ignored because they were not returned. The overall response rate for this study is 82%, while the response rate actually used is 78%.

This is regarded as relatively high, since the respondents are managers who are supposed to be too busy to answer questionnaires. However, it is found that the sample is sufficient to represent the regression analysis conducted.

### *D. Operationalisation And Measurement Strategy Of The Model Variables:*

The measures of model variables in this study were analyzed using statistical procedures starting with internal consistency test and establishing constructs reliability. Statistical procedures are common among many researchers such as Malhotra et al. [23].

1) Internal consistency to assess the reliability of the scale using Cronbach's alpha.

2) Developing a structural model based on the composite measures to linking the hypothesized model's constructs.

3) Descriptive analysis of the mean and standard deviation of the investigated hypotheses.

### *E. Internal Reliability*

The internal consistency measures and Cronbach's alpha are obtained in order to assess the reliability of the measurement instruments. Table 1 shows the Cronbach's alpha value for each scale. It is clear that Cronbach alpha is acceptable statistically and managerially because α values are greater than the accepted 0.65%.

TABLE 1. Cronbach alpha

| Variable | Cronbach's alpha |
|---|---|
| Implementation M-commerce | 0.756 |
| Social influence | 0.732 |
| Usefulness | 0.678 |
| Ease of use | 0.692 |

### DESCRIPTIVE STATISTICS

Descriptive statistics such as means frequencies, and standard deviation, were used to identify the major characteristics of respondents in terms of their gender, age, educational level and working experience.

TABLE 2. DESCRIPTIVE STATISTICS

| Demographic object | Valid item | % |
|---|---|---|
| Gender | Male | 64.7 |
| | Female | 35.3 |
| Employee age (years) | 20-30 | 41.0 |
| | 30-40 | 35.38 |
| | 40-50 | 15.17 |
| | Above 50 | 8.45 |
| Education level | College degree | 1.5 |
| | Bachelor degree | 88.2 |
| | Postgraduate degree | 10.3 |
| Work experience (years) | Less than 6 | 29.0 |
| | From 1 - 2 | 25.0 |
| | 5 – 10 | 20.37 |
| | More than 10 | 25.63 |

## V. RESULTS

This study attempts to identify the technology acceptance model effects on implementation of mobile commerce e-banking services. The statistical results indicated the TAM effects of mobile services implementation. All analyses were conducted with SPSS. Frequency and percentage were used to describe the samples of the study and multiple regression analysis was conducted to test the research hypotheses.

### *A. Measuring The Effect Between The Independent Variable And Dependent Variable Simple Regression*

Most of the respondents agreed that there is a good and close relationship between TAM factor and mobile services implementations (Mean=3, 69; SD=0.98). The result of the regression analysis shows that there is a significant positive effect at the function level ($\alpha \leq 0.01$) which means that there is a relationship between TAM and mobile commerce implementation for the independent variables with a variance of 47.4%; thus, Hypothesis 1 was supported.

### *B. The Effect Between Social Influence And Mobile Commerce Implementation*

From the results in Table 3 which relate to correlation between the independent variable social influence and the dependent variable mobile commerce implementation, a positive and significant effect can be found at function level ($\alpha \leq 0.01$) which supports the hypothesis ($H_2$), where (r=0.499).

Based on the results in Table 3 which represent simple regression analysis, Simple regression was used to test the above hypothesis, and it was found that the calculated t (8.112) is significant at ($p<0.001$) level, which means that there is a relationship between social influence and mobile commerce implementation. It was observed that the variance was 41.1%, as such, Hypothesis 2 was supported.

TABLE 3. THE EFFECT BETWEEN SOCIAL INFLUENCE AND MOBILE COMMERCE IMPLEMENTATION

| Variable | Beta | Sig |
|---|---|---|
| Social influence and mobile commerce implementation | 0.499 | 0.000 |

### C. Effect Between Usefulness And Mobile Commerce Implementation

Based on the results in Table 4 which relate to correlation relationship between the independent variable usefulness and the dependent variable mobile commerce implementation, a positive and significant effect can be found at function level ($\alpha \leq 0.01$) which supports the hypothesis ($H_3$). Based on the results in Table 4 which represent simple regression analysis, simple regression was used to test the above hypothesis, and it was found that the calculated t (89.932) is significant at (0.01) level, which means that there is a relationship between them. A significant effect can be noticed at function level ($\alpha \leq 0.01$) to the independent variable usefulness in the dependent variable mobile commerce implementation. The variance was observed as 61.7%. Thus, hypothesis 3 is supported.

TABLE 4. THE EFFECT USEFULNESS AND MOBILE COMMERCE IMPLEMENTATION

| Variable | Beta | Sig |
|---|---|---|
| Usefulness and mobile commerce implementation | 0.393 | 0.000 |

### D. Effect between ease of use and mobile commerce implementation

Based on the results in Table 5 which relate to correlation relationship between the independent variable services, ease of use and the dependent variable mobile commerce implementation, a positive and significant effect can be found at function level ($\alpha \leq 0.01$), which supports the hypothesis ($H_4$) (Table 5). Simple regression was used to test the above hypothesis and it was found that the calculated t (7.641) is significant at (0.01) level; a significant effect can be found at function level ($\alpha \leq 0.01$). The variance is (0.681). Thus, hypothesis 4 is supported partially.

TABLE 5. The Effect between Ease of Use and Mobile Commerce Implementation

| Variable | Beta | Sig |
|---|---|---|
| Ease of use and mobile commerce implementation | 0.329 | 0.000 |

## VI. CONCLUSION

The researcher investigated the Technology acceptance model statistical in order find new way that can that supports involved employees and allows them to understand the m-commerce services in Jordan banks. The researcher argued that developing countries have to apply the technology acceptance model in order to help employees accept the new types of e-services.

A technology acceptance model is a success model for support using new services for electronic commerce. Social influence plays a significant role in influencing the mobile services in banks. Managers should focus on relative advantage of the usefulness and ease of use, in order to develop the mobile commerce implementation. Banks should introduce m-commerce service to revise their customer service need. In addition, banks should focus on service model and its applications. Managers should be demanding and challenging, so long as they are consistent in their treatment to usefulness of mobile commerce services. Finally, bank managers should make continuous evaluation and honest dialogue about the ease of use of mobile commerce services.

However, regardless of their limitations, it is useful to continue analysing the new types of e-services through the technology acceptance model. By doing so, first it is expected that deeper insights would be gained into the banking industry itself and its evolution over time. Secondly, this study hopes to develop the evaluation framework that was used, which can be transferred to other types of industries.

## ACKNOWLEDGEMENT

The author is grateful to the Applied Science Private University, Amman, Jordan, for the financial support granted to cover the publication fee of this research article.

## REFERENCES

[1] Alawneh, A., and Hattab, E. (2008b). **"**E-business Value Creation Jordanian Banking Services Industry: An Empirical Analysis of Key Factors". International Arab Conference on e- technology. Arab Open University, Amman-Jordan. October 15-10, 2010.

[2] Deng, H., Li, W. and Agrawal, D.P. (2002a) 'Routing security in Ad Hoc networks', IEEE Communications Magazine, Special Topics on Security in Telecommunication Networks, October 2002, Vol. 40, No. 10, pp.70–75.

[3] Fisher, R.J. and Price, L.L. (1992) 'An investigation into the social context of early adoption behavior', Journal of Consumer Research, Vol. 19, pp.477–486.

[4] Kim, H., Lee, I. and Kim, J. (2008) 'Maintaining continuers vs. converting discontinuers: relative importance of post-adoption factors for mobile data services', Int. J. Mobile Communications, Vol. 6, No. 1, pp.108–132.

[5] Lin, Y-M. and Shih, D-H. (2008) 'Deconstructing mobile commerce service with continuance intention', Int. J. Mobile Communications, Vol. 6, No. 1, pp.67–87.

[6] Liu, C-C. (2008) 'Mobile phone user types by Q methodology: an exploratory research', Int. J. Mobile Communications, Vol. 6, No. 1, pp.16–31.

[7] Miranda, F.J. Cortés, R and Barriuso, C. (2006) "Quantitative Evaluation of e-Banking Web Sites: an Empirical Study of Spanish Banks" The Electronic Journal Information Systems Evaluation Volume 9 Issue 2, pp 73 - 82, available online at www.ejise.com

[8] Novak, T. P., and Hoffman, D. (1997) "Modeling the Structure of Flow Experience Among Web Users," Vanderbilt University.

[9] Orr, B. (2004) E-Banking job one: Give customers a good ride. ABA Banking Journal, Vol. 96, Iss. 5, pp 56-57.

[10] pearlson , C, Saunders ,S (2009) strategic management information systems , Forth edition , Johan Wiley & sons, INC.

[11] Scornavacca, E., Prasad, M. and Lehmann, H. (2006) 'Exploring the organizational impact and perceived benefits of wireless Personal Digital Assistants in restaurants', Int. J. Mobile Communications, Vol. 4, pp.558–567.

[12] Selz, D. and Schubert, P. (1997): "Web assessment: A model for the evaluation and the assessment of successful electronic commerce applications". Electronic Markets, vol 7, nº 3, pp.46-48.

[13] Stuart, J.B. and Sid, L.H. (2003) 'Rising sun: iMode and the wireless internet', Communications of the ACM, Vol. 46, pp.78–84.

[14] [Venkatesh, V., Ramesh, V. and Massey, A.P. (2003) 'Understanding usability in mobile commerce', Communications of the ACM, Vol. 46, pp.53–56.

[15] venkatesh,,V., Morreis,MG., Davis, F.D. (2003) User acceptance of information technology :Toward a unifeied view. MIS Quarterly, 27(3), 435- 478.

[16] zMiranda, F.J. Cortés, R and Barriuso, C. (2006) "Quantitative Evaluation of e-Banking Web Sites: an Empirical Study of Spanish Banks" *The Electronic Journal Information Systems Evaluation* Volume 9 Issue 2, pp 73 - 82, available online at www.ejise.

[17] Watson, R.T., Pitt, L.F., Berthon, P. and Zinkhan, G.M. (2002) 'U-commerce: expanding the universe of marketing', Journal of the Academy of Marketing Sciences, Vol. 30, pp.333–347.

[18] Weber, M., Denk, M., Oberecker, K., Strauss, C. and Stummer, C. (2008) 'Panel surveys go mobile', Int. J. Mobile Communications, Vol. 6, No. 1, pp.88–107.

[19] Xie, B., Kumar, A. and Agrawal, D.P. (2008) 'enabling multi-service on 3G and beyond: challenges and future directions', IEEE Wireless Communication Magazine.

[20] Xie, B., Kumar, A., Zhao, D., Reddy, R. and He, B. (2010) 'On secure communication in integrated heterogeneous wireless networks', Int. J. Information Technology, Communications and Convergence, Vol. 1, No. 1, pp.4-23.

[21] Wells, N., and J. Wolfers, "Finance with a personalized touch," Communication of the ACM, Vol. 43, No. 8: 30-34, 2000.

[22] Sekaran ,U. (1992 ). Research Methods for Business: A skill Building Approach, Second Edtion, alynow university press, USA p (285-301).

[23] Malhotra, K.and Briks D., (2000). Marketing Research: An Applied Approach, European Edition , prentice – Hall.

# Clone Detection Using DIFF Algorithm For Aspect Mining

Rowyda Mohammed Abd El-Aziz
Department of Computer Science
Faculty of Computers and Information
Helwan University
Cairo, Egypt

Amal Elsayed Aboutabl
Department of Computer Science
Faculty of Computers and Information
Helwan University
Cairo, Egypt

Mostafa-Sami Mostafa
Department of Computer Science
Faculty of Computers and Information
Helwan University
Cairo, Egypt

***Abstract*— Aspect mining is a reverse engineering process that aims at mining legacy systems to discover crosscutting concerns to be refactored into aspects. This process improves system reusability and maintainability. But, locating crosscutting concerns in legacy systems manually is very difficult and causes many errors. So, there is a need for automated techniques that can discover crosscutting concerns in source code. Aspect mining approaches are automated techniques that vary according to the type of crosscutting concerns symptoms they search for. Code duplication is one of such symptoms which risks software maintenance and evolution. So, many code clone detection techniques have been proposed to find this duplicated code in legacy systems. In this paper, we present a clone detection technique to extract exact clones from object-oriented source code using Differential File Comparison Algorithm (DIFF) to improve system reusability and maintainability which is a major objective of aspect mining.**

***Keywords- aspect mining; reverse engineering; clone detection; DIFF algorithm.***

## I. INTRODUCTION

In software engineering, it is essential to manage the complexity and evolution of software systems. Hence, decomposing large software systems into smaller units is required. The result of this decomposition is separation of concerns that leads to facilitating parallel work, team specialization, quality assurance and work planning [1].

However, there are some functionalities that cannot be assigned to a single unit because the code implementing them is scattered over many units and tangled with other units. Such functionalities are called *crosscutting concerns* [2]. The existence of these crosscutting concerns leads to reducing maintainability, evolution and reliability of software systems.

Aspect Oriented Software Development (AOSD) is a new programming paradigm that solves the problem of crosscutting concerns existence in legacy systems. Aspect oriented programming modularizes such crosscutting concerns in new units called *aspects* and introduces ways for weaving aspect code with the system code at the appropriate places [3]. The success of aspect oriented programming directs software engineers to a new research area called *aspect mining*. Aspect mining is a specialized reverse engineering process which aims at discovering crosscutting concerns automatically in existing systems. This process improves system maintainability and evolution and reduces system complexity. It also enables migration from object-oriented to aspect-oriented systems in an efficient way [4][5][6]. Aspect mining approaches vary according to the type of crosscutting concerns symptoms they search for. Code duplication is one of the main symptoms of crosscutting concerns. It is considered a major problem for large industrial software systems because it increases their complexity and maintenance cost. So, many clone detection techniques are used to find this duplicated code in legacy systems and will be discussed in details in section 2. In this paper, we present a clone detection technique to extract exact clones from object-oriented source code using Differential File Comparison Algorithm (DIFF).

The basic idea is to find different lines of code between two source code files using Diff Algorithm. As a consequence, the remaining lines of code in both files are identical and considered clones. Clones can then be extracted from files. Finding clones in source code as a symptom of crosscutting concerns helps in improving system reusability and maintainability which is the aim of aspect mining. In section 2, previous work on clone detection techniques is presented. In section 3, we describe the basic idea of the used technique to detect clones in source code. In section 4, experimental work and results are discussed. Finally, conclusion and future work are presented in section 5.

