IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 5 Issue 9

September 2014

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

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

SAI

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the August 2014 Issue of International Journal of Advanced Computer Science and Applications.

Today, it is incredible to consider that in 1969 men landed on the moon using a computer with a 32-kilobyte memory that was only programmable by the use of punch cards. In 1973, Astronaut Alan Shepherd participated in the first computer "hack" while orbiting the moon in his landing vehicle, as two programmers back on Earth attempted to "hack" into the duplicate computer, to find a way for Shepherd to convince his computer that a catastrophe requiring a mission abort was not happening; the successful hack took 45 minutes to accomplish, and Shepherd went on to hit his golf ball on the moon. Today, the average computer sitting on the desk of a suburban home office has more computing power than the entire U.S. space program that put humans on another world!!

Computer science has affected the human condition in many radical ways. Throughout its history, its developers have striven to make calculation and computation easier, as well as to offer new means by which the other sciences can be advanced. Modern massively-paralleled super-computers help scientists with previously unfeasible problems such as fluid dynamics, complex function convergence, finite element analysis and real-time weather dynamics.

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.

Lastly, we would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that materials contained in this volume will satisfy your expectations and entice you to submit your own contributions in upcoming issues of IJACSA

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 5 Issue 9 September 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Associate Editors

Reviewer Board Members

CONTENTS

Cloud Based Public Collaboration System in Developing Countries

Sherif M. Badr
Chairman of Information and Decision Support Center
The Egyptian Cabinet
Cairo, Egypt.

Sherif E. Hussein
Computer and Systems Department
Mansoura University
Mansoura, Egypt.

***Abstract*—Governments in developing countries are increasingly making efforts to provide more access to information and services for citizens, businesses, and civil servants through smart devices. However, providing strategically high impact m-services is facing numerous challenges, such as complexity of different mobile technologies, creating secured networks to deliver reliable service, and identifying the types of services that can be easily provided on mobile devices. Those problems could be solved by applying cloud computing model to the business process of E-government to build a government cloud. This research, proposes an environment for citizens to have greater access to their government and, in theory, makes citizen-to-government contact more inclusive. In addition, it examines an application that allows anyone to report and track non-emergency issues via the internet. It can also encourage citizens to become active in improving and taking care of their community by reporting issues in their neighborhood in order to improve the Egyptian e-government development index.**

Keywords—e-government; m-government; cloud computing

I. Introduction

The emergence of new information and communication technologies has not only revolutionized the way business is conducted but also transformed the delivery mechanism of governmental services. Since the 1990s, public-sector organizations across the globe have been applying internet technology and other computing technologies in innovative ways to deliver services, engage citizens, and improve efficiency using a set of practices commonly known as electronic government (e-government) [1].

E-Government is rapidly becoming one of the government's critical means for the provision of seamless services for public agencies, businesses, and citizens [2, 3]. It is the use of information and communication technologies as a tool to achieve better government. Its innovation and development can position the public sector as a driver of demand for Information and Communication Technology (ICT) infrastructure and applications in the broader economy [4]. Moreover, an explosion in the use of mobile technologies, such as smart phones, laptops, and tablets to connect to wireless networks has enabled governments to change from e-government to m-government. M-Government is an emerging trend in government services and applications delivery [5].

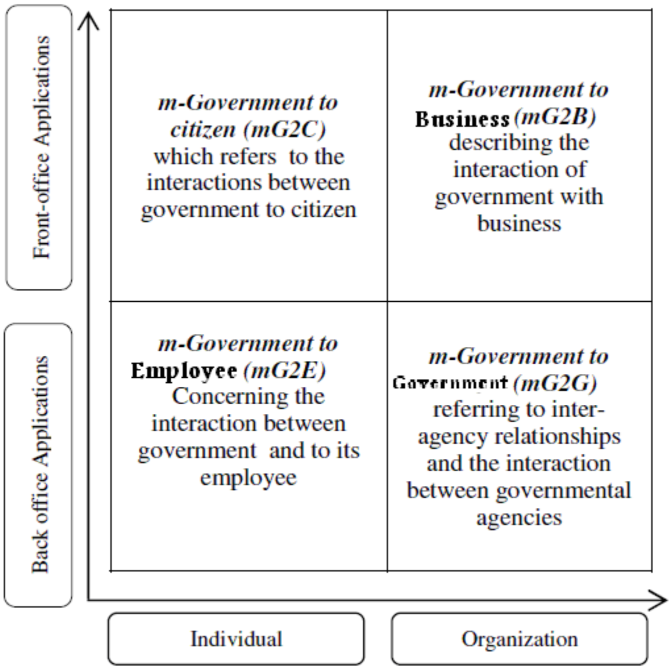
It creates and guarantees mobility and portability for the public, business, and government. Furthermore, real-time access to information and personalization of information access are guaranteed to maximize benefits of using information and, in turn, create further advanced e-Government services [6, 7]. However, providing strategic impact, and secured m-services, dealing with the complexity of different mobile technologies, and creating secure networks to realize reliable services are the major challenges for government institutions [8].

The potential for m-Government in developing countries, however, remains largely unexploited, even though, governments in developing countries are increasingly making efforts to expand mobile networks infrastructure to provide more access to information and services for citizens through wireless devices [9, 10]. While e-Government encompasses usage of all technologies to deliver services to citizens and improve the activities of government and streamline their processes, m-government presents an expansion to the e-Government to use mobile technologies in delivering services.

In addition, m- Government is a better option compared to e-Government in delivering services and public information to citizens due to its nature of being available anywhere, anytime and from any internet enabled device. The trend towards m-Government has been facilitated by growing capabilities of mobile technologies and their associated infrastructures, devices and systems and their acceptance in both developed and developing countries [11].

M-Government can be applied to four main purposes such as m-services, m-communications, m-administration, and m-democracy in the public sector. Like e-Government, m-Government operates on four different levels of interactions as identified in figure 1.

Figure 1 describes the four levels of interactions which are m-government-to-citizen (mG2C), that refers to the interaction between government and citizens; m-government-to-business (mG2B), describing the interaction between government and businesses; m-government-to-employee (mG2E), also known as internal effectiveness and efficiency (IEE), is concerning the government and its employees; and m-government-to-government (mG2G), referring to inter-agency relationships and the interactions between government agencies [12].



m-Government to citizen (mG2C) which refers to the interactions between government to citizen

m-Government to **Business** *(mG2B)* describing the interaction of government with business

m-Government to **Employee** *(mG2E)* Concerning the interaction between government and to its employee

m-Government to **Government** *(mG2G)* referring to inter-agency relationships and the interaction between governmental agencies

Fig. 1. The four levels of interactions are m-government-to-citizen (mG2C), m-government-to-business (mG2B), m-government-to-employee (mG2E), and m-government-to-government (mG2G).