## II. PREVIOUS WORK

Previous studies report that about 5% to 20% of software systems contain code duplication which is a consequence of copying existing code fragments and then reusing them by pasting with or without minor modifications instead of rewriting similar code from scratch [7]. Therefore, it is considered a common activity in software development. Developers perform this activity to reduce programming time and effort. However, this activity results into software systems which are difficult to maintain. The reason is that if a bug is detected in a code fragment, other similar code fragments have to be checked for the same bug. Consequently, there is a need

for automated techniques that can find duplicated code fragments in source code such as clone detection techniques.

### A. Clone Detection Techniques

Clone detection techniques can be categorized into the following [8]:

- String-based techniques (also called text-based techniques): at the beginning, little or no transformation in raw source code is performed; for example, white spaces and comments are ignored. Then, the source code is divided into a number of strings (lines). These strings are compared according to the used algorithm to find duplicated ones [9].
- Token-based techniques: use lexical analysis for tokenizing source code into a stream of tokens used as a basis for clone detection.
- AST-based techniques: use parsing to represent source code as an abstract syntax tree (AST) [10]. Then, clone detection algorithm compares similar sub-trees in this tree.
- PDG-based techniques: use Program Dependence Graphs (PDGs) to represent source code [11]. PDGs describe the semantic nature of source code in high abstraction such as control and data flow of the program.
- Metrics-based techniques: hashing algorithms are used in such techniques [12]. A number of metrics are calculated for each code fragment in source code. Then, code fragments are compared to find similar ones.

### B. Clone Terminology

When two code fragments are identical or similar, they are called *clones*. There are four types of clones: Type I, Type II, Type III and Type IV. Each of these four types of clones belongs to one of two classes according to the type of similarity it represents: textual similarity or functional similarity. In this context, clones of Type I, Type II and Type III are categorized under textual similarity and Type IV is categorized under functional similarity [13].

- Type I: is called exact clones where a copied code fragment is identical to the original code fragment except for some possible variations in whitespaces and comments.
- Type II: a copied code fragment is identical to the original code fragment except for some possible variations about user-defined identifiers (name of variables, constants, methods, classes and so on), types, layout and comments.
- Type III: a copied code fragment is modified by changing the structure of the original code fragment, e.g. adding or removing some statements.
- Type IV: in this type, clones have semantic similarity between code fragments. Clones, according to this type, are not necessarily copied from the original code because sometimes, they have the same logic and are similar in their functionalities but developed by different developers.

## III. Proposed Technique

In this paper, a clone detection technique is presented using Differential File Comparison Algorithm (DIFF) [14] to detect exact clones in source code files. Our clone detection technique passes through three stages:

- Source code normalization: this stage acts as a preprocessing stage. Our clone detection technique is text-based and, therefore, a little transformation of the source code is needed. White spaces and comments are removed at this stage.
- Differential File Comparison: This is the main stage of the proposed technique. The Differential File Comparison algorithm (DIFF) [14] determines differences of lines between two files. It solves the problem of 'longest common subsequence' by finding the lines that are not changed between files. So, its goal is to maximize the number of lines left unchanged. An advantage of the DIFF algorithm is that it makes efficient use of time and space. So, this idea is used to find differences in source code lines between two files.
- Extracting exact clones: After finding differences in source code lines between the two given source code files using the DIFF Algorithm, the remaining lines of code in both files are identical and considered clones. The complement of the difference between 2 files is determined which results in extracting exact clones from two given source code files.

The main steps of DIFF algorithm are summarized as follows [14]:

1. Determine equivalence classes in file 2 and associate them with lines in file 1. Hashing is used to get better optimization when comparing large files (thousands of lines).
2. Find the longest common subsequence of lines.
3. Get a more convenient representation for the longest common subsequence.
4. Weed out spurious sequences called jackpots.

## IV. Experimental Work And Results

Our experiment was conducted on a simple case study consisting of two source code files implemented in the C# programming language. These files have some differences and similarities in their lines of code as shown in figure 1. At the beginning, the two files are normalized by removing white spaces and comments. Then, they are compared using DIFF algorithm and the differences in source code lines between both files are highlighted as shown in figure 2.

```
class Program {
 public int sumElements(int[] arr){
 int sum = 0;
 for (int i = 0; i < 5; i++)
 {
 sum += arr[i];
 }
 return sum;
 }
 static void Main(string[] args)
 {
 Program p = new Program();
 int result;
 int avg;
 int arr = new int[5];
 int size = arr.Length;
 Console.WriteLine("Enter
numbers:");
 for (int i = 0; i < 5; i++)
 arr[i]=
int.Parse(Console.ReadLine());
// sum of array elements
 result = p.sumElements(arr);
 // average of array elements
  avg = result / size;
 Console.WriteLine("Addition is:"
+  result);
 Console.WriteLine("Average is:"
+ avg);
 }}
```

```
class Prog {
 public float sumElement(float[] arr) {
 int sum = 1;
 for (int i = 0; i < 5; i++)
 {
 sum += arr[i];
 }
return sum;
}
 static void Main(string[] args)
 {
 Prog p = new Prog();
 float result;
 float avg;
 float arr = new float[5];
  int size = arr.Length;
 Console.WriteLine("Enter numbers:");
for (int j = 0; j < 5; j++)
arr[j] = int.Parse(Console.ReadLine());
// sum of array elements
  result = p.sumElements(arr);
 // average of array elements
  avrg = result / size;
Console.WriteLine("Addition is:" +
result);
Console.WriteLine("Average is:" +
avg);
 }}
```

Figure1. Two source code files

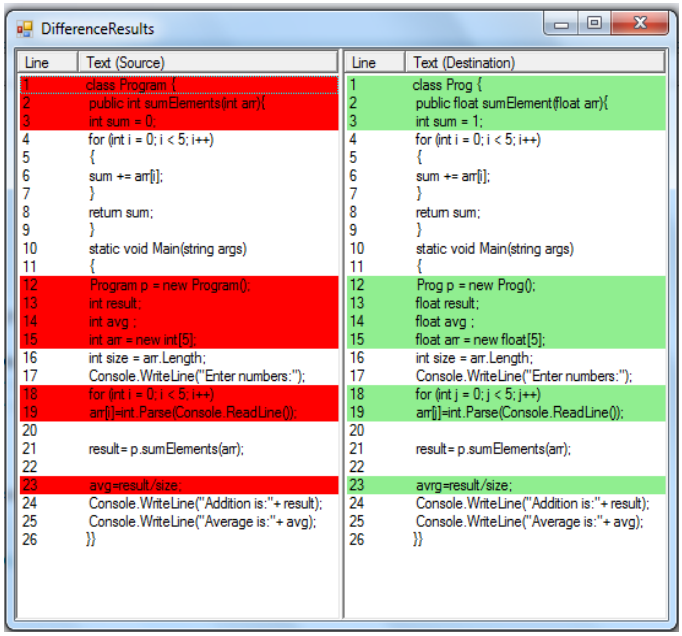



Figure2. Difference between lines of code

Finally, exact cloned lines of code are detected in both files after removing those differences from source code lines as shown in figure 3.

Clone Detective tool [15] [16] is a Visual Studio integration that allows analyzing C# projects for source code that is duplicated somewhere else. Clone Detective tool is supposed to detect type I and type II clones but it may miss some clones as explained in [17].

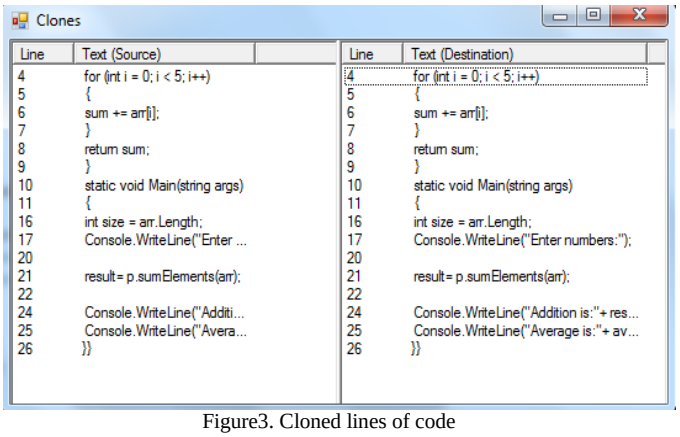



Figure3. Cloned lines of code

By comparing our results with those obtained from the Clone Detective tool for Visual Studio 2008 using the same case study; it is found that the Clone Detective tool cannot detect all the differences in lines of code whereas our proposed technique can do that.

Table 1 shows the results of comparing the two tools regarding the total number of lines in each file and the total number of cloned lines between two files with setting clone minimum length equals to one. It is noticed that our proposed technique can detect all exact cloned lines which are actually 14 lines but Clone Detective tool detects 24 cloned lines and this is not accurate because only 14 lines are exact clones and other lines are different.

Table1.Comparison of results obtained by the proposed technique and the Clone Detective tool

| Comparison | | Total number of lines | Total number of cloned lines |
|---|---|---|---|
| Proposed Technique | Source | 26 | 14 |
| | Destination | 26 | 14 |
| Clone Detective | Source | 26 | 24 |
| | Destination | 26 | 24 |

## V. Conclusion And Future Work

We present a simple clone detector to discover code cloning which is a symptom of crosscutting concerns existence in software systems. Detection of code clones decreases maintenance cost, increases understandability of the system and helps in obtaining better reusability and maintainability which is the aim of aspect mining .The technique is experimented on a simple case study (two source code files) and finally exact clones are extracted from source code.

We consider this tool as a starting point towards a complete clone detection system. In the future, this tool can be extended to detect type II and type III clones and mine source code written in other programming languages, not only C#. It can also be extended to work on more than two source code files.

## REFERENCES

[1] Arie van Deursen, Marius Marin and Leon Moonen, "Aspect Mining and Refactoring", In Proceedings of the First International Workshop on REFactoring: Achievements, Challenges, Effects (REFACE03), 2003.

[2] Bounour Nora and Ghoul Said, "A model-driven Approach to Aspect Mining", Information Technology Journal ,vol.5, 2006 , pp. 573-576.

[3] M.Marin, A.vanDeursen and L.Moonen ,"Identifying Crosscutting Concerns Using Fan-In Analysis",ACM Transactions on Software Engineering and Methodology, Vol. 17, December 2007.

[4] Bounour Nora, Ghoul Said and Atil Fadila, "A Comparative Classification of Aspect Mining Approaches", Journal of Computer Science,vol. 2 , pp. 322-325, 2006.

[5] Chanchal Kumar Roy, Mohammad Gias Uddin, Banani Royand Thomas R. Dean,"Evaluating Aspect Mining Techniques: A Case Study", 15th IEEE International Conference on Program Comprehension (ICPC'07), 2007.

[6] Andy Kellens, Kim Mens, and Paolo Tonella, "A Survey of Automated Code-Level Aspect Mining Techniques",In Transactions on Aspect Oriented Software Development, Vol. 4 (LNCS 4640), pp. 145-164, 2007.

[7] Chanchal Kumar Roy and James R. Cordy, "A Survey on Software Clone Detection Research", Technical Report No.2007-541, School of Computing,Queen's University, KingstonOntario, Canada, September 2007.

[8] Magiel Bruntink, "Aspect Mining using Clone Class Metrics", In Proceedings of the 1st Workshop on Aspect Reverse Engineering, 2004.

[9] Kunal Pandove,"Three Stage Transformation for Software Clone Detection", Master Thesis,Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Deemed University,May 2005.

[10] Ira D. Baxter, Andrew Yahin,Leonardo Moura, Marcelo Sant'Anna and Lorraine Bier,"Clone Detection Using Abstract Syntax Trees", In Proceedings of the 14th International Conference on Software Maintenance (ICSM'98), pp. 368-377, Bethesda, Maryland, November 1998.

[11] Jens Krinke, "Identifying Similar Code with Program Dependence Graphs", In Proceedings of the 8th Working Conference on Reverse Engineering (WCRE'01), pp. 301-309,Stuttgart, Germany, October 2001.

[12] Jean Mayrand, Claude Leblanc and Ettore M. Merlo, "Experiment on the Automatic Detectionof Function Clones in a Software System Using Metrics", In Proceedings of the International Conference on Software Maintenance (ICSM '96),1996.

[13] Yogita Sharma "Hybrid Technique for Object Oriented Software Clone Detection", Master Thesis,Computer Science and Engineering Department,Thapar University, June 2011.

[14] J.W.Hunt and M.D.McIlroy, "An Algorithm for Differential File Comparison", Bell Laboratories, Murray Hill, New Jersey, 1976.

[15] http://clonedetectivevs.codeplex.com, last accessed Augest 2012.

[16] Elmar Juergens, Florian Deissenboeck and Benjamin Hummel, "CloneDetective–A Workbench for Clone Detection Research", In Proccedings of the 30th International Conference on Software Engineering (ICSE), 2009.

[17] Chanchal K. Roy, James R. Cordy and Rainer Koschke,"Comparison and Evaluation of Code Clone Detection Techniques and Tools: A Qualitative Approach", Science of Computer Programming Journal, February 2009.

## AUTHORS PROFILE

**Rowyda Mohammed Abd El-Aziz** is currently a Software Developer at the Ministry of Planning, Cairo, Egypt. She worked as Teaching Assistant in Modern Sciences and Arts University in Egypt for four years. She is a Masters Student at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. Her current research interests include software engineering and Human Computer Interaction.

**Amal Elsayed Aboutabl** is currently an Assistant Professor at the Computer Science Department, Faculty of Computers and Information, Helwan University, Cairo, Egypt. She received her B.Sc. in Computer Science from the American University in Cairo and both of her M.Sc. and Ph.D. in Computer Science from Cairo University. She worked for IBM and ICL in Egypt for seven years. She was also a Fulbright Scholar at the Department of Computer Science, University of Virginia, USA. Her current research interests include parallel computing, image processing and software engineering.

**Mostafa-Sami M. Mostafa** is currently a Professor of computer science, Faculty of Computers and Information, Helwan University, Cairo, Egypt. He worked as an Ex-Dean of faculty of Computers and Information Technology, MUST, Cairo. He worked also as an Ex-Dean of student affairs and Ex-Head of Computer Science Department, faculty of Computers and Information, Helwan University, Cairo, Egypt. He is a Computer Engineer graduated 1967, MTC, Cairo, Egypt. He received his MSC 1977 and his PhD 1980 from University of Paul Sabatier, Toulouse, France. His research activities are in Software Engineering and Computer Networking. He is awarded supervising more than 80 Masters of Sc. and 18 PhDs in system modeling and design, software testing, middleware system development, real-time systems, computer graphics and animation, virtual reality, network security, wireless sensor networks and biomedical engineering.

# On the Projection Matrices Influence in the Classification of Compressed Sensed ECG Signals

Monica Fira, Liviu Goras
Institute of Computer Science
Romanian Academy
Iasi, Romania

Liviu Goras, Nicolae Cleju, Constantin Barabasa
Faculty of Electronics, Telecommunications and Information Technology
"Gheorghe Asachi" Technical University of Iasi
Iasi, Romania

***Abstract*— In this paper the classification results of compressed sensed ECG signals based on various types of projection matrices is investigated. The compressed signals are classified using the KNN (K-Nearest Neighbour) algorithm. A comparative analysis is made with respect to the projection matrices used, as well as of the results obtained in the case of the original (uncompressed) signals for various compression ratios. For Bernoulli projection matrices it has been observed that the classification results for compressed cardiac cycles are comparable to those obtained for uncompressed cardiac cycles. Thus, for normal uncompressed cardiac cycles a classification ratio of 91.33% was obtained, while for the signals compressed with a Bernoulli matrix, up to a compression ratio of 15:1 classification rates of approximately 93% were obtained. Significant improvements of classification in the compressed space take place up to a compression ratio of 30:1.**

***Keywords- ECG; compressed sensing; projection matrix; classification; KNN.***

## I. INTRODUCTION

In the last decade, a new concept regarding the acquisition, analysis, synthesis and reconstruction of signals was introduced. Known under several equivalent names: **compressed/compressive sampling** or **sensing** (acquisition/detection by compression), it speculates the sparsity of various classes of signals with respect to certain basis or dictionaries. In the following we refer to a signal f (including a biomedical one) which is a member of a class $F \subset R^N$ of ND discrete signal, in particular 1D temporal or 2D spatial signals (images). We ask the question of correlating the properties of the class F to the minimum number of measurements necessary for coding the signal f with a Euclidean metric recovery error, ε, imposed, respectively $\|f-f_\varepsilon\|_{l2} \le \varepsilon$. The compressed sensing concept relies on an important result obtained by Candes and Tao [1-4] namely that if the signals of the class F admit representations through a small number of components in an adequately selected base, i.e. they are *sparse* in that basis, it is possible to reconstruct them with a very good precision from a small number of random measurements by solving a simple problem of linear programming. Specifically, it is shown that *if* the the n-th component f(n) of a signal in a given base, whose values in descending order satisfy the relation $|f|(n) \le Rn^{-1/p}$ with R,p > 0 (which represents a constraint on the descending speed of the components) and K measurements (projections) of the form

$$yk = <X_k, f>,\ k=1,\ldots,K,$$

are performed, where $X_k$ are N-dimensional Gaussian independent vectors with normal standard distribution, *then* any signal that meets the mentioned constraint for a given *p* can be reconstructed with a very high probability in the form of a $f^{\#}$ signal defined as a solution of minimum norm $l_1$ of the system $yk = <Xk, f^{\#}>$ with the relationship

$$\|f-f\#\|l2 \le C_p R(K/\log N)^{-r}$$

where

$$r = 1/p-1/2.$$

The result is optimal in the sense that it is generally impossible to obtain a better precision out of K measurements regardless of the mode in which these measurements are performed.

Reformulating the main problem, the situation can be regarded as the one of recovering a signal $f \in R^N$ using a minimum number of measurements, i.e. of linear functionals associated to the signal, so that the Euclidean distance $l_2$ between the initial and the reconstructed signal to be lower than an imposed value ε.

## II. METHODOLOGY AND OBJECTIVE

Assuming the existence of a dictionary D of elements $\{\underline{d}_k\}^L{}_{k=1}$ with L>N, each column of the dictionary is a normalized vector ($\|d_k\|^2 = <d_k, d_k> = 1$) belonging to $C^N$ that will be called atom. The dictionary contains L vectors and can be viewed as a matrix of size N*x*L. An example is the Coifman dictionary which contains L=N*log*N elements consisting of attenuated harmonic waveforms of various durations and localizations. Other types of dictionaries are those proposed by Ron and Shen [5] or the combined ridglet/wavelet systems proposed by Starck, Candes and Donoho [6].

For a given sparse signal $\underline{S} \in C^N$ the determination of the vector of coefficients $\underline{\gamma}$ with the highest number of null elements belonging to $C^L$ so that $D\underline{\gamma} = \underline{S}$ is envisaged.

Formally the problem consists of solving the optimization problem:

$$(P_0) \quad \min \|\underline{\gamma}\|_0 \quad \text{subject to} \quad \underline{S} = D\underline{\gamma}$$

where the norm $l_0$ is the number of non-zero elements in γ. Unfortunately the problem is rarely easy to solve. Since in general L>>N the solution is not unique. Determining the solution of the problem ($P_0$) requires enumerating all subsets of the dictionary and finding the smallest subset which can be able to represent the signal.

A remarkable result [2] is that for a large number of dictionaries, the determination of sparse solutions can be achieved based on the convex optimization, respectively by solving the problem

$$(P_1) \quad \min \|\underline{\gamma}\|_1 \quad \text{subject to} \quad \underline{S} = D\underline{\gamma}$$

Intuitively, using the norm $l_1$ can be regarded as a convexification of the problem ($P_0$). The convex optimization problems are well studied and there are numerous algorithms and software; as already mentioned, the problem ($P_1$) is a linear programming problem and can be solved by interior point type methods even for large values of N and L. The possibility of solving a problem P0 by solving problem ($P_1$) may seem surprising. However, there are results which ensure in a rigorous manner the fact that, if there is a highly sparse solution for the problem ($P_0$) then it is identical to the solution of the problem ($P_1$). Conversely, if the solution of the problem ($P_1$) is sparse enough, i.e., if the sparsity degree is below a certain threshold, then it is ensured the fact that this is also the solution for the problem ($P_0$).

In order to obtain the representation of the signals in overcomplete dictionaries several methods have been proposed in the past few years, such as the „*method of frames*", „*matching pursuit*", „*basis pursuit*" (BP), as well as the „*method of best orthogonal basis*" [2].

A possibility of improving the results of the reconstruction when using the concept of compressed sensing is to use specific dictionaries, constructed according to the nature, particularities, statistics or the type of the compressed signal. Thus, there are algorithms [7] which on reconstruction will use a certain dictionary selected from a series of several available dictionaries, namely, the dedicated dictionary constructed for that particular class of signals. These types of reconstruction algorithms have the advantage of a good reconstruction, but they require additional information related to the initial signal, based on which it will be decided on the dictionary used on reconstruction. A solution to this problem would be the correct classification of the original signal or of the compressed signal. For biomedical signals this classification of the signal involves placing the signal into one of several predefined pathological classes for which there exist specific dictionaries. In practical applications, this classification of the original signal is not possible or it requires an additional effort. Therefore, the ideal solution (which does not require an extra effort in the compression stage) is to classify the compressed signal during the reconstruction stage. In other words, for the classification of the compressed signal [10], the problem of classification is moved from the compression stage into the reconstruction stage [8].

In this paper we investigate the possibility of classification of the ECG signals after their compression based on the concept of compressive sensing. In order to obtain good results both from the classification point of view and from the point of view of the reconstruction, we will segment the ECG signal into cardiac cycles which will be further compressed. In other words, ECG segments will be used (cardiac cycles) and the ECG signal will be reconstructed by concatenating these cardiac segments (cycles). According to the algorithm described in [9] the segmentation of the ECG signal into cardiac cycles is achieved based on the R waves detection. Thus, one cardiac cycle is represented by the ECG signal between the middle of a RR segment and the middle of the next RR segment, where the RR segment means the ECG waveform between two successive R waves. Figure 1 represents the segmentation of the ECG signal. After the segmentation of the ECG signal there is a centering of the R wave which is made by resampling on 150 samples on both sides of the R wave. In this way all cardiac cycles will have size 301 and the R wave will be positioned on the sample 151 [9].

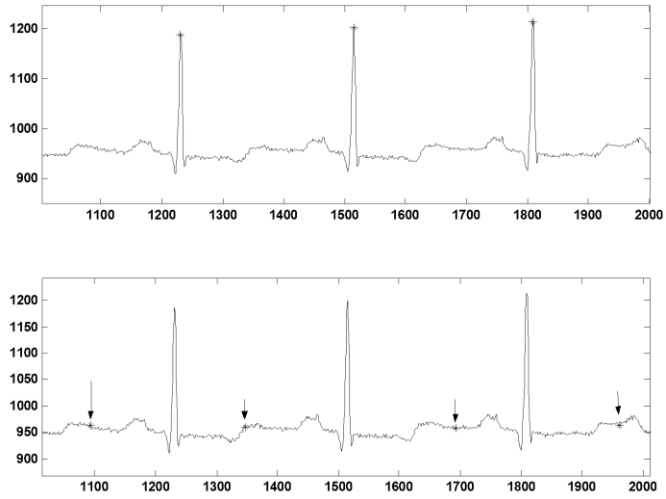



Figure 1. Segmented ECG signal [9]

In order to compress the signals obtained this way, based on the concepts of compressed sensing, a KxN projection matrix of measurements has been used. The compression ratio depends on the value of K. Due to the fact that the original ECG segments have the size of 301 (because there was a resampling of the cardiac cycles and all cycles have been resampled on 301 samples), the projection matrix will have one of the dimensions 301, *N = 301*, and the other dimension of the matrix, K, will represent the number of measurements. Thus, if the projection matrix has the size 20x301, it means that for the compression of any cardiac cycle of size 301 only 20 measurements will be taken, resulting a compressed version of any cardiac cycle of size 20, which means a compression ratio of 15:1.

For the classification of the compressed cardiac cycles we used the KNN classifier with an Euclidean distance type, and

the decision of belonging to a certain class was based on the nearest neighbor.

A data set of 5601 compressed cardiac cycles, 701 cardiac cycles from each of all the 8 classes (normal and 7 pathological classes) was constructed.

In order to train the KNN classifier we used 1500 cardiac cycles, the testing being made on the rest of the data from the database.

We tested several types of projection matrices (Gaussian random, Fourier, random with elements of -1, 0, 1, etc). Together with the type of matrices, the number of measurements was varied from 2 to 60 (equivalent to compression ratio between 150:1 and 5:1). Thus, using different types of matrices, an analysis of the classification of the compressed cardiac cycles for various compression ratios was performed.

The following types of matrices were used:

- *Random projection matrix* (marked on graphs with *random*): all entries of the K × N projection matrix are independent standard normal distributed random variables.
- *Matrices with zeros and ones, with a predefined number of ones (3, 5, 7, 10, 50 or 150) randomly distributed across each measurement* (marked on graphs with *V1_3, V1_5, V1_7, V1_10 , V1_50 or V1_150*)
- *Matrices with zeros and ones, with a predefined number of ones (3, 5, 7, 10, 50 or 150) randomly distributed across each of the N matrix columns (*marked on graphs with *V1m_3, V1m_5, V1m_10 or V1m_15*)
- *Random projection matrices with values of -1, 0 and 1 uniformly distributed* (marked on graphs with *V_1_0_-1 (1/3 1/3 1/3)*) i.e. Bernoulli matrix with constant distribution
- *Random projection matrix with values of -1, 0 and 1, and unequal distribution* (marked on graphs with *V_1_0_-1 (1/4 1/2 1/4)*) i.e. Bernoulli matrix
- *Matrices with 1 and -1, with a predefined (5, 50 or 150) number of 1's randomly distributed across each measurement* (note on graphs with *V-1_5, V-1_50 or V-1_150*)
- *Random Fourier matrix:* The signal is a discrete function f on Z/NZ, and the measurements are the Fourier coefficients at a randomly selected set of frequencies of size K (K < N).
- *Random projection matrix with 0 and 1* (marked on graphs with *V_0_1_random*): all entries of the N × K projection matrix are independent standard normally distributed random variables.

## III. Experimental Results And Discussions

A number of 24 ECG annotated recordings from the MIT-BIH Arrhythmia database have been used to test the possibility of the classification of compressed patterns [30]. The ECG signals were initially digitized through sampling at 360 samples per second, quantized and encoded with 11 bits and then resampled as described above.

Based on the database annotations, eight major classes have been identified, namely a class of normal cardiac beats and seven classes of pathological beats: atrial premature beat, left bundle branch block beat, right bundle branch block beat, premature ventricular contraction, fusion of ventricular and normal beat, paced beat, fusion of paced and normal beat.

For the resampled cardiac cycles, but without compression, using for training 1500 cycles and using the KNN algorithm, we found a classification ratio of 91.33%.

In Figure 2 the classification curves for various projection matrices are represented. Very good results have been obtained for the Bernoulli matrix, namely for projection matrices with values of -1, 0 and 1, in equal proportion (⅓,⅓,⅓) or variable proportions (¼, ½ , ¼). Also, very good results were obtained for the projection matrix containing only the elements of 0 and -1, in equal proportions (½ and ½), which, in fact, is a custom Bernoulli matrix.

From the point of view of the results, the second best projection matrix is random with independent standard normal distributed random variables entries.

The weakest results are obtained with the matrix containing values of 0 and -1, with a number of 5 non-zero elements. The difference between the results obtained with this matrix and the next matrix from the classification point of view are high, namely from 50% in case of a compression of 30:1 obtained with the matrix V-1_5, to approximately 70% for compression of 30:1 with the Fourier matrix.

In Figure 3 the results for three compression ratios, 20:1, 30:1 and 60:1 are presented.

It is also observed that for a compression ratio lower than 20:1 the results of the classification do not improve significantly, i.e. one observes a stabilization of the classification ratio. Also, between the compression of 20:1 and 30:1 the improvement of the classification ratio is small, therefore choosing the classification ratio will be based on the sparsity of the signal, which will implicitly influence the reconstruction errors also.

Another aspect to be mentioned, and which is especially important for hardware implementations of compressed sensing devices, is that in the case of projection matrices which contain only elements of -1, 0 and 1 there is the advantage of reducing the number of calculations required for compression. If in the case of random matrices used for compression a significant number of multiplications is necessary, for matrices with elements -1, 0 and 1 (Bernoulli matrices) we need only a small number of additions.