There are several attractive features that prompt shift toward m-Government in developing countries [13]:

- Number of mobile users and increasing penetration: More people than ever have ownership of mobile devices capable of accessing e-services and e-contents.
- Mobiles connecting people to the Internet: Urban users are using mobiles to receive an “Internet experience” through Wireless Application Protocol (WAP) services provided over General packet radio service (GPRS).
- Mobility: Enables people to access content wherever they are.
- Inclusiveness and Remote area access: Mobile phones, can reach those areas where the infrastructure necessary for Internet services or wired phone services is difficult to setup. In the developing countries mobile government applications may become a key method for reaching citizens in far and wide areas and promoting exchange of communications. In such countries with insufficient conventional telecom infrastructures and greater acceptance of mobile phones, the ability of reaching rural areas may be considered as an important feature of m-Government. Mobile technologies increase inclusion of the most marginalized people in society.
- Low Cost: Mobile phone is relatively a low cost technology, which the common person can afford to have it as compared to internet technology.
- Ease of Learning: Usage of mobile devices is fairly simple, making it easy for any person to use and to access information.
- Easy Infrastructure Setup: Due to the simple architecture of mobile telephony, new mobile phone networks can be easily installed in countries where infrastructure complexity and cost are not feasible.
- Improvement on e-Government effort: M-government is not a replacement to e-Government but complementary to it. Also, it helps in expanding the scope of e-Governance in areas like e-Democracy, e-Participation, e-Voting and many other forms of communication between the citizen and the government.

The application designed and evaluated here is a Public Collaboration System (PCS). It is a collaboration between citizen and government. The proposed application aims to use wireless/mobile technologies to provide information and services for citizens (mG2C). PCS allows anyone to report and track non-emergency issues via the internet as well as encouraging citizens to become active in improving and taking care of their community by reporting issues in their neighborhood. There are already a lot of involved citizens and hard-working local authorities and service people. The proposed application seeks to use the potentials of the internet and mobile computing to bring them closer together and reach even more people.

II. Methodology

To achieve citizen approval and extensive use of m-government services, the selection of both technology and services should match the actual requests of people.

Consequently, limitations and strong points inherent to the use of mobile and wireless technologies for public service provision should be considered through the prism of user centricity.

The crucial aim of introducing government services via new technological means is to create an added value. Undeniably, the added value of m-government is mobility itself. Nevertheless, the value given by mobility comes, not from the accessibility of certain services through mobile phones, but from capability of these services to support mobility of the user. Different aspects of mobility and its inferences to government environment are vital for appreciating the individuality of m-government concept, as well as obstacles and motivations to m-government service adoption [14].

The same technological features of a mobile device (small screen, miniaturized keyboard, etc.) constraining the application of enhanced services to mobile environment, on the other hand, constitute additional value for the user, as they increase the mobility level of the device.

The United Nations Public Administration Programme (UNPAP) has developed a number of indices to measure the e-government performance. Those measures are updated annually since its creation in 2003. It covers all Member states of the UN. Among those measures is the E-Government Readiness Index (EGDI) which is a composite measure of the capacity and willingness of countries to use e-government for ICT-led development countries to use e-government for ICT-led development. The EGDI looks at the most important dimensions of e-government: (i) scope and quality of online services, (ii) telecommunication connectivity, and (iii) human capacity [15].

Governments' efforts are ranked, while countries size, infrastructure availability and ICT penetration, and the level of education and skill development are taken into account. Closely connected to the survey, the UNPAP also produces an E-Participation Index. The index rates the performance of national governments relative to one another by averaging three other indices: the Online Service Index, the Telecommunication Index and the Human Capital Index. The maximum possible value is one and the minimum is zero. Though the basic model has remained constant, the precise meaning of these values varies from one survey to the next as understanding of the potential of e-government changes and the underlying technology evolves.

The United Nations Survey 2012 assessment of progress indicates that e-government is increasingly being viewed among countries in the vanguard as going beyond service delivery towards a framework for a smart, inclusive and sustainable growth for future generations. In countries that follow that trend, a focus on institutional integration coupled with online citizen orientation in public service continues to be dominant. Both in terms of information and services, the citizen is increasingly viewed as ‘an active customer of public services’ with borrowed private sector concepts being applied to improve public sector governance systems. A key driver for this approach is the need to achieve efficiency in government at the same time that services are being expanded. Advances in technology, which allow data sharing and efficient streamlining of cross-agency governance systems are forming the back-end of integrated portals where citizens find a myriad of relevant information arranged by theme, life cycle or other preferred use. The trend towards personalization of services has gained momentum with more countries tailoring substance and presentation in accord with varied preferences. Multichannel service delivery features were found on several portals in 2012 through which the government conducted business with citizens. Citizen inclusion is also expanding both horizontally and vertically with more governments around the world in 2012 accepting and promoting the need to inform – and involve – the citizen in the public decision making process [16].

E-government innovation and development can position the public sector as a driver of demand for ICT infrastructure and applications in the broader economy. The effect will be more pronounced in cases where government programmes constitute a significant proportion of a country’s GDP and where the regulatory environment is conducive to expansion of ICT manufacturing, software and related services. E-government programmes can be a catalyst in boosting productivity, thereby speeding up the benefits of newer technologies to the people. In the last few years many countries have employed ICT in areas such as entrepreneurship, innovation, research and development, promoting distance learning, e-health, e-agriculture, e-trade and other fields. Accessing these new technologies for development is being recognized as one of the key sources of economic growth. Of particular importance is the effect of cellular technologies. Where national governments have taken a lead, rapid mobile technology proliferation has contributed as much as a one percent annual increase in economic growth over the last few years. Not with standing these trends, progress remains uneven. In the current recessionary climate some countries have been better able to continue to invest in ICT infrastructure and service improvement. Others are evaluating the marginal utility of such investment, especially taking into account low user uptake of existing services, and reassessing service portfolios where demand for online services is low. Many countries with low levels of infrastructure and human capital remain at lower levels of e-government development with serious issues of digital divide [17].

In all cases, e-government take a prominent role in shaping development making it more in tune with people’s needs and driving the whole process based on their participation.

Africa has seen improvement in e-government with countries in the region looking to increase their online presence through developing websites for government ministries and agencies. Table 1 shows that Seychelles (0.5192) climbed several points to number one in the region in 2012 followed by Mauritius (0.5066) and South Africa (0.4869).

TABLE I. E-GOVERNMENT TOP RANKED COUNTRIES IN AFRICA

Rank	Country	E-gov. development index ranking		World e-gov. development	
		2012	*2010*	*2012*	*2010*
1	Seychelles	0.5192	0.4179	84	104
2	Mauritius	0.5066	0.4645	93	77
3	South Africa	0.4869	0.4306	101	97
4	Tunisia	0.4833	0.4826	103	66
5	Egypt	0.4611	0.4518	107	86
6	Cape Verde	0.4297	0.4054	118	108
7	Kenya	0.4212	0.3338	119	124
8	Morocco	0.4209	0.3637	120	126
9	Botswana	0.4186	0.3637	121	117
10	Namibia	0.3937	0.3314	123	125
Regional Average		0.2780	0.2733		
World Average		0.4882	0.4406		

It is notable that all of the leading African countries increased their e-government development index value in 2012 but lost in comparative performance around the world, except for Seychelles, Kenya and Morocco, which gained in the world rankings from 104 to 84, 124 to 119 and from 126 to 120 respectively. Tunisia (0.4833) and Egypt (0.4611) declined in rank substantially as did Cape Verde (0.4297) because their improvements did not keep pace with those of other countries around the world [15].

An example of a big improvement in EGDI was in Kazakhstan. Kazakhstan has improved from 2010 in terms of providing online features that allows citizens to engage with government. An interesting online participation feature is the government’s Blogs site, where citizens can communicate with the government agencies’ executives by posting comments and questions. The executives may then respond and post their answers on the blog. The site also contains statistical information on the questions and comments an agency executive has received as well as how many times he/she has responded. Another example was Bahrain. Bahrain's e-government strategy is based upon "delivering customer value through collaborative government." The government sees citizens as customers who have different needs and demand different services and at the same time demand value for money. Thus the aim of e-government is to provide all services, integrated, to all citizens and upon their choice of channel. The Kingdom provides delivery of services through the following channels: e-government portal, mobile portal, national contact centre (a 24-7 call centre) and e-services centers and kiosks. Bahrain has introduced the "Listen" feature, which enables people with visual disabilities to hear any text available on the website with the click of a button. Another very innovative feature is the e-government toolbar, which can be downloaded permanently to your browser. This allows direct access to e-services and RSS feeds without having to go to the main portal [18].

Though, there is considerable progress in the expansion of online services, one of the primary challenges that remain in Least Developed Countries (LDC's) is integration of back-end processes with efficient, user friendly, and target oriented services delivery. The proposed application tackles those limitations using a back-end deployed over the cloud. That kind of deployment takes advantage of both unlimited scalability and minimum cost as will be explained in the implementation section.

Moving from improving public sector efficiency, Europe looks to take this role further in adapting innovative technologies to human development and economic sustainability in the future.

E-government innovation and development can position the public sector as a driver of demand for ICT infrastructure and applications in the broader economy. For e-participation to contribute to sustainable development and the socio-economic uplift of the people, the role of government requires a shift from that of a controller of information and services to that of a proactive facilitator. In this context, it is imperative that information and services are geared toward promoting user uptake, addressing the needs and concerns of the citizenry, especially the vulnerable. It also requires viewing the citizens not only as passive receivers of information through web based services, but also as active partners who are engaged and supported to interact with the government through ICT-based dissemination of relevant government information. The best performing countries in e-participation appear in table 2. Once again the Republic of Korea tops the list, but this year it is joined by the Netherlands. Kazakhstan (0.9474), a developing country, which was noted in the 2010 Survey for its commitment to e-participation, moved up 16 places to be ranked second and tied with Singapore. Among this group several other countries were tied for the same spot, such as Australia, Estonia, and Germany, which were all at the 5th position. With the use of consultation tools, including social media, other developing countries have also caught up to the developed countries as e-leaders. Notable among these are Bahrain, Egypt, the United Arab Emirates, Colombia, and Chile. Europe's share of the top ten fell from 51 per cent in 2010 to 38 percent this year. This change was primarily the result of the Americas increasing from 14 per cent to 19 per cent with Chile and Colombia joining the leaders, along with the appearance of Egypt from Africa, and Bahrain and the United Arab Emirates from Western Asia [15].

TABLE II. WORLD E-PARTICIPATION RANKING

E-participation index	Country	Rank
1.0000	Netherlands	1
1.0000	Republic of Korea	
0.9474	Kazakhstan	2
0.9474	Singapore	
0.9211	United Kingdom	3
0.9211	United States	
0.8947	Israel	4
0.7632	Australia	5
0.7632	Estonia	
0.7632	Germany	
0.7368	Columbia	6
0.7368	Finland	
0.7368	Japan	
0.7368	United Arab Emirates	
0.6842	Egypt	7
0.6842	Canada	
0.6842	Norway	
0.6842	Sweden	
0.6579	Chile	8
0.6579	Russian Federation	
0.6579	Bahrain	

Table 2 shows the potentials for some countries to improve their EDGI. Governments with high e-participation values and low EDGI need to re-evaluate their e-government policies and to take advantage of the willingness of their citizens to be an active part in the government decision making.

III. SYSTEM ANALYSIS

As an attempt to increase the world e-government development ranking in developing countries, a system has been designed to take advantage of the already existing mobile users that are willing to participate in the government policy making and take advantages of any available mobile services. The proposed system deals with many government factions and departments that offer public services. It allows users to report different problems they can find (road hazard, crowd, environmental pollution or risk, unlawful use of public resources, etc.) and vote for published ones and at the same time trace the government procedures to solve them. The system had to include a validation module to screen, classify and group problems to be easily administered. It had to use as well an expert system module to support both users and government departments for the best procedure for a particular problem. The expert system knowledgebase increases as the system serves more users. Another important feature for the system is the ability to load balance user requests when number of users increase. That feature uses a platform as a Service (PaaS) which can allocate resources unlimitedly with proper security measures. The platform is typically represented as a single box [19]. Since the platform usually acts as if it were a single box, it's much easier to work with, and generally there is no need to change much in the application to be able to run on a PaaS environment. PaaS doesn’t only offer CPU, memory or file storage; but also offers other parts of the infrastructure, such as databases, either in the form of a scaling traditional RDBMS system, or one of the ’NoSQL’ databases [20, 21].