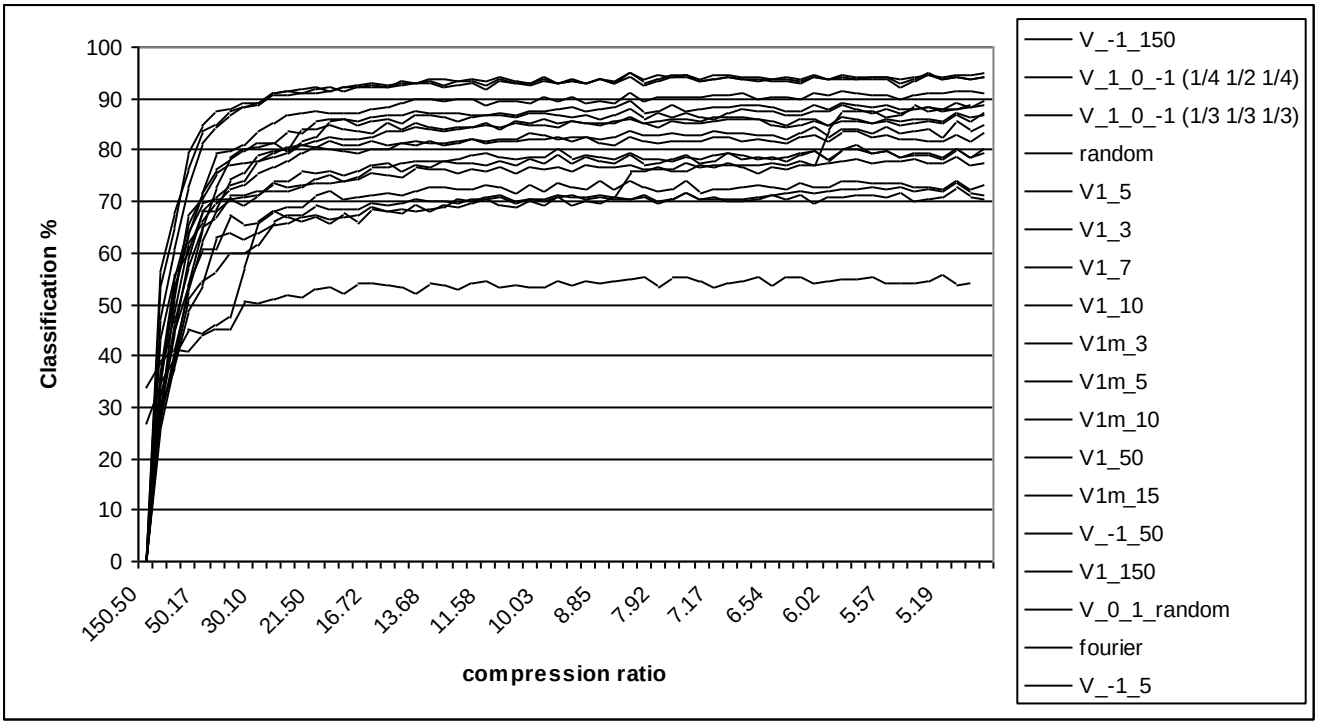



Figure 2. The compression ratio vs. classification% for various projection matrices

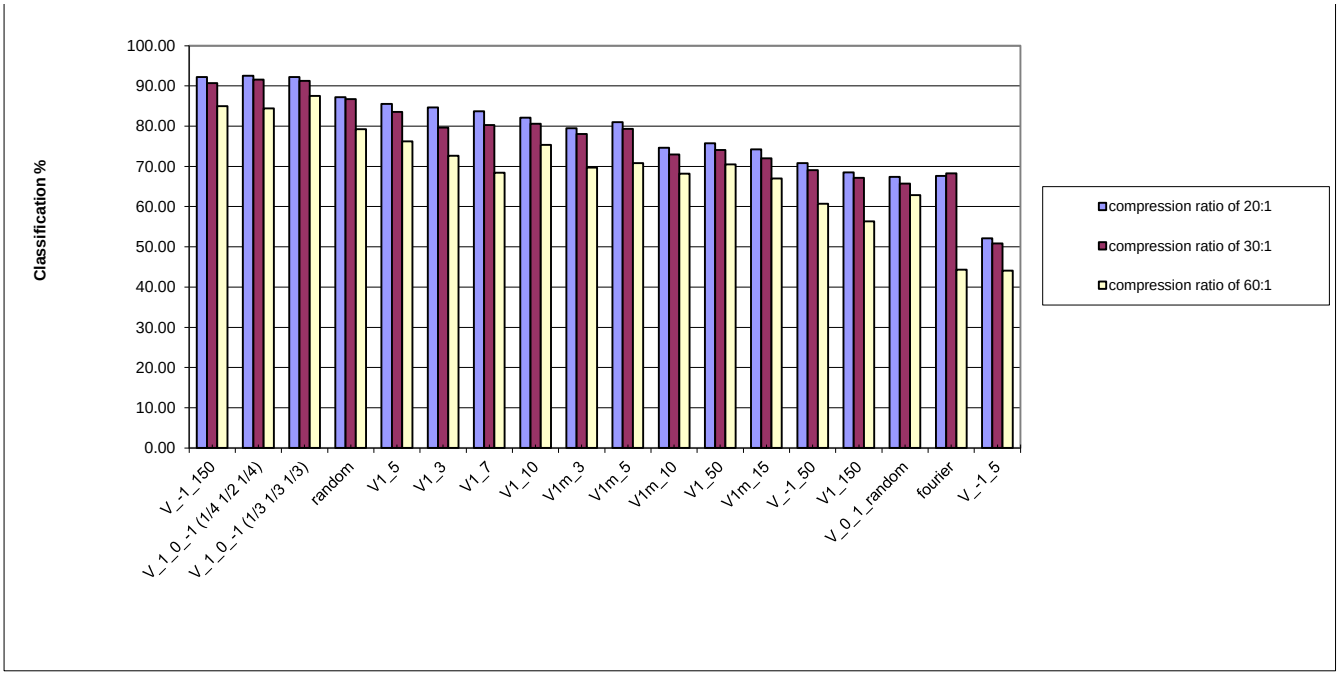



Figure 3. The compression ratio of 20:1, 30:1 and 60:1 vs. classification% for various projection matrices

## IV. CONCLUSIONS

This paper presents a comparative analysis of the classification results for compressively sensed cardiac cycles, using different project matrices and a variable number of measurements.

The classification of cardiac cycles is made using the KNN algorithm and the construction of the projection matrices is varied, including random matrices with real numbers, Bernoulli matrices, random matrices with elements of -1, 0 and 1 with different probabilities, random matrices with values of 0 and 1 and normal distribution, etc.

For Bernoulli projection matrices it has been observed that the classification results for compressed cardiac cycles are comparable to those obtained for uncompressed cardiac

cycles. Thus, for normal uncompressed cardiac cycles a classification ratio of 91.33% was obtained, while for the signals compressed with a Bernoulli matrix, up to a compression ratio of 15:1 classification rates of approximately 93% were obtained.

Significant improvements of classification in the compressed space take place up to a compression ratio of 30:1.

## ACKNOWLEDGMENT

This work has been supported by CNCSIS –UEFISCSU, project PNII – RU - PD 347/2010 (M. Fira)

This paper was realized with the support of EURODOC "Doctoral Scholarships for research performance at European level" project, financed by the European Social Found and Romanian Government (N. Cleju, C. Barabasa).

## REFERENCES

[1] D. Donoho, "Compressed sensing," IEEE Transactions on Information, Theory, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] S.S. Chen , D.L. Donoho , M.A. Saunders , "Atomic Decomposition by Basis Pursuit", SIAM Journal on Scientific Computing, Vol. 43, No. 1, 2005

[3] J. Haupt, R. Nowak, "Signal reconstruction from noisy random projections", IEEE Trans. on Information Theory, 52(9), pp. 4036-4048, September 2006)

[4] E. Candès, M. Wakin, "An introduction to compressive sampling", IEEE Signal Processing Magazine, 25(2), pp.21 - 30, March 2008)

[5] A. Ron, Z. Shen, "Affine systems in L2(Rd): the analysis of the analysis operator", J. Funct. Anal. 148 (1997) 408–447.

[6] J.-L. Starck, M. Elad, D.L. Donoho, "Redundant multiscale transforms and their application for morphological component analysis", Adv. Imag. Elect. Phys. 132 (2004).

[7] M. Fira, L. Goras, C. Barabasa, N. Cleju, "On ECG Compressed Sensing using Specific Overcomplete Dictionaries", Advances in Electrical and Computer Engineering, Vol. 10, Nr. 4, 2010, pp. 23- 28

[8] C. Monica Fira, L. Goras, C. Barabasa, N. Cleju, „ECG compressed sensing based on classification in compressed space and specified dictionaries", EUSIPCO 2011 (The 2011 European Signal Processing Conference), 29 august – 2 septembrie 2011, Barcelona, Spania

[9] M. Fira, L. Goras, "An ECG Signals Compression Method and Its Validation Using NNs", IEEE Transactions on Biomedical Engineering, Vol. 55, No. 4, 1319 – 1326, April 2008

[10] Yi-Haur Shiau, Chaur-Chin Chen, "A Sparse Representation Method with Maximum Probability of Partial Ranking for Face Recognition", International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 1, 2012

# An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree

Md. Hedayetul Islam Shovon
Department of Computer Science and Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh

Mahfuza Haque
Department of Computer Science and Engineering
Rajshahi University of Engineering & Technology
Rajshahi-6204, Bangladesh

***Abstract*—Improving student's academic performance is not an easy task for the academic community of higher learning. The academic performance of engineering and science students during their first year at university is a turning point in their educational path and usually encroaches on their General Point Average (GPA) in a decisive manner. The students evaluation factors like class quizzes mid and final exam assignment lab - work are studied. It is recommended that all these correlated information should be conveyed to the class teacher before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. In this paper, we present a hybrid procedure based on Decision Tree of Data mining method and Data Clustering that enables academicians to predict student's GPA and based on that instructor can take necessary step to improve student academic performance**

***Keywords- Database; Data clustering; Data mining; classification; prediction; Assessments; Decision tree; academic performance.***

## I. INTRODUCTION

Graded Point Average (GPA) is a commonly used indicator of academic performance. Many universities set a minimum GPA that should be maintained. Therefore, GPA still remains the most common factor used by the academic planners to evaluate progression in an academic environment. Many factors could act as barriers to student attaining and maintaining a high GPA that reflects their overall academic performance, during their tenure in university. These factors could be targeted by the faculty members in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance [1]. With the help of clustering algorithm and decision tree of data mining technique it is possible to discover the key characteristics for future prediction. Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets. The amount of data stored in educational databases is increasing rapidly. Clustering technique is most widely used technique for future prediction. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities (Kifaya, 2009). These applications can help both instructor and student to enhance the education quality. This study makes use of cluster analysis to segment students into groups according to their characteristics [2]. Decision tree analysis is a popular data mining technique that can be used to explain different variables like attendance ratio and grade ratio. Clustering is one of the basic techniques often used in analyzing data sets [3]. This study makes use of cluster analysis to segment students in to groups according to their characteristics and use decision tree for making meaningful decision for the student's.

## 2. Methodology

### A. *Data Clustering*

Data Clustering is unsupervised and statistical data analysis technique. It is used to classify the same data into a homogeneous group. It is used to operate on a large data-set to discover hidden pattern and relationship helps to make decision quickly and efficiently. In a word, Cluster analysis is used to segment a large set of data into subsets called clusters. Each cluster is a collection of data objects that are similar to one another are placed within the same cluster but are dissimilar to objects in other clusters.

#### *a. Implementation Of K-Means Clustering Algorithm*

K-Means is one of the simplest unsupervised learning algorithms used for clustering. K-means partitions "n" observations in to k clusters in which each observation belongs to the cluster with the nearest mean. This algorithm aims at minimizing an objective function, in this case a squared error function. The algorithm and flow-chart of K-means clustering is given below…

**Algorithm 1** Basic K-means Algorithm.

1: Select $K$ points as the initial centroids.
2: **repeat**
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Fig.1 Traditional K-Means Algorithm [4].

From the algorithm it is easily seen that, initially we have only the raw data. So, it is clustered around a single point. If the cluster number K is fixed then we need to cluster around that point. If the cluster is not fixed then it is continued until the centered is not changed. Initially the students are all in a same group. But when K-means clustering is applied on it then

it clusters the student's into three major categories, one is good, one is medium, and the other is low standard student.

The flow chart of the k-means algorithm that means how the k-means work out is given below.

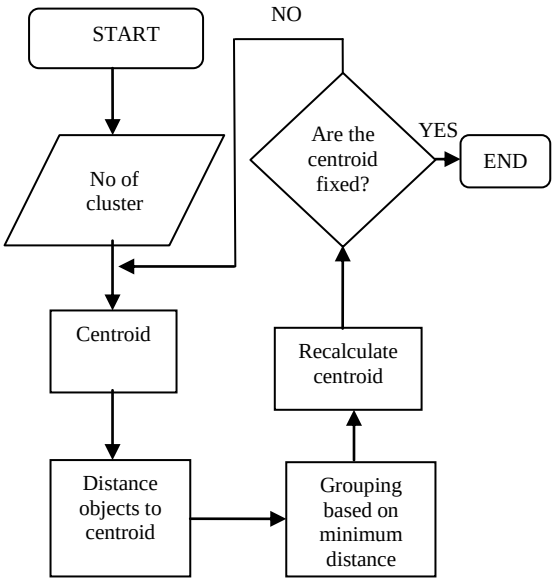



Fig.2 Flow-Chart Of K-Means Clustering.

B. Data mining

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The sequences of steps identified in extracting knowledge from data are shown in fig.3

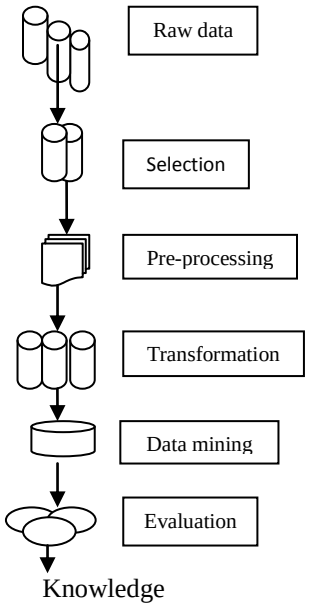



Fig.3 Steps Of Knowledge Extraction

*I. Decision Tree*

Decision tree induction can be integrated with data warehousing techniques for data mining. A decision tree is a predictive node ling technique used in classification, clustering, and prediction tasks. A decision tree is a tree where the root and each internal node are labeled with a question. The arcs emanating from each node represent each possible answer to the associated question. Each leaf node represents a prediction of a solution to the problem under consideration.

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. Decision Tree Algorithm: generate a decision tree from the given training data.

1 Create a node N
2 If samples are all of the same class, C then
3 Return N as a leaf node labeled with the class C;
4 If attribute-list is empty then
5 Return N as a leaf node labeled with the most common class in samples.
6 Select test-attribute, the attribute among attribute-list with the highest information gain;
7 Label node N with test-attribute;
8 For each known value $a_i$ of test-attribute.
9 Grow a branch from node N for the condition test attribute = $a_i$;
10 Let Si be the set of samples for which test-attribute = $a_i$;
11 If Si is empty then
12 Attach a leaf labeled with the most common class in samples;
13 Else attach the node returned by generate-decision-tree ($S_i$,attribute-list-attribute);

Each internal node tests an attribute, each branch corresponds to attribute value, and each leaf node assigns a classification.