Figure 2 describes the features the moment citizens start to participate in the system. At first the user from the client side usually a mobile device or a desktop takes a picture of the problem he sees, adds notations, and sends it through the web to inform the authorities about it. He can also review other problems that have been published before. As an added feature to the system, is the ability of the user, if he is in a place without internet coverage, to automatically save the problem in the internal database of the client. When internet connection becomes available, the system sends the saved problems upon user request. After the problem is sent, a copy is saved in the back-end database to be opened and filtered by a validation module to check for its consistency before to be saved in the database published on the website. The government reviews the approved problems and sends their progress to the web server.

The user can review other problems using his dashboard from a mobile or a desktop through a website URL or can search for certain problems or just check the most recent problems and vote for them.

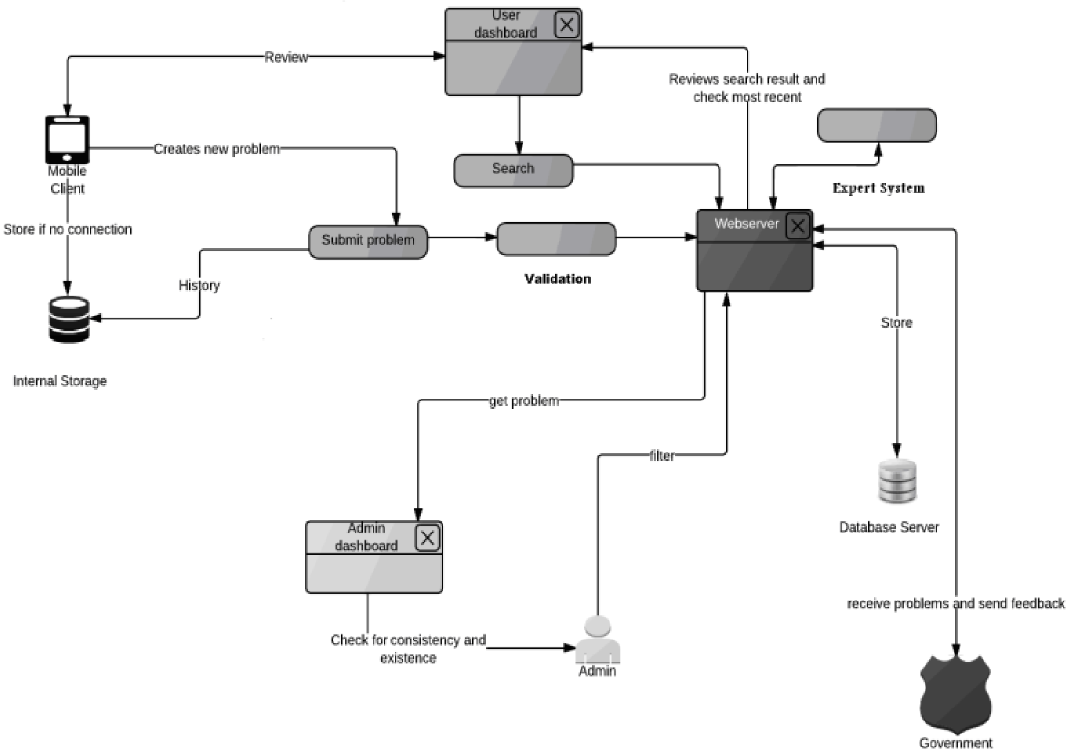


Fig. 2. PCS life cycle diagram

A. Sequence diagram

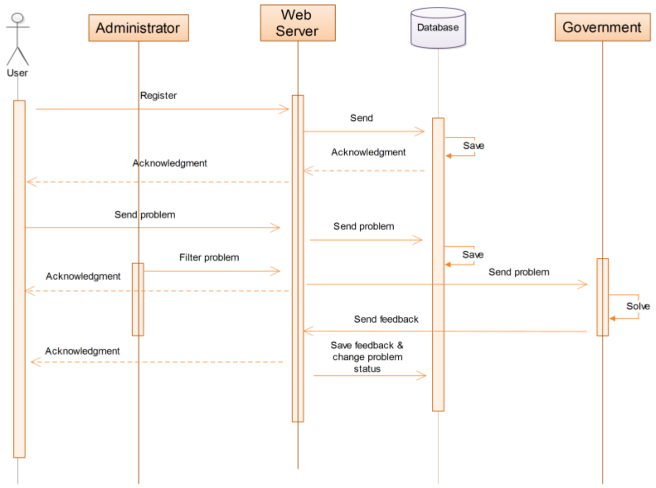


Fig. 3. PCS sequence diagram

The sequence diagram shown in figure 3 describes objects interactions arranged in time sequence. It depicts the objects involved in the scenario and the sequence of actions exchanged between the objects needed to carry out its functionality. If the user is not a registered member, the system allows him to register before the problem data is sent to the web server. The validation module then checks the consistency and validity of the problem sent. If the problem is inconsistent with the registration regulation the user is suspended after administrative review and no further action will be taken to process the problem by the government. Filtered problems will be published and an acknowledged will be sent to the user. The government will review the problems and change its status according to the progress it has achieved. The user can see his problem progress or other published problems through the front-end.

B. Context diagram

The Context Diagram shown in figure 4 describes the scope of the system which consists of two entities (citizen and government), the citizen can register and send a problem to the system, then receive acknowledgement of the process status. Government receives the problem details from the system then provides the system with a feedback for the problem progress. There are few levels that depict the context diagrams which are summarized as follows:

Level 0: In this level, the system is decomposed into front-end (mobile application or web application) and web server system, citizen can register through the web or the mobile then receives an acknowledgement by the process status.

Level 1: In this level, the front-end is decomposed into three processes. The first process checks whether the user is registered or not, and if he is already registered, it sends user data to the database server. In the second process, the problem system receives problem data including (image, description, location, and category) and responds by an acknowledgement to notify users by the process status. In case the citizen has no internet connection, the problem data will be saved into an internal database until connection is available, and then will be sent to the web server.

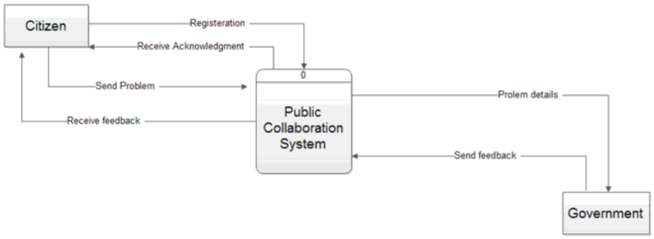


Fig. 4. PCS context diagram

Level 2: In this level, the system is decomposed into three processes. Receive and send data, manage reports, check the registration status. The first module receive problem that and send it to the database server. The manage report change data status according to a validation filtration and the government solution progress. The registration module can register a new user or check whether he is a registered member or not.

Level 3: In this level, manage report is decomposed into three processes; filter report, publish report, and feedback system. The problem is sent to be filtered by a validation module then detects if the problem exists and whether it is consistent, then sends it to the publish report. If it is really a problem, then it sends it to the government. If not, it discards it before notifying the user with its status and changes the user status to be suspended.

C. The validation module

As more and more users use the service, overload could occur (too many submissions, duplicate submissions—same problem). Thus, for administrators and government not to waste too much time and effort organizing and collating submissions, a validation module is needed. It can organize submissions: group similar problems (and create statistical data) according to keywords in submission (text), pattern recognition (image), and statistics based on location, time, etc. The module can assign higher priority to problems from trusted users (users with previous popular submission according to public votes). It can also increase the priority of problems according to votes or number of occurrence.

D. The expert system module

An expert system gives advices to guide citizens and government the best procedure to solve a problem according to previous solutions stored in the expert system's knowledgebase. The expert system can serve as well to group problems data and help users categorizing their problems. ES can continuously learns new solutions as they are solved by the government and tracks unsolved problems reasons and suggests solutions.

IV. System Design

In the proposed system, the architecture was separated into software and service architecture:

Software architecture: The set of structures needed to reason about the software system, which comprises the software elements, the relations between them, and the properties of both elements and relations. Figure 5 shows the PCS mobile interface with its different application components, It needs to have consistent look irrespective to which smart phone is using the PCS application.

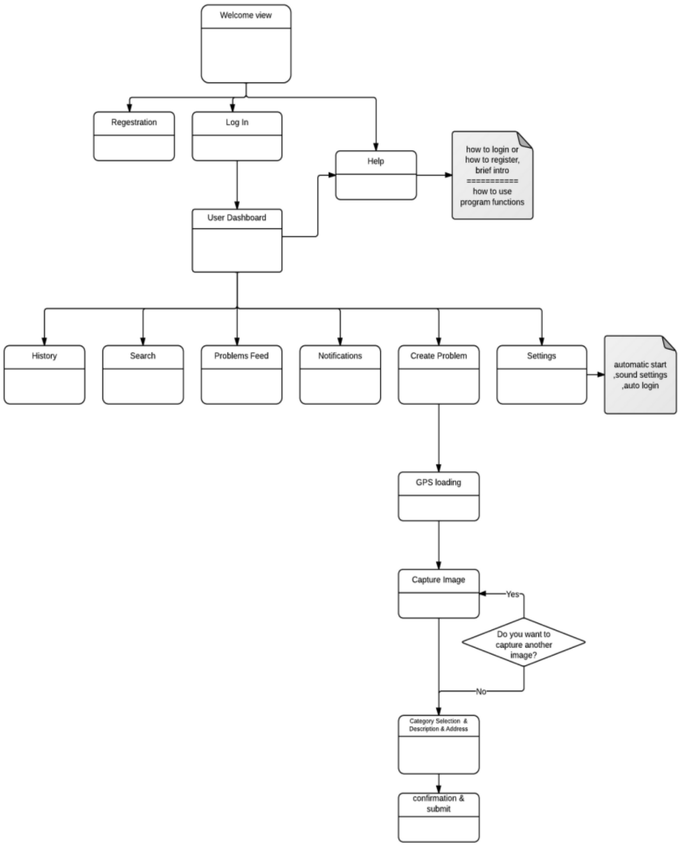

Welcome view
Regestration
Log In
Help
how to login or how to register, brief intro
===========
how to use program functions
User Dashboard
History
Search
Problems Feed
Notifications
Create Problem
Settings
automatic start ,sound settings ,auto login
GPS loading
Capture Image
Yes
Do you want to capture another image?
No
Category Selection & Description & Address
confirmation & submit

Fig. 5. The PCS mobile front-end structure chart

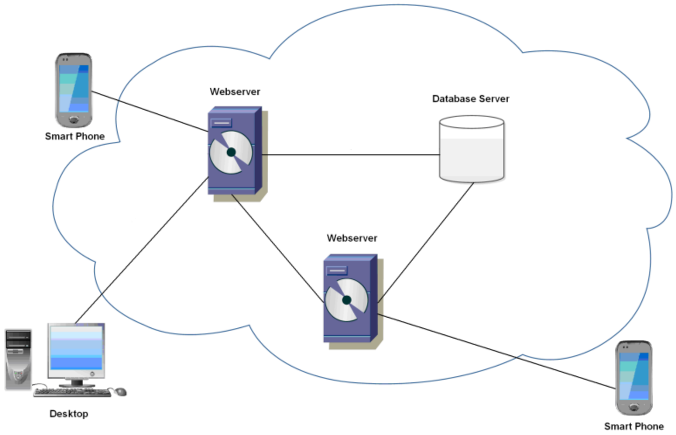


Fig. 6. The PCS front-end (the mobile interface and the application website) and the back-end (the application servers and the data storage).

Service architecture: The physical design of an individual service that encompasses all the resources used by a service. This would normally include databases, software components, legacy systems, identity stores, XML schemas and any backing stores. It is also beneficial to include any service agents employed by the service, as any change in these service agents would affect the processing capabilities of the service.

The PCS application design was divided into the front-end and back-end design as shown in figure 6. The front-end was either an application installed on the smart phone or a website accessible from any device with a browser. On the other hand, the back-end was the server side part that manages the PCS's services and databases. The back-end is supported by ColdFusion and Amazon Machine Image (AMI) to allow the PCS to scale over the cloud to accommodate the possible increase in the number of users.

Front-end Design: In the design stage, Adobe Flash Builder has been used to simplify the integration of Adobe ColdFusion and Adobe Flex framework to create the PCS as a SWF application. ColdFusion components (CFCs) have been imported into the application and the ability to access data was done through service calls to the CFC functions. The mobile interface design was implemented using the Flash builder and the integrated MXML and FX components.

Android and IOS are both supported by Adobe Integrated Runtime environment (Adobe AIR) which can give a consistent feel for users. The functions that manipulate data on the mobile, retrieved from the web service through REST and URL requests, were coded by Actionscript integrated into Flash builder [22].

ActionScript, used in PCS implementation, is a dialect of ECMAScript (it is a superset of the syntax and semantics of the language more widely known as JavaScript), and is used primarily for the development of websites and software targeting the Adobe Flash Player platform, used on Web pages and mobile devices in the form of embedded SWF files.