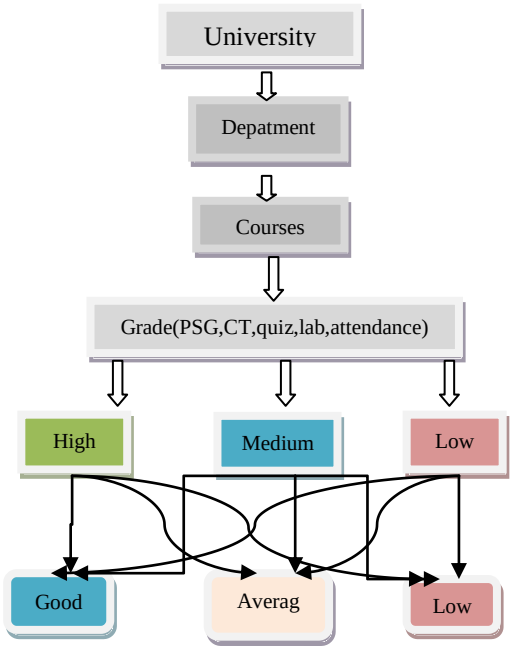



Fig.4 Decision Tree.

From 50 training sample here only 20 samples are shown.

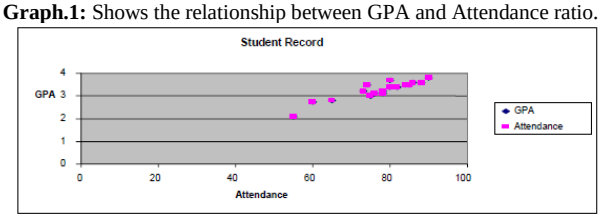
Table 1: Training Sample.

| Roll | GPA | CT | Attendance | Assignment | Lab_per | Quiz |
|---|---|---|---|---|---|---|
| 1 | 3.89 | 19 | 10 | Y | good | Y |
| 2 | 3.53 | 12 | 10 | Y | avg | Y |
| 3 | 3.2 | 10 | 10 | Y | good | Y |
| 4 | 3.6 | 16 | 10 | N | avg | N |
| 5 | 3.54 | 12 | 10 | Y | bad | Y |
| 6 | 3.5 | 10 | 10 | Y | good | N |
| 7 | 3 | 5 | 5 | N | avg | Y |
| 8 | 3.74 | 12 | 10 | Y | good | N |
| 9 | 3.67 | 14 | 10 | Y | good | Y |
| 10 | 2.05 | 2 | 6 | N | bad | N |
| 11 | 3.25 | 3 | 6 | N | avg | Y |
| 12 | 3.56 | 5 | 8 | N | good | Y |
| 13 | 3.2 | 9 | 5 | Y | avg | Y |
| 14 | 3.5 | 14 | 10 | Y | avg | Y |
| 15 | 3.2 | 10 | 10 | N | good | Y |
| 16 | 2.99 | 0 | 0 | Y | avg | N |
| 17 | 2.98 | 0 | 0 | N | avg | N |
| 18 | 3.87 | 3 | 5 | Y | avg | N |
| 19 | 3.45 | 8 | 8 | Y | avg | N |
| 20 | 3.21 | 9 | 6 | N | avg | N |

## II. RESULT AND DISCUSSION

From the training data GPA and the attendance ration of the student is given below

**Graph.1:** Shows the relationship between GPA and Attendance ratio.




If we apply K-means clustering algorithm on the training data then we can group the students in three classes "High" "Medium" and "Low" according to their new grade. New grade is calculated from the previous semester grade that means external assessment and internal assessment. The table and corresponding graph is given below.

Table 2. Percentage of students according to GPA.

| Class | GPA | No of student | Percentage |
|---|---|---|---|
| 1 | 2.00-2.20 | 5 | 8.33 |
| 2 | 2.20-3.00 | 10 | 16.67 |
| 3 | 3.00-3.32 | 17 | 28.33 |
| 4 | 3.32-3.56 | 15 | 25 |
| 5 | 3.56-4.00 | 13 | 21.67 |

Here, I cluster student among their GPA that means, from GPA 2.00- 2.20 we have 8.33% student. From 2.20-3.00 student percentage is 16.67%.

From 3.00-3.32 we have 28.33%. From 3.32-3.56 percentage is 25% .The percentage is 21.67% between GPA 3.56-4.00.

The graphical representation of GPA and the percentage of student's among the student are given below.

**Graph 2:** Number and percentage of students regarding to GPA

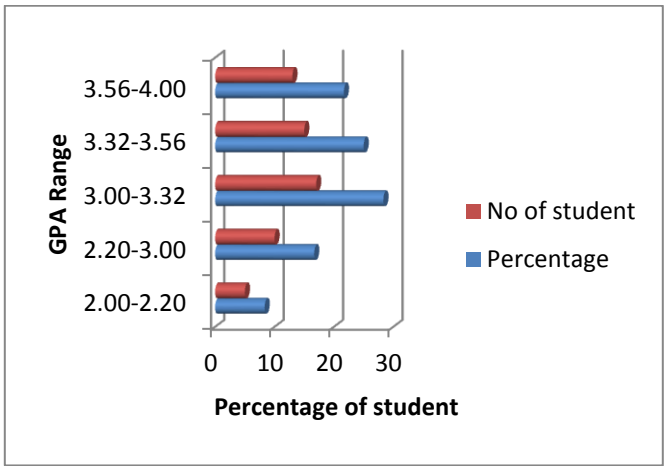


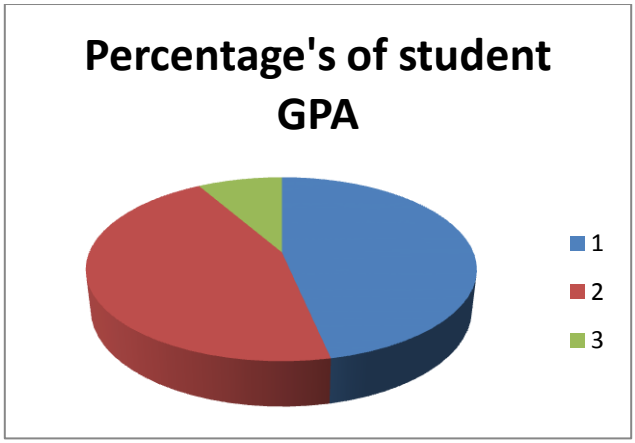

Table 2

| Class | GPA | No of student | Percentage |
|---|---|---|---|
| High | >=3.50 | 28 | 46.67 |
| Medium | 2.20<=GPA <3.5 | 27 | 45 |
| Low | <=2.20 | 5 | 8.33 |

After clustering the student, we group the student into three categories. One is High, second is Medium, and the last one is Low.

Graphical representation of these three categories is given below.

**Graph 3:** Shows the percentage of students getting high, medium and low GPA




If we apply data mining technique decision tree then it will help us to make correct decision about the student which is need to take by the instructor. The decision step is given below.

Table 3: Decision based on the student categories:

| Step No | Grade | Effort |
|---|---|---|
| S-01 | A+ | He/She is a good student. Need not to take special care. |
| S-02 | A,A- | Is not so good. Need to take care of CT & Quiz. |
| S-03 | B+,B | Is a medium student. Should take care of CT,quiz and lab performance also. |
| S-04 | Below B grade | Is a lower standard student. Need lot of practice of his/her lesson and also take care of all the courses ct,lab,quiz ,attendance carefully. |

## III. Conclusion and Future Work

In this study we make use of data mining process in student's database using k-means clustering algorithm and decision tree technique to predict student's learning activities. We hope that the information generated after the implementation of data mining and data clustering technique may be helpful for instructor as well as for students. This work may improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education. For future work, we hope to refine our technique in order to get more valuable and accurate outputs, useful for instructors to improve the students learning outcomes.

## References

[1] Oyelade, Oladipupo & Obagbuwa, "Application of K-means clustering algorithm for prediction of student's academic performance.", IJCSIS2010,vol.7,No.1, pp-292.

[2] Research, ISSN 1450-216X Vol.43 No.1 (Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar & M.Inayat Khan, "Data Mining Model for Higher Education System **",**European Journal of Scientific 2010), pp.24.

[3] Dr.Vuda Sreenivasaro & Capt.Genetu Yohannes ,"Improving academic performance of student of defence university based on data warehousing and data mining", Global Journal of computer science and technology, v.12,Issue2,Version.1,pp-29.

[4] Research, ISSN 1450-216X Vol.43 No.1 (Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar & M.Inayat Khan, "Data Mining Model for Higher Education System **",**European Journal of Scientific 2010), pp.27.

[5] Dr.Vuda Sreenivasaro & Capt.Genetu Yohannes ,"Improving academic performance of student of defence university based on data warehousing and data mining", Global Journal of computer science and technology, v.12,Issue2,Version.1,pp-33.

[6] Kifaya(2009) Mining student evaluation using associative classification and clustering.

[7] ZhaoHui. Maclennan.J, (2005). Data Mining with SQL Server 2005 Wihely Publishing, Inc.

## Authors Profile

**Md. Hedayetul Islam Shovon** was born in 1986 in Bangladesh. He received his B.Sc Engineering degree from Rajshahi University of Engineering and Technology (RUET) in 2009. He joined in the department of Computer Science and Engineering of RUET in October 2009. Since then he involved in different research oriented activities. His research interests include Image Processing, Data Clustering, Computer Vision and Optimization etc.

**Mahfuza Haque** was born in Bangladesh. He received his B.Sc Engineering degree from the department of Computer Science and Engineering of Rajshahi University of Engineering and Technology (RUET) in 2012. His research interest includes Pattern recognition, Data clustering, Data mining.

# Prevention and Detection of Financial Statement Fraud – An Implementation of Data Mining Framework

Rajan Gupta
Research Scholar, Dept. of Computer
Sc. & Applications, MaharshiDayanand University,
Rohtak (Haryana) – India.

Nasib Singh Gill
Professor,Dept. of Computer
Sc. & Applications, MaharshiDayanand University,
Rohtak (Haryana), India.

***Abstract*— Every day, news of financial statement fraud is adversely affecting the economy worldwide. Considering the influence of the loss incurred due to fraud, effective measures and methods should be employed for prevention and detection of financial statement fraud. Data mining methods could possibly assist auditors in prevention and detection of fraud because data mining can use past cases of fraud to build models to identify and detect the risk of fraud and can design new techniques for preventing fraudulent financial reporting. In this study we implement a data mining methodology for preventing fraudulent financial reporting at the first place and for detection if fraud has been perpetrated. The association rules generated in this study are going to be of great importance for both researchers and practitioners in preventing fraudulent financial reporting. Decision rules produced in this research complements the prevention mechanism by detecting financial statement fraud.**

***Keywords- Data mining framework; Rule engine; Rule monitor***.

## I. INTRODUCTION

Financial statement fraud is a deliberate misstatement of material facts by the management in the books of accounts of a company with the aim of deceiving investors and creditors. This illegitimate task performed by management has a severe impact on the economy throughout the world because it significantly dampens the confidence of investors.

The magnitude of this problem can be evaluated by the fact that a number of Chinese companies listed on US stock exchanges have of faced accusations accounting fraud, and in June 2011, the U.S. Securities and Exchange Commission warned investors against investing with Chinese firms listing via reverse mergers. While over 20 US listed Chinese companies have been de-listed or halted in 2011, a number of others have been hit by the resignation of their auditors [1].

Association of certified fraud examiners (ACFE) in its report to the nation on occupational fraud and abuse (2012) [2] suggests that the typical organization loses 5% of its revenue to fraud each year. The median loss caused by occupational fraud cases was $140,000.

This study by ACFE reveals that perpetrators with higher levels of authority tend to cause much larger losses. The median loss among frauds committed by owner / executives was $573,000, the median loss caused by managers was $180,000 and the median loss caused by employees was $60,000. The report by the ACFE also measured the common methods of detecting fraud and found that in more than 43 % cases tips and complaints have been the most effective means of detecting frauds.

Prevention and detection of financial statement fraud has become a major concern for almost all organisations globally. Though, it is a fact that prevention of financial statement fraud is the best way to reduce it, but detection of fraudulent financial reporting is critical in case of failure of prevention mechanism.

The aim of this paper is to provide a methodology for prevention and detection of financial statement fraud and to present the empirical results by implementing the framework. In this research, we test the applicability of data mining framework for prevention and detection of financial statement fraud. As per the recommendations of the framework we apply descriptive data mining for prevention and predictive data mining techniques for detection of financial statement fraud.

This paper is organized as follows. Section 2 summarizes the contribution in the field of prevention and detection of financial statement fraud. Section 3 implements the data mining framework for detection of fraud if prevention techniques have failed followed by conclusion (Section 4).

## II. RELATED WORK

Cost of financial statement fraud is very high both in terms of finance as well as the goodwill of the organization and related country. In order to curb the chances of fraud and to detect the fraudulent financial reporting, number of researchers had used various techniques from the field of statics, artificial intelligence and data mining.

For instance, Spathis et al [3] compared multi-criteria decision aids with statistical techniques such as logit and discriminant analysis in detecting fraudulent financial statements. Neural Network based support systems was proposed by Koskivaara [4] in 2004. He demonstrated neural network as a possible tool for use in auditing and found that the main application areas of NN were detection of material errors, and management fraud.

A decision tree was constructed by Koh and Low [5] in order to predict the hidden problems in financial statements by examining the following six variables: quick assets to current liabilities, market value of equity to total assets, total liabilities to total assets, interest payments to earnings before interest and tax, net income to total assets, and retained earnings to total assets. Kirkos et al [6], carry out an in-depth analysis of publicly available data of 76 Greek manufacturing firms for detecting fraudulent financial statements by using three Data Mining classification methods namely Decision Trees, Neural Networks and Bayesian Belief Networks. They investigated the usefulness of these techniques in identification of FFS.

In 2007, a genetic algorithm approach to detecting financial statement fraud was presented by Hoogs et al [7]. An innovative fraud detection mechanism is developed by Huang et al. [8] on the basis of Zipf's Law. This technique reduces the burden of auditors in reviewing the overwhelming volumes of datasets and assists them in identification of any potential fraud records. A novel financial kernel using support vector machines for detection of management fraud was developed by Cecchini et al [9].

In 2008, the effectiveness of CART on identification and detection of financial statement fraud was examined by Belinna et al [10] and found CART as a very effective technique in distinguishing fraudulent financial statement from non-fraudulent. Juszczak et al. [11] apply many different classification techniques in a supervised two-class setting and a semi-supervised one-class setting in order to compare the performances of these techniques and settings.

Further, Zhou & Kapoor [12] in 2011 applied four data mining techniques namely regression, decision trees, neural network and Bayesian networks in order to examine the effectiveness and limitations of these techniques in detection of financial statement fraud. They explore a self – adaptive framework based on a response surface model with domain knowledge to detect financial statement fraud.