The last phase was to construct the website using the flex MXML components in Flash builder which provides flexibility and scalability of running on any browser supported by Adobe Flash Player [23].

Back-end Design: Adobe ColdFusion on Cloud offers a way to easily leverage ColdFusion as a scalable service through Amazon Web Services (AWS). The ability for a Flex application to directly access data from the server is powerful because it allows for dynamic content to update the Flex application user interface without refreshing the HTML page. It is also a great advantage because it decreases number of requests to the server and the amount of content that is transferred over the network as shown in figure 7.

The AMI is available for Windows and Ubuntu operating systems. ColdFusion Windows AMI comes with the recent releases of windows operating system, configured with IIS web server (Internet Information Services) and a pre-installed MySQL database server. ColdFusion Ubuntu AMI comes with Ubuntu operating system, configured with Apache web server and a pre-installed MySQL database server. Both AMIs are available for Large and Extra-large instance types of AWS. Both AMIs have JRE built into them along with the ColdFusion Hot Fix Update.

ColdFusion is pre-installed on the Ubuntu and Windows AMIs. They can be configured using jumpstart tool. The credentials for MySQL database server were changed after installing the ColdFusion AMI [24].

After accessing ColdFusion AMI on AWS EC2 (Amazon Elastic Compute Cloud) instance, the instance is connected and the jumpstart tool launches and runs automatically.

Amazon provides a security mechanism to AWS instance through "security group." You can add rules to each security group that control the inbound traffic allowed to reach the instances associated with the security group. You can select and apply only those rules which are required. For Windows instances, to enable FTP-related functionality, Turned On option, have been selected in Windows Firewall and Notify me when Windows Firewall blocks a new program option has been selected as well.

By default, for ColdFusion instance set up on AWS, maximum JVM heap size is set to 512 MB. This can be configured according to our requirement based on memory available for selected instance type. For large instance, memory available is 7.5 GB and for extra-large instance it is 15 GB [25].

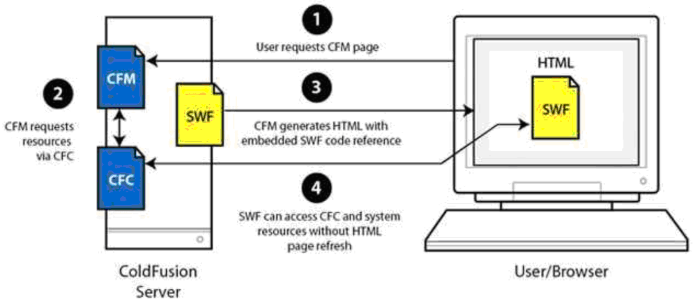

1
User requests CFM page
2
CFM requests resources via CFC
CFM
SWF
CFC
3
CFM generates HTML with embedded SWF code reference
HTML
SWF
4
SWF can access CFC and system resources without HTML page refresh
ColdFusion Server
User/Browser

Fig. 7. The PCS front-end and back-end interaction

V. Discussion and Conclusion

M-government is the new frontier of the development agenda which needs to be prioritized by the development community at large. Developing countries should take a closer look at the potential of mobile technologies to enable better access to public information and services for the masses and adjust their current strategies, programs and processes accordingly. M-Government needs to be implemented as an integral part of e-Government. Moreover, a priority list of several high-impact m-services and a larger list need to be developed for rapid implementation by each government. The proposed system offers more collaboration between citizens and government. That system can be one of the e-government frontiers which take advantage of mobile devices to reach the majority of Egyptian citizens and improves Egypt e-government development ranking.

The use of cloud technology in the proposed system offers reliability, sustainability as the cloud relies on government resources or public networks, so network paralysis, bandwidth bottleneck or unexpected problems could be avoided and stability of data access and sustainability of business could be achieved.

As explained in this research, e-government systems need to offer data security, as cloud technology has its security risks which need to be carefully taken into account.

References

[1] T. Almarabeh, A. AbuAli, "A General Framework for E-Government: Definition Maturity Challenges, Opportunities, and Success", European Journal of Scientific Research, Vol. 39, No. 1, pp.: 29-42, 2010.

[2] NSN/CTO "Towards effective e-governance: The delivery of public services through local econtent, 2008, Global Summery report", http://www.cto.int/Portals/0/docs/research/towards-effective-egovernance/ Towards-Effective-eGovernance.pdf (accessed July, 2014).

[3] Siau, K., and Y. Long, "Using Social Development Lenses to Understand E-Government Development.", Journal of Global Information Management, Vol. 14, No. 1, 2006, pp.47-62.

[4] Lallan E, "eGovernment for Development, M-Government Definitions and Models", http://www.egov4dev.org/mgovdefn.htm (accesed June, 2014)

[5] Sulaiman A. Alateyah, Richard M Crowder and Gary B Wills, "An Exploratory study of proposed factors to Adopt e-government Services" International Journal of Advanced Computer Science and Applications(IJACSA), 4(11), 2013.

[6] Johan Hellström, "Mobile phones for good governance - challenges and way forward", http://www.w3.org/2008/10/MW4D-WS/papers/ hellstrom-gov.pdf, (accessed July, 2014).

[7] NSN/CTO "Towards effective e-governance: The delivery of public services through local econtent, 2008, Global Summery report", http://www.cto.int/Portals/0/docs/research/towards-effective-egovernance/ Towards-Effective-eGovernance.pdf (accessed July, 2014).

[8] Silvana Trimi AND Hong Sheng, "Trends in M-GOVERNMENT" , COMMUNICATIONS OF THE ACM May 2008/Vol. 51, No. 5.

[9] Nasim Qaisar and Hafiz Ghufran Ali khan, "E-Government Challenges in Public Sector:A case study of Pakistan", IJCSI International Journal of Computer Science Issues, Vol. 7, No. 5, 2010, pp. 310-317.

[10] Hiba M., Tamara A., Amer A, "E-government in Jordan", European Journal of Scientific Research, Vol. 35, No.2, 2009, pp.188-197.

[11] Johan Hellström, "Mobile phones for good governance - challenges and way forward", http://www.w3.org/2008/10/MW4D-WS/papers/ hellstrom-gov.pdf, (accessed July 2014).

[12] A. Farshid Ghyasi, "m-Government: Cases of Developing Countries", Mobile Government Lab. http://www.mgovlab.org (accessed July 12, 2014).

[13] Mohammed Alshehri and Steve Drew, "A Comprehensive Analysis of E-government services adoption in Saudi Arabia: Obstacles and Challenges " International Journal of Advanced Computer Science and Applications(IJACSA), 3(2), 2012.