Ravisankar et al [13] applied six data mining techniques namely Multilayer Feed Forward Neural Network (MLFF), Support Vector Machines (SVM), Genetic Programming (GP), Group Method of Data Handling (GMDH), Logistic Regression (LR), and Probabilistic Neural Network (PNN) to identify companies that resort to financial statement fraud on a data set obtained from 202 Chinese companies. They found Probabilistic neural network as the best techniques without feature selection. Genetic Programming and PNN outperformed others with feature selection and with marginally equal accuracies.

Recently, Johan Perols [14] compares the performance of six popular statistical and machine learning models in detecting financial statement fraud. The results show, somewhat surprisingly, that logistic regression and support vector machines perform well relative to an artificial neural network in detection and identification of financial statement fraud.

The review of the existing literature reveals that the research conducted till date is solely in the field of detection and identification of financial statement fraud and a very little or no work has been done in the field of prevention of fraudulent financial reporting.

Therefore, in the present research we implement a data mining framework for prevention along with detection of financial statement fraud.

The major objective of this research is to test the applicability of predictive and descriptive data mining techniques for detection and prevention of fraud respectively by implementing a data mining framework. In order to feel the sense of fraud, we implement association rule mining and to detect fraudulent financial reporting we apply three classification techniques namely decision trees, naïve Bayesian classifier and Genetic programming.

## III. The Methodology: Applicability & Its Implementation

The methodology applied in this paper is a data mining framework of Gupta & Gill (2012) [15]. The framework is presented as Fig 1.

The first step of the framework is **feature selection**. We selected 62 financial ratios / variables as features to be used as input vector in further analysis.

These features represent behavioural characteristics along with measures of liquidity, safety, profitability and efficiency of the organisations under consideration. Table 1 present the list of 62 features.

Figure 1: A data mining framework for prevention and detection of financial statement fraud.

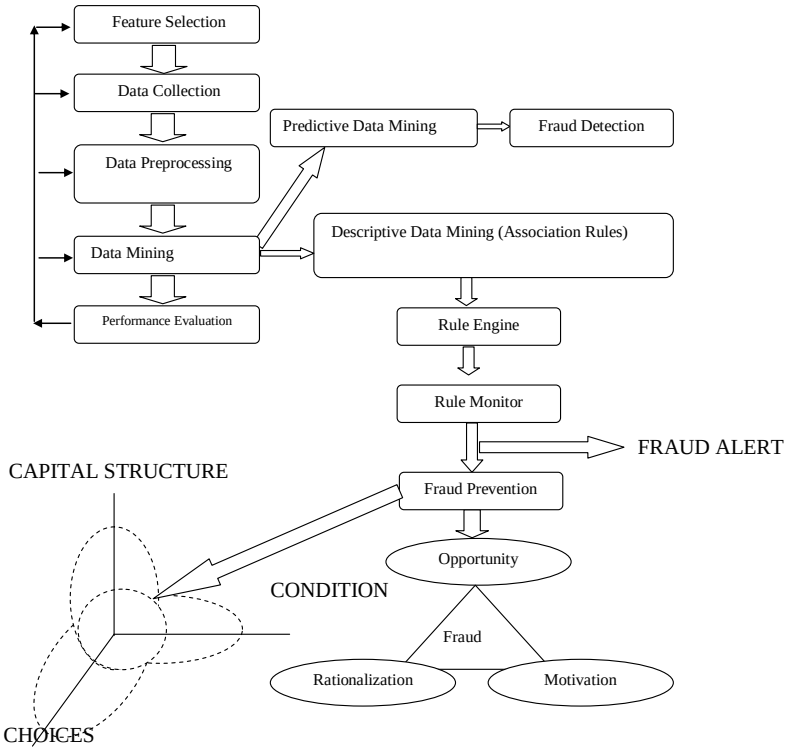


Table 1: Features For Prevention & Detection Of Financial Statement Fraud

| S.No. | Financial Items / Ratios |
|---|---|
| 1 | Debt |
| 2. | Total assets |
| 3 | Gross profit |
| 4 | Net profit |
| 5 | Primary business income |
| 6 | Cash and deposits |
| 7 | Accounts receivable |
| 8 | Inventory/Primary business income |
| 9 | Inventory/Total assets |
| 10 | Gross profit/Total assets |
| 11 | Net profit/Total assets |
| 12 | Current assets/Total assets |
| 13 | Net profit/Primary business income |
| 14 | Accounts receivable/Primary business income |
| 15 | Primary business income/Total assets |
| 16 | Current assets/Current liabilities |
| 17 | Primary business income/Fixed assets |
| 18 | Cash/Total assets |
| 19 | Inventory/Current liabilities |
| 20 | Total debt/Total equity |
| 21 | Long term debt/Total assets |
| 22 | Net profit/Gross profit |
| 23 | Total debt/Total assets |
| 24 | Total assets/Capital and reserves |
| 25 | Long term debt/Total capital and reserves |
| 26 | Fixed assets/Total assets |
| 27 | Deposits and cash/Current assets |
| 28 | Capitals and reserves/Total debt |
| 29 | Accounts receivable/Total assets |
| 30 | Gross profit/Primary business profit |
| 31 | Undistributed profit/Net profit |
| 32 | Primary business profit/Primary business profit of last year |
| 33 | Primary business income/Last year's primary business income |
| 34 | Account receivable /Accounts receivable of last year |
| 35 | Total assets/Total assets of last year |
| 36 | Debit / Equity |
| 37 | Accounts Receivable / Sales |
| 38 | Inventory / Sales |
| 39 | Sales – Gross Margin |
| 40 | Working Capital / Total Assets |
| 41 | Net Profit / Sales |
| 42 | Sales / Total Assets |
| 43 | Net income / Fixed Assets |
| 44 | Quick assets / Current Liabilities |
| 45 | Revenue /Total Assets |
| 46 | Current Liabilities / Revenue |
| 47 | Total Liability / Revenue |
| 48 | Sales Growth Ratio |
| 49 | EBIT |
| 50 | Z – Score |
| 51 | Retained Earnings / Total Assets |
| 52 | EBIT / Total Assets |
| 53 | Total Liabilities / Total assets |
| 54 | Cash return on assets |
| 55 | Interest expense / Total Liabilities |
| 56 | EBIT / sales |
| 57 | Age of the company (Number of years since first filing available from provider) |
| 58 | Change in cash scaled to total assets |
| 59 | Change in current assets scaled by current liabilities |
| 60 | Change in total liabilities scaled by total assets |
| 61 | Size of company on the basis of assets |
| 62 | Size of company on the basis of revenue |

During the second step of **Data Collection,** all the financial ratios of Table 1 have been collected from financial statements namely balance sheet, income statement and cash flow statement for 114 companies listed in different stock exchanges globally. The dataset used in this study has been collected from www.**wikinvest**.com. The companies accused of fraudulent financial reporting has been identified by analysing Accounting and Auditing Enforcement Releases published by S.E.C. (U.S. Securities and Exchange Commission) for the period of five years starting from 2007. All the incidents of violation of the Foreign Corrupt Practices Act (FCPA) have been removed from the sample, because FCPA prohibits the practice of bribing foreign officials and most of the AAERs issued because of FCPA do not reflect which financial statement viz. balance sheet or income statement, is affected.

We identified 29 organisations with charges of issuing fraudulent financial statements and hence termed as fraudulent in this study. 85 organisations out of total of 114 have been marked as non – fraudulent since no indication or proof of falsifying financial statement has been reported. However, absence of any proof does not guarantee that these firms have not falsified their financial statements or will not do the same in future.

In order to make dataset ready for mining, data need to be pre - processed. Data has been transformed in to an appropriate format for mining during the step of **Data preprocessing**. Dataset is cleaned further by replacing missing values with the mean of the variable. Each of the independent financial variables has been normalized by using range transformation (min = 0.0, max = 1.0).

We compiled all the 62 input variables given in Table 1.In order to reduce dimensionality of the dataset we applied one way ANOVA. The variables with p – value <=0.05 are considered significant and informative and with high p – value are deemed to be non – informative. Informative variables are tested further using descriptive data mining methods. The input variables which are considered significant are given in Table 2 along with respective F- values and p – values.

TABLE 2: LIST OF INFORMATIVE VARIABLES

| S.No. | Financial Items / Ratios | F - value | P – value |
|---|---|---|---|
| 1 | Debt | 1.345 | .028 |
| 2 | Inventory/Primary business income | 3.031 | .001 |
| 3 | Inventory/Total assets | 17.468 | .000 |
| 4 | Net profit/Total assets | 3.035 | .001 |
| 5 | Accounts receivable/Primary business income | 6.099 | .018 |
| 6 | Primary business income/Total assets | 3.038 | .001 |
| 7 | Primary business income/Fixed assets | 3.055 | .001 |
| 8 | Cash/Total assets | 2.918 | .001 |
| 9 | Inventory/Current liabilities | 6.744 | .001 |
| 10 | Total debt/Total assets | 2.851 | .001 |
| 11 | Long term debt/Total capital and reserves | 4.266 | .014 |
| 12 | Deposits and cash/Current assets | 2.932 | .001 |

| | | | |
|---|---|---|---|
| 13 | Capitals and reserves/Total debt | 2.213 | .003 |
| 14 | Gross profit/Primary business profit | 3.847 | .008 |
| 15 | Accounts Receivable / Sales | 1.702 | .021 |
| 16 | Working Capital / Total Assets | 2.906 | .001 |
| 17 | Sales / Total Assets | 12.818 | .003 |
| 18 | Net income / Fixed Assets | 3.038 | .001 |
| 19 | Quick assets / Current Liabilities | 1.839 | .050 |
| 20 | Revenue /Total Assets | 12.818 | .003 |
| 21 | Capital and Reserves / Total Debt | 1.130 | .049 |
| 22 | Retained Earnings / total assets | 3.039 | .001 |
| 23 | EBIT | 4.363 | .023 |
| 24 | EBIT / Total Assets | 3.043 | .001 |
| 25 | Z – score | 3.054 | .001 |
| 26 | Total liabilities / Total Assets | 3.154 | .002 |
| 27 | Cash flow from operations | 1.720 | .018 |
| 28 | Cash return on assets | 3.940 | .002 |
| 29 | Interest Expenses | 1.806 | .010 |
| 30 | Interest exp / Total Liabilities | 1.440 | .042 |
| 31 | Size of the company on the basis of assets | 1.179 | .043 |
| 32 | Change in cash scaled by total assets | 2.967 | .001 |
| 33 | Current Liabilities of the previous year | 1.391 | .028 |
| 34 | Total Liabilities of the previous year | 1.346 | .022 |
| 35 | Change in Total Liabilities scaled by Total Assets | 3.188 | .001 |

The step of data preprocessing is followed by selection of an appropriate data mining technique. The framework suggests the use of descriptive data mining technique for prevention and predictive methods for detection of financial statement fraud. Therefore, we first apply association rule mining for preventing fraudulent financial reporting at the first place.

We implement association rules by using RapidMiner version 5.2.3. All the informative variables have been converted into nominal variables. Nominal variables further converted into binomial variables because it is the preliminary requirement for rule engine. In the next stage of the framework, **Rule engine** generates the required association rules.

In the process of rule generation, frequent itemsets is being generated using FP Growth. The minimum support for FP Growth has been set to 0.95. The frequent itemsets generated has been used for creating the association rules. The minimum confidence for generating rules is 0.8. Table 3 lists the association rules generated by rule engine.

Now, the **rule monitor** module will monitor the financial ratios of each organisation and compare the values of the ratios with the values given in the association rules for indicating the anomaly. Anomalies detected by rule monitor are reflected as number of non fraud companies identified as fraud in Table 3. The results generated by rule monitor are able to raise an alarm regarding fraud.

In view of the whistle blown by rule monitor, organisations should consider the presence or absence of conditions which refers to certain financial pressures exhibited by the management. Such organisations should think in terms of providing employees the working environment that values honesty because irresponsible and ineffective corporate governance could increase the chances of financial statement fraud. The absence of effective corporate governance may provide enough opportunity to the managers / employees for selecting an option of fraudulent financial reporting. Hence, this unlawful practice of fraudulent financial reporting could be prevented by checking or taking away the opportunity to commit fraud and by avoiding the combination of opportunity, pressure and motive in an organisation.

TABLE 3: ASSOCIATION RULE

| S.N | Association Rule | Support | Confidence | Lift | Conviction | Number of non-fraud companies identified as fraud |
|---|---|---|---|---|---|---|
| 1. | Inventory / total assets > 0.033 ➔ Fraud | 43% | 86% | 1.153 | 1.812 | 30 |
| 2. | Cash / Total Assets <0.198 ➔Fraud | 42.1% | 84.2% | 1.129 | 1.611 | 36 |
| 3. | Inventory / Current Liabilities >0.190 ➔Fraud | 43.9% | 87.7% | 1.176 | 2.071 | 31 |
| 4. | Deposits and cash / Current Assets <0.408 ➔Fraud | 43.9% | 87.7% | 1.176 | 2.071 | 33 |
| 5. | Sales / Total Assets >.553 ➔ Fraud | 44.7% | 89.5% | 1.200 | 2.417 | 23 |
| 6. | Revenue / Total Assets >.553 ➔Fraud | 44.7% | 89.5% | 1.200 | 2.417 | 25 |
| 7. | Inventory / current Liabilities >0.190 && Deposits and cash / Current Assets < 0.408 ➔ Fraud | 43.9% | 87.7% | 1.176 | 2.071 | 15 |

Once the prevention mechanism has failed to prevent fraud then the framework suggest the usage of predictive data mining for detection and identification of financial statement fraud. In this study three data mining techniques namely CART, Naive Bayesian Classifier and Genetic Programming have been used for detection of fraudulent financial statements and differentiating between fraud and non fraud reporting.In order to have better reliability of the result, ten – fold cross validation has been implemented.

A decision tree (CART) has been constructed in this study by using SIPINA Research edition software version – 32 bit. The complete dataset has been used as training data for constructing the tree given as Figure 2. The confidence level was set to 0.05. CART manages to classify 95 % cases. This method well classifies 98 % non fraud cases and misclassifies only 4 fraud cases. The percentage of classification for fraud cases is 86 %.

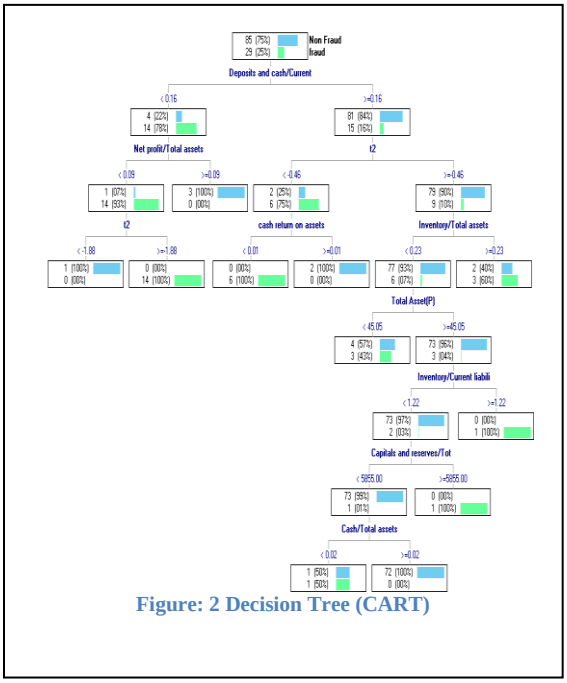



Figure: 2 Decision Tree (CART)

The financial ratio namely Deposits and cash to current assets has been used as the first splitter by the decision tree constructed in this research. This ratio is an indicator for the measurement of capability of a company in converting its non – liquid assets into cash. At second level of the tree, retained earnings / total assets (t2) and net profit / total assets has been used as a splitter. The ratios used by tree are given in Table 4.

TABLE 4

| S. No. | Financial Ratios / Items |
|---|---|
| 1 | Net profit/Total assets |
| 2 | Size of the company on the basis of assets (Total Assets P) |
| 3 | Deposits and cash/Current assets |
| 4 | Capital & Reserves / Total Assets |
| 5 | Inventory / Current Liabilities |
| 6. | Cash return on Assets |
| 7. | Cash / Total Assets |
| 8. | Inventory / Total Assets |
| 9. | Retained earnings / Total Assets (t2) |

We applied Naïve Bayesian Classifier, the second method of classification by using SIPINA Research edition software version – 32 bit. The method correctly classifies 88% cases.

Third method of classification, Genetic programming has been implemented using a data mining tool Discipulus version 5.1. The process begins with division of dataset in to two datasets namely training data and validation data. The training data set has been used to train the sample and validation dataset is used exclusively for the purpose of validation. In this study, 80% of the whole dataset is designated as training data for training the sample, whereas, rest 20% is assigned exclusively for the purpose of validation. Since our dependent variable (target output) is binary, we select "hits then fitness" as a fitness function. Every single run of Discipulus has been set to terminate after it has gone 50 generations with no improvement in fitness.

**Performance evaluation**, the final step of the framework is used for measuring the performance and judging the efficacy of data mining methods. Performance of association rules generated in this study has been measured with the help of support, confidence, lift and conviction (Table 3). The rules generated by rule engine have support of more than 40% and confidence more than 80%.

Sensitivity and specificity have been used as a metrics for performance evaluation of classification techniques used in this research. The confusion matrix for Decision trees, Naïve Bayesian classifier and Genetic programming is given below.

TABLE: 5 (CONFUSION MATRIX FOR DECISION TREE)

| Label | Non Fraud | Fraud |
|---|---|---|
| NF (Non Fraud) | 83 | 2 |
| F (Fraud) | 4 | 25 |

TABLE: 6 (CONFUSION MATRIX FOR NAIVE BAYSIAN CLASSIFIER)

| Label | Non Fraud | Fraud |
|---|---|---|
| NF | 79 | 6 |
| F | 8 | 21 |

TABLE: 7 (CONFUSION MATRIX FOR GENETIC PROGRAMMING)

| Label | Non Fraud | Fraud |
|---|---|---|
| NF | 84 | 1 |
| F | 13 | 16 |

Performance matrix indicating the sensitivity (type 1 error) and specificity (type II error) of the three methods used in this study is given in Table 8.

Table: 8 (Performance Matrix)

| S.No. | Predictor | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| 1 | Decision Tree | 86.2 | 97.7% |
| 2 | Naïve Bayesian Classifier | 84 | 92.9 |
| 3 | Genetic Programming | 53 | 99.2 |

Decision tree (CART) classifies 25 fraud cases as fraud from a total of 29 such cases correctly therefore, produces best sensitivity. The following are the decision rules generated by using decision tree (Figure 2).

1. If ((Deposits and cash / Current assets >=0.16) && (Retained Earnings / Total Assets> = -0.46) && (Inventory / Total Assets > = 0.23)) then Fraud
2. If ((Deposits and cash / Current assets >=0.16) && (Retained Earnings / Total Assets> = -0.46) && (Inventory / Total Assets < 0.23) && (Size of the company on the basis of assets >=45.05) &&(Inventory / Current Liabilities >=1.22)) then Fraud
3. If ((Deposits and cash / Current assets >=0.16) && (Retained Earnings / Total Assets> = -0.46) && (Inventory / Total Assets < 0.23) && (Size of the company on the basis of assets >=45.05) && (Inventory / Current Liabilities >=1.22) && (Capital and Reserves / Total Assets > = 5855.00)) then Fraud
4. If ((Deposits and cash / Current assets < 0.16) && (Retained Earnings / Total Assets> = -1.88) && (Net profit / Total Assets < 0.09)) then Fraud
5. If ((Deposits and cash / Current assets >=0.16) && (Retained Earnings / Total Assets< -0.46) && (Cash return on Assets < 0.01)) then Fraud

Genetic programming outperforms the other two techniques by correctly classifying 84 cases out of 85 non fraud organisations, hence produces best specificity. Table 9 represents the input impact of various input parameters on the model generated by Genetic Programming.

TABLE: 9 IMPACT OF INPUT VARIBALES (GENETIC PROGRAMMING)

| S.No. | Variable | Frequency | Average Impact | Maximum Impact |
|---|---|---|---|---|
| 1 | Debt | 0.06 | 00.00000 | 00.00000 |
| 2 | Inventory/Primary business income | 0.35 | 22.52747 | 53.84615 |
| 3 | Inventory/Total assets | 0.35 | 09.70696 | 20.87912 |
| 4 | Net profit/Total assets | 0.06 | 02.19780 | 02.19780 |
| 5 | Cash/Total assets | 0.29 | 03.84615 | 05.49451 |
| 6 | Total debt/Total assets | 0.12 | 00.00000 | 00.00000 |
| 7 | Fixed assets/Total assets | 0.00 | 00.00000 | 00.00000 |
| 8 | Deposits and cash/Current assets | 0.18 | 06.59341 | 06.59341 |
| 9 | Working Capital / Total Assets | 0.06 | 00.00000 | 00.00000 |
| 10 | Sales / Total Assets | 0.00 | 00.00000 | 00.00000 |
| 11 | Net income / Fixed Assets | 0.41 | 07.69231 | 09.89011 |
| 12 | Revenue /Total Assets | 0.29 | 09.01099 | 14.28571 |
| 13 | EBIT | 0.06 | 05.49451 | 05.49451 |
| 14 | Z score | 0.06 | 19.78022 | 19.78022 |
| 15 | Accounts receivable/Primary business income | 0.29 | 00.54945 | 01.09890 |
| 16 | Primary business income/Total assets | 0.18 | 02.74725 | 03.29670 |
| 17 | Primary business income/Fixed assets | 0.41 | 03.29670 | 08.79121 |
| 18 | Capitals and reserves/Total debt | 0.00 | 00.00000 | 00.00000 |
| 19 | Gross profit/Primary business profit | 0.53 | 05.65149 | 09.89011 |
| 20 | Accounts Receivable / Sales | 0.00 | 00.00000 | 00.00000 |
| 21 | Retained earnings / Total Assets | 0.18 | 02.93040 | 04.39560 |
| 22 | EBIT / Total Assets | 0.24 | 03.29670 | 05.49451 |

Since Decision trees are capable of identifying type 1 error in more than 86% and Genetic programming correctly detect type II error for almost all the cases present in the dataset, therefore, we arrive at a conclusion that data mining techniques used in this study are capable enough for identification and detection of financial statement fraud in case of failure of prevention mechanism.

## IV. CONCLUSION

Prevention along with detection of financial statement fraud would be of great value to the organizations throughout the world. Considering the need of such a mechanism, we employ a data mining framework for prevention and detection of financial statement fraud in this study. The framework used in this research follow the conventional flow of data mining.

We identified and collected 62 features from financial statements of 114 organizations. Then we find 35 informative variables by using one way ANOVA.

These informative variables are being used for implementing association rule mining for prevention and three predictive mining techniques namely Decision Tree, Naïve Bayesian Classifier, Genetic programming for detection of financial statement fraud. Rule Engine module of the framework generated 7 association rules. These rules are used by rule monitor module for raising an alarm regarding fraud and hence preventing it at the first place.

The three data mining methods used for detection of financial statement fraud are compared on the basis of two important evaluation criteria namely sensitivity and specificity. Decision tree produces best sensitivity and Genetic programming best specificity as compared with other two methods. These techniques will detect the fraud in case of failure of prevention mechanism. Hence, the framework used in this research is able to prevent fraudulent financial reporting and detect it if management of the organization is capable of perpetrating financial statement fraud despite the presence of anti fraud environment.

## REFERENCES

[1] Atkins Matt, Accounting Fraud in US listed Chinese companies (September 2011). Available at: http://www.financierworldwide.com

[2] ACFE, 2012 ACFE Report to the nations on ocupational fraud and abuse, *Technical report- Global fraud survey 2012*, 2012.

[3] C. Spathis, M. Doumpos, C. Zopounidis, Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques, European Accounting Review 11 (3) (2002) 509–535.

[4] E. Koskivaara, Artificial neural networks in auditing: state of the art, The ICFAI Journal of Audit Practice 1 (4) (2004) 12–33.

[5] H.C. Koh, C.K. Low, Going concern prediction using data mining techniques, Managerial Auditing Journal 19 (3) (2004) 462–476.

[6] Efstathios Kirkos, Charalambos Spathis &Yannis Manolopoulos (2007), Data mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications 32 (23) (2007) 995–1003

[7] Hoogs Bethany, Thomas Kiehl, Christina Lacomb and DenizSenturk (2007). A Genetic Algorithm Approach to Detecting Temporal Patterns Indicative Of Financial Statement Fraud, Intelligent systems in accounting finance and management 2007; 15: 41 – 56, John Wiley & Sons, USA, available at: www.interscience.wiley.com

[8] S.-M. Huang, D.C. Yen, L.-W.Yang, J.-S.Hua, An investigation of Zipf's Law for fraud detection.Decision Support Systems 46 (1) (2008) 70–83.

[9] M. Cecchini, H. Aytug, G.J. Koehler, and P. Pathak. Detecting Management Fraud in Public Companies.http://warrington.ufl.edu/isom/docs/papers/DetectingManagementFraudInPublicCompanies.pdf

[10] BelinnaBai, Jerome yen, Xiaoguang Yang, False Financial Statements: Characteristics of china listed companies and CART Detection Approach, International Journal of Information Technology and Decision Making , Vol. 7, No. 2(2008), 339 – 359

[11] Juszczak, P., Adams, N.M., Hand, D.J., Whitrow, C., & Weston, D.J. (2008). Off-the-peg and bespoke classifiers for fraud detectionl, Computational Statistics and Data Analysis, vol. 52 (9): 4521-4532

[12] Wei Zhou, G. Kappor, Detecting evolutionary financial statement fraud. Decision Support Systems 50 (2011) 570 – 575.

[13] P.Ravisankar, V. Ravi, G.RaghavaRao, I., Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems, 50(2011) 491 – 500

[14] Johan Perols, Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms, A Journal of Practice & Theory 30 (2), 19 (2011), pp. 19-50

[15] Gupta Rajan& Gill Nasib S. " A data mining framework for prevention and detection of financial statement fraud", International Journal of Computer Application, 50(8): 7 - 14, July 2012. Published by Foundation of computer science, New York, U.S.A.

## AUTHORS PROFILE

Rajan Gupta obtained master's degree in computer application from Department of Computer Science & Application, Guru Jambheshwar University, Hisar, Haryana, India and Master Degree of Philosophy in Computer Science from Madurai Kamraj University, Madurai, India. He is currently pursuing Doctorate degree in Computer Science from Department of Computer Science & Application, Mahrshi Dayanand University, Rohtak, Haryana, India.

Dr Nasib S. Gill obtained Doctorate degree in computer science and Post-doctoral research in Computer Science from Brunel Univerrsity, U.K. He is currently working as Professor and Head in the Department of Computer Science and Application, Mahrshi Dayanand University, Rohtak, Haryana, India. He is having more than 22 years of teaching and 20 years of research experience. His interest areas include software metrics, component based metrics, testing, reusability, Data Mining and Data warehousing, NLP, AOSD, Information and Network Security.

# Review of Remote Terminal Unit (RTU) and Gateways for Digital Oilfield delpoyments

Francis Enejo Idachaba
Department of Electrical and Information Engineering
Covenant University Ota.
Ogun state
Nigeria

Ayobami Ogunrinde
SPDC Nigeria

***Abstract*— The increasing decline in easy oil has led to an increasing need for the optimization of oil and gas processes. Digital oilfields utilize remote operations to achieve these optimization goals and the remote telemetry unit and gateways are very critical in the realization of this objective. This paper presents a review of the RTUs and gateways utilized in digital oilfield architectures. It presents a review of the architecture, their functionality and selection criteria. It also provides a comparison of the specifications of some popular RTUs.**

***Keywords—Digital Oilfield; Gateway; HMI; i-fields; RTU; Smartfields.***

## I. INTRODUCTION

The advent of Digital Oilfields, Smartfields or i-fields has led to an increase in the need to monitor, control and automate various systems at remote oil and gas production sites to increase production, reduce overall production cost, and reduce employee exposure. Control systems such as SCADA (Supervisory Control and Data Acquisition), or DCS (Distributed Control System) using Remote Telemetry Units and Gateways are deployed to achieve these control functions. The RTUs and gateways comprises of various components such as

- HMI (Human Machine Interface)
- RTU (Remote Terminal Unit): This collects the site data and sends it to a station via a communications system.
- Supervisory systems/ Master station: this collects the information from the process and control the process. This is usually a computer.
- Communication system that provides a means by which all components communicate securely without loss of data and information.

Digital oilfield installations require bidirectional transmission of data from the sensors located in the field and control signals from the control room or the office domain to these sensors and devices located in the field. The data from the sensors are transmitted at defined intervals or by exception while the control algorithms used for the field devices can either be the on/off control or a variable control. This paper Ease of Use

## II. RTU

Remote Terminal Unit (RTU) is a microprocessor-based device connected to sensors, transmitters or process equipment for the purpose of remote telemetry and control.