[14] Raed Kanaan and Ghassan Kanaan, "The Failure of E-government in Jordan to Fulfill Potential" International Journal of Advanced Computer Science and Applications(IJACSA), 4(12), 2013.

[15] United Nations Public Administration Network, "Government survey 2012- United Nations Public Administration Programme", http://unpan1.un.org/intradoc/groups/public/documents/un/unpan048065 .pdf (accessed June 2014)

[16] Rinku Dixit "m-Government : Ruling the High - Tech Way", http://www.egovonline.net/articles/articledetails. asp?Title= m%E2%80%93Government-:-Ruling-the-High-%E2%80%93-Tech-Way&ArticalID= 2103&Type=MCONNECT, (accessed July 2014)

[17] Alateyah, Sulaiman, Crowder, Richard M. and Wills, Gary B. (2013) An exploratory study of proposed factors to adopt e-government services. *International Journal of Advanced Computer Science and Applications*, Vol. 4, No. 11, 2013, pp. 57-66.

[18] Efraim Turban, David King, Electronic Commerce 2012: Managerial and Social Networks Perspectives, USA:Pearson, 2012.

[19] David Villegas, Norman Bobroff, Ivan Rodero, Javier Delgado, Yanbin Liu, Aditya Devarakonda, Liana Fong, S. Masoud Sadjadi, Manish Parashar, "Cloud federation in a layered service model," Journal of Computer and System Sciences, Vol. 78, Issue 5, September 2012, pp. 1330-1344.

[20] Wei Liu, Feiyan Shi,Wei Du and Hongfeng Li "A Cost-Aware Resource Selection for Data intensive Applications in Cloud-oriented Data Centers" in the International Journal of Information Technology and Computer Science 2011.

[21] Perera, S.; Kumarasiri, R.; Kamburugamuva, S.; Fernando, S.; Weerawarana, S.; Fremantle, P., "Cloud Services Gateway: A Tool for Exposing Private Services to the Public Cloud with Fine-grained Control," Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International, 2012 , pp.: 2237 - 2246.

[22] Costanzo, A. ; Faro, A. ; Giordano, D. ; Spampinato, C., "Context aware services for mobile users: JQMobile vs Flash Builder implementations," Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on, pp.: 1185 - 1192, 2012.

[23] Rainer, B. ; Lederer, S. ; Muller, C. ; Timmerer, C., "A seamless Web integration of adaptive HTTP streaming," Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pp.: 1519 - 1523, 2012.

[24] Design and Implementation of Coldfusion-Based Web Application Firewall Feng Fangmei ; Changgeng Shao ; Dan Liu Computer Science & Service System (CSSS), 2012 International Conference on, pp.: 659 – 662, 2012.

[25] von Laszewski, G., Diaz, J., Fugang Wang, Fox, G.C., "Comparison of Multiple Cloud Frameworks," Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on Communication, Networking & Broadcasting ; Components, Circuits, Devices & Systems ; Computing & Processing Hardware/Software, 2012 , pp.: 734 - 741.

Energy Efficient Cluster-Based Intrusion Detection System for Wireless Sensor Networks

Manal Abdullah, Ebtesam Alsanee, Nada Alseheymi
Computer Science Department, Faculty of Computing and Information Technology FCIT,
King Abdul-Aziz University KAU,
Jeddah, Saudi Arabia

***Abstract*—Wireless sensor networks (WSNs) are network type where sensors are used to collect physical measurements. It has many application areas such as healthcare, weather monitoring and even military applications. Security in this kind of networks is a big concern especially in the applications that required confidentiality and privacy. Therefore, providing a WSN with an intrusion detection system is essential to protect its security from different types of intrusions, cyber-attacks and random faults. Clustering has proven its efficiency in prolong the node as well as the whole WSN lifetime. In this paper we have designed an Intrusion Detection (ID) system based on Stable Election Protocol (SEP) for clustered heterogeneous WSNs. The benefit of using SEP is that it is a heterogeneous-aware protocol to prolong the time interval before the death of the first node. KDD Cup'99 data set is used as the training data and test data. After normalizing our dataset, we trained the system to detect four types of attacks which are Probe, Dos, U2R and R2L, using 18 features out of the 42 features available in KDD Cup'99 dataset. The research used the K-nearest neighbour (KNN) classifier for anomaly detection. The experiments determine K = 5 for best classification and this reveals recognition rate of attacks as 75%. Results are compared with KNN classifier for anomaly detection without using a clustering algorithm.**

Keywords*—wireless sensor networks WSN; intrusion detection ID; clustering protocols; stable election protocol SEP; KDD cup'99; KKN

I. Introduction

Due to their easy and inexpensive deployment features, Wireless Sensor Networks (WSNs) are applied to various fields of science and technology. These applications include to gather information about human activities and behavior, such as healthcare, military surveillance and reconnaissance, highway traffic; to observe physical and environmental phenomena, such as ocean and wildlife, earthquake, pollution, wild fire, water quality; to monitor industrial sites, such as building safety, manufacturing machinery performance, and so on [1]. On the other hand, security in WSNs is an important issue, particularly if they have mission-critical jobs. For example, a confidential patient health record should not be unrestricted to third parties in a healthcare applications. Securing WSNs is critically important in military applications where security crack in the network would cause causalities of the friendly armies in a battlefield [1]. Security attacks against WSNs are categorized into two main branches: Active and Passive. In passive attacks, attackers are normally hidden and either tap the communication link to collect data; or destroy the functioning elements of the network. Passive attacks can be grouped into eavesdropping, node malfunctioning, node tampering/ destruction and traffic analysis types. In active attacks, an adversary actually affects the operations in the attacked network. This effect may be the objective of the attack and can be detected. Active attacks can be grouped into Denial-of-Service (DoS), jamming, hole attacks (black hole, wormhole, sinkhole, etc.), flooding and Sybil types [1].