RTUs find applications in oil and gas remote instrumentation monitoring, networks of remote pump stations, Environmental monitoring systems, Air traffic equipment etc. [1]

RTUs with the aid of appropriate sensors, monitors production processes at remote site and transmits all data to a central station where it is collated and monitored. An RTU can be interfaced using serial ports (RS232, RS482, and RS422) or Ethernet to communicate with the central stations. They also support various protocol standards such as Modbus, IEC 60870, DNP3 making it possible to interface with 3rd party software.

## III. RTU ARCHITECTURE

The RTU architecture comprises of a CPU, volatile memory and nonvolatile memory for processing and storing programs and data. It communicates with other devices via either serial ports or an onboard modem with I/O interfaces. It has a power supply with a backup battery, surge protection against spikes, real-time clock and a watchdog timer to ensure that it restarts when operating in the sleep mode.[2]

Figure 1 shows the block diagram of a typical RTU configuration. A typical RTU hardware module includes a control processor and associated memory, analog inputs, analog outputs, counter inputs, digital inputs, digital outputs, communication interfaces and power supply [3]

### *A. Central Processing Unit (Cpu)*

Current RTU designs utilize a 16bit or 32 bits microprocessor with a total memory capacity of 256kbytes expandable to 4 Mbytes. It also has two or three communication ports (RS232, RS422 and RS485) or Ethernet link. This system is controlled by a firmware and a real-time clock with full calendar is used for accurate time stamping of events. A watchdog timer provides a check that the RTU program is executing regularly. The RTU program regularly resets the watchdog timer and if this is not done within a

certain time-out period the watchdog timer flags an error condition and can sometimes reset the CPU. [3]

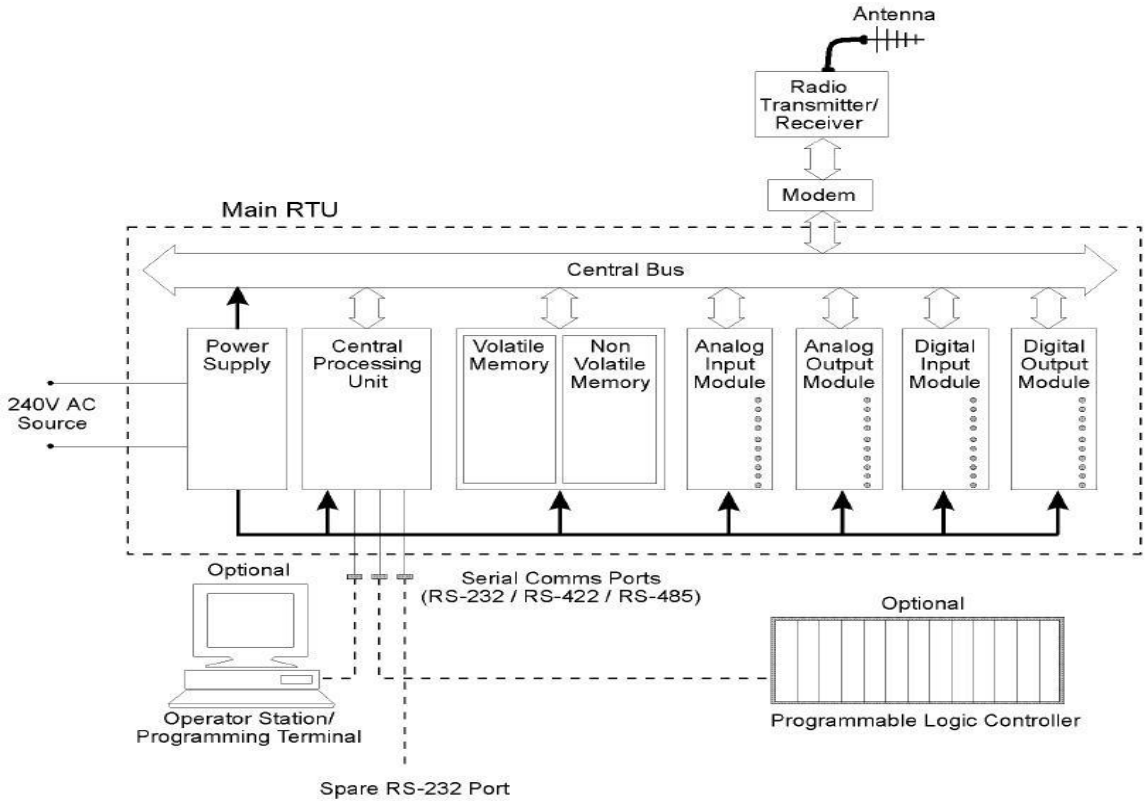



Figure. 1 RTU Hardware Structure [3]

*B. Analog Input Modules*

An analog input signal is generally a voltage or current that varies over a defined value range, in direct proportion to a physical process measurement. 4-20 milliamp signals are most commonly used to represent physical measurements like pressure, flow and temperature. [4]. Five main components that makes up the analog input module are as follows:

- Input multiplexer: This samples several analog inputs in turn and switches each to the output in sequence. The output goes to the analog digital converter.
- Input signal amplifier: This amplifies the low-level voltages to match the input range of the board's A/D converter
- Sample and hold circuit
- A/D converter: This measures the input analog voltage and output a digital code corresponding to the input voltages.
- Bus interface and board timing system.

Typical analog input modules features include:

- 8, 16, or 32 analog inputs
- Resolution of 8 to 12 bits
- Range of 4-20 mA
- Input resistance typically 240kohms to 1 Mohms
- Conversion rates typically 10 microseconds to 30 milliseconds.

*C. Analog Output Module*

Analog Output modules function is to convert a digital value supplied by the CPU to an analog value by means of a digital to analog converter. This analog representation can be used for variable control of actuators.

Analog output modules features are as follow:

- 8, 16 or 32 analog outputs
- Resolution of 8 or 12 bits
- Conversion rate from 10µ seconds to 30 milliseconds
- Outputs ranging from 4-20 mA/0 to 10 volts

*D. Digital Input Modules*

These are used to indicate status and alarm signals [5]

*E. Digital Output Modules*

These modules are used to drive an output voltage at each of the appropriate output channels with three approaches possible.

- Triac Switching: Triacs are used to achieve AC power control. The Triac responds primarily as a switch, the AC energy source for a portion of each alternation can be controlled [6]
- Read Relay Switching

- TTL voltage outputs

### F. Power Supply Module

RTUs need a continuous power supply to function, but there are situations where RTUs are located at quite a distance from an electric power supply. In these cases, RTUs are equipped with alternate power source and battery backup facilities in case of power losses.

Solar panels are commonly used to power low-powered RTUs, due to the general availability of sunlight. Thermo electric generators can also be used to supply power to the RTUs where gas is easily available like in pipelines. [4]

### G. Communication interfaces

Modern RTU are designed to be flexible enough to handle multiple communication media such as

- RS 232/RS 442/RS 485
- Ethernet
- Dial up telephone lines/dedicated landlines
- Mircrowave/MUX
- Satellite
- X.25 packet protocols
- Radio via trunked/VHF/UHF/900 Mhz

## IV. GATEWAY

A gateway is a device with dedicated hardware and software that translates between two different protocols, making communication possible between networks of different architectures and protocols. The job of a gateway is much more complex than that of a network router or switch due to this conversion functions. Gateways in digital oil fields collate data from the RTUs in the field and remote sites and integrate these data into the Companies IT network

Gateways are necessary for communication between terminals connected to heterogeneous networks using different protocols and having different network characteristics. They provide the connectivity between systems at remote locations with the target system to enable different network applications

A gateway can function as a protocol gateway which converts between protocols; an application gateway which accepts inputs in one format translates it and then sends it, or a security gateway which basically acts as a firewall securing and filtering packets. [7]

Components of a gateway

- Microprocessor
- Motherboard
- RAM
- Flash
- Interface boards as I/O ports

## V. SELECTING THE RIGHT EQUIPEMENT

The selection of RTUs and Gateways are based on the specifications of the implementation in terms of data type, capacity and transmission rate. These parameters include [8]

*1)* Capacity: *The RTU must be able to support the data transmission frequency and the data rates. The RTU must also have sufficient spare IOs to allow for expansion*.

*2)* Environment Factor: *The RTU must be able to withstand the environmental factors and be designed to the required ingress protection ratings and installed in the appropriate hazardous area classification.*

*3)* Control: *RTUs are also used for applications requiring different control schemes such as on/off and variable control. Control relays serve as the control element and are connected to the RTU and activated remotely to achieve the desired control. During RTU selection, the type of control required must be defined to ensure that the selected RTU supports the control system.*

*4)* Connectivity: *The network connectivity requirements and data formats must defined before the RTU and Gateways are selected. The data exchange format must also be defined before the selection of the RTU and gateways are finalized.*

*5)* Upgradeability*: The ease of firmware upgrade is also another key parameter as it will be desired for the RTU and gateway to be upgradeable over the air without the requirement of un-installation and office based upgrade.*

*6)* Transmission range: *The RTU frequency and range are to be confirmed to be suitable for the application and this is usually confirmed by network plan. The equipment are expected to operate within the approved frequency band and with the required licenses.*

*7)* Power: *The power supply requirements should also be within the required capacity and quality and these power supply systems must be able to with stand the environmental factors, cost and weather conditions and also be resistant to vandalization. Options include solar panels, batteries or other power source. Also putting into consideration cost as well*.

## VI. RTU MANUFACTURERS

There are various manufacturers of Remote Terminal Unit for various functions and industries. A list of some RTU manufactures and their products are presented in the Table 1

*1)* Vmonitor iX-S8 Wireless RTU: *An intelligent remote terminal unit and wireless technology with low power consumption to provide a reliable and cost effective means of remotely monitor and automate your applications in the oil and gas fields. [9]*

*2)* ControlWave® Micro Hybrid RTU/PLC: *A highly programmable controller that combine the unique capabilities of a programmable logic controller (PLC) and a remote terminal unit (RTU) into a single hybrid controller. [10]*

*3)* Zetron Model 1732 RTU: *A cost-effective solution for applications that need to connect widely distributed remote sites to a central control program using radio, telephone and wire line communications media. [11]*

*4)* Brodersen RTU32: *Brodersen RTU32 RTU, PLC and controller series based on a 32-bit platform provides RTU/PLC with power and leading edge functionality. [12]*

*5)* Siemens Vicos RTU: *A telecontrol with standard SIMATIC S7 Programmable Logic controller [13]*

*6)* Oleumtech Wireless RTU/Modbus Gateway: *Wio wireless RTU products are low cost remote terminal units that combine traditional remote IO functionality of a standard [14]*

## REFERENCES

[1] Borin Manifacturing. Borin Manifacturing, Inc. Borin [online] http://www.borin.com/remote-terminal-unit/. [Accessed 24th July 2012]

[2] TheWater Environment Federation, "Automation of waste water treatment facilities." WEF Press McGraw-Hill Companies, 2007.

[3] Clarke, G.R., Reynders,D and Edwin W, "SCADA protocols: DNP3, 60870.5 and related systems." Elsevier, 2004. pp.15 - 25

[4] Shaw, W.T. "Cybersecurity for SCADA systems." Pennwell books, 2006. pp.25 - 32

[5] Sumathi S. and Surekha..P. "LabVIEW based advances instrumentation sustems." Springer, 2007. pp 242 - 246

[6] Patrick, D.R. and Fardo, S.W. Electricity and electronics fundamentals.: Fiarmont Press, Inc, 2008.

[7] Zhang, P and William A. "Advnced Industrial Control Technology." 2010. pp 378 – 379

[8] DPS Telecom. [online] http://www.dpstele.com. [Accessed 24th July 2012]

[9] VMonitor. VMonitor.[online] http://www.vmonitor.com. [Accessed 24th July 2012]

[10] Emerson Process Mangement. [Online] http://www.documentation.emersonprocess.com [Accessed 24th July 2012]

[11] Zetron. [online] http://www.zetron.com. [Accessed 24th July 2012]

[12] Borderson. [online] http://brodersensystems.com. [Accessed 24th July 2012]

[13] Siemens. [Online] https://concert.siemens.com. [Accessed 24th July 2012]

## AUTHORS PROFILE

Francis Idachaba (idachabafe@yahoo.com) a lecturer with Covenanrt University Ota is currently with on a Fellowship with Shell Petroleum Development Company in Nigeria.

Ayobami Ogunrinde (ayo.ogunrinde@gmail.com ) is currently with SPDC in Nigeria.

TABLE 1. RTU MANUFATURERS SPECIFICATION

| | Vmonitor iX-S8 | Control Wave Micro Hybrid | Zetron 1732 RTU | Broderson RTU 32 | Siemens Vicos RTU |
|---|---|---|---|---|---|
| *Processor* | 8051 Micro-Controller running at 12.58Mhz 64KB Data Memory | 32-bit ARM 9 Processor, 33 MHz CPU Module (Low power) 150 MHz Module | PLC | 32-bit 500 Mhz CPU with 128K L2 Cache | SIMATIC S7 – 300 Central Units |
| **Serial Port** | RS232<br>RS232/RS485<br>I2C | RS232<br>RS485<br>Ethernet | RS232 | Dual Ethernet<br>RS232<br>RS232/RS422/RS485<br>USB | RS232<br>Ethernet |
| **Comm. Protocol** | Vmonitor Proprietary Protocol<br>MODBUS | MODBUS, Foundation Fieldbus. HART, DFI, CIP, DNP3, Serial ASCII | MODBUS | MODBUS, DNP3 Suite, ProfiNET Client, PROFIBUS DP Master, COMLI, IEC 61400-25 | IEC60870-5, SINAUT8-FW PCM, Profibus, ProfiNET |
| **Operation Modes** | Continuous, Stand alone and Sleep Mode | Run Mode, Remote Mode, Local Mode and Sleep Mode for Low power application | Fail-Safe Mode | | |
| **Radio** | 900 MHz or 2.4 GHz | | 902-928MHz | | 902-928MHz |
| **Distance** | Up to 5 Miles | | | | |
| **Operating Temperature** | (-40 to 85)°C Humidity (5 – 90%) | (-40 to 70)°C Humidity (15 – 95%) | (0 to 50)°C Humidity (0 – 90%) | (-40 to 70)°C | (0 to 60)°C Humidity (5 – 95%) |
| **Power Supply** | Input Power 8-15VDC (Battery Powered) | 9 – 30 VDC | 16VDC 2A Max, 120 VAC | 24 – 48VDC<br>115-230 VAC/DC<br>115-230 VAC/DC with UPS and Battery | 24VDC, 230 VAC |
| **Integrated Analogue/Digital Input** | (AI): 1-5 V, 4-20mA<br>(DI): 3-24VAC | (AI): 1-5 V, 4-20m A<br>(DI): 12-24VDC | (AI): +/- 10 VDC, 4-20 mA | (AI): 0-10V, 0-5V, +/- 5V, +/- 10V, 0-20mA, 4-20mA<br>(DI): 10-30VDC | |