Solutions to security attacks against wireless sensor networks involve many components such as prevention, detection and mitigation. First, we discuss the intrusion detection components. According to [1], detection means being aware of the attack that is present. So if an attacker manages to pass the measures taken by the 'prevention' step, then it means that there is a failure to defend against the attack. At this time, the security solution would immediately switch into the 'detection 'phase of the attack in progress and specifically identify the nodes that are being compromised. ID systems are used to monitor both user and system activities to analysis any abnormal activity patterns and recognize patterns of typical attacks. In WSN, sensor nodes use batteries as power supply so battery power is a significant resource for sensor devices. The sensor nodes can be installed in an extensive geographical space to observe physical phenomenon with adequate precision and dependability. After installed, the minor sensor nodes are usually unapproachable to the operator. Therefore, conservation of energy and energy efficient routing must be taken into account when choosing a clustering algorithm. Contribution in this paper is to build an intrusion detection system that combines three main features:

- Use an energy efficient cluster-based WSN that guarantee prolong the life time of the single sensor node and the whole network as well. SEP protocol works based on election of the node which have the highest energy within each cluster as a cluster head. This technique has proven to prolong the life time of the network.
- Use of KNN classifier that has the advantage of having simple classifier and reduce the computation of detecting the attacks. Reducing the computation is an important advantage toward saving the network energy in general.
- Use of KDD-NSL[2] dataset that has a specific feature of avoiding the redundant attributes by removing irrelevant and redundant features that are inter-

correlated. This technique helps to achieve high detection rate and accurate results.

The rest of the paper is organized as follows. Section 2 discusses the literature review and related works. In section 3, the proposed ID system is introduced. The experimental work is discussed in section 4 and finally in section 5, the paper is concluded.

II. Literature Review

In this section it is required to review the LEACH protocol as basic clustering protocol where it is used to compare the results. The research relies on three main parts which are the SEP cluster-based WSN, the ID system and the classification technique. The three parts are discussed in the following subsections then some related work are introduced.

A. LEACH Clustering protocol: advantages and problems

The core idea of LEACH protocol is to split the whole network into numerous clusters. The cluster head node is arbitrarily selected, the chance of every node to be selected as cluster head is equal, and energy consumption of the entire network is averaged. Thus, LEACH can extend network life-cycle. LEACH algorithm is cyclical; it provides a conception of rounds. Every round contains two states: cluster setup state and steady state. In setup state, it forms cluster in self-adaptive mode and in steady state, it transfers data. The selection of cluster head depends on decision made 0 or 1. If the number is less than a threshold, the node turns into a cluster head for the present round. The threshold is set as shown in formula (1) [3]:

$$T(n) = \begin{cases} \frac{p}{1-p*(r*mod1/p)} & if\ n\in G \\ 0 & else \end{cases} \qquad (1)$$

where P is the preferred percentage of cluster head (e.g. 4 or 5%), r is the present round, and G is the set of nodes that have not been cluster heads in the last 1/p rounds. Using this threshold, every node will be a cluster head at some point within 1/p rounds. Nodes that have been cluster heads cannot become cluster heads for a second rounds 1/(p-1). Each node has 1/p probability of becoming a cluster head in each round. At the end of every round, every normal node that is not a cluster head select the nearest cluster head and joins that cluster to transfer data. The cluster heads combine and compress the information and forward it to the base station, thus it extends the life span of main nodes. In this algorithm, the energy consumption will be assigned uniformly among all nodes and the non-head nodes are turning off as much as possible. LEACH assumes that all nodes are in range of wireless transmission of the base station which is not the case in many sensor deployments. 5% of the entire nodes play as cluster heads in each round. Time Division Multiple Access (TDMA) is deployed for better management and scheduling.

One problem in the traditional LEACH protocol is that the cluster head node is randomly selected [4]. After several rounds, the node with more remaining energy and the node with less remaining energy have same probability to be selected as cluster head. If the node that has less energy is chosen as cluster head, it will run out of energy and die rapidly, so that network's robustness will be affected and network lifetime will be short [5].

B. Stable Election Protocol SEP

The SEP (Stable Election Protocol) preserves a clustering hierarchy. SEP is an improvement over LEACH in the way that it took into account the heterogeneity of networks. In SEP, some of the high energy nodes are referred to as advanced nodes and the probability of advanced nodes to become CHs is more as compared to that of non-advanced nodes[5]. In SEP, the clusters are re-established in every "round". New cluster heads are selected in every round and as a result the load is well distributed and balanced among the nodes of the network. Furthermore every node transfers to the closest cluster head so as to divide the communication cost to the sink (which is tens of times greater than the processing and operation charge). Just the cluster head has to report to the sink and may consume a large amount of energy, but this happens periodically for every node. In SEP there is an ideal percentage (determined a priori) of nodes that has to become CH in every round, according to [5] we denote this ideal percentage as "Popt". When the nodes are homogeneous, that means all the nodes in the field have the same primary energy, the SEP protocol assurances that each one of them will become a cluster head exactly once each 1/Popt rounds. According to [5] 1/Popt is denoted as "epoch" of the clustered sensor network. On average, n × Popt nodes need become cluster heads per round per epoch where n is the whole number of nodes. Nodes that are chosen to be CH in the present round can no longer become CH in the same epoch. The probability of non-elected nodes belong to the group G to become a CH growths after every round in the same epoch. This maintains a stable number of CHs per round. The choice is made at the beginning of every round by every node S ∈ G independently where picking an arbitrary number between [0,1]. If the arbitrary number is less than a threshold T(S), then the node turn into a CH in the present round. The threshold is set as in equation (2) [5], where r is the present round number.

$$T(s) = \begin{cases} \frac{Popt}{1-Popt(r\ mod\ 1/Popt)} & if\ s\in G \\ 0 & otherwise \end{cases} \qquad (2)$$

C. KNN Classifier

Nearest neighbor rule is widely used in identifying the category of unknown data point on the basis of its nearest neighbor whose class is already known [6]. In KNN, the nearest neighbor is calculated on the basis of value of k that specifies how many nearest neighbors are to be considered to define class of a sample data point [7]. Success of the KNN classifier depends on the least distant between instance features, which are determined by its distance function such as the ordinal Euclidean distance. The Euclidean distance between points is defined by equation (3) [8]:

$$E(P,Q)= \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}$$

$$= \sum_{i=1}^{n}((p_i-q_i)^2) \qquad (3)$$

Where P = (p_1,p_2,p_3 ,…,p_n) and Q = (q_1,q_2,q_3 ,…,q_n)

D. *Intrusion Detection System*

Proposed ID system detects four types of attacks which are [9]:

- Denial of Service (DOS): Attacker tries to prevent legitimate users from using a service.
- Remote to Local (R2L): Attacker does not have an account on the victim machine, hence tries to gain access.
- User to Root (U2R): Attacker has local access to the victim machine and tries to gain super user privileges.
- Probe: Attacker tries to gain information about the target host.

It is important to note that the test data includes specific attack types not in the training data which make the task more realistic. The datasets have a total number of 24 training attack types, with extra 14 types in the test data only. The name and classifications of the training attack types are listed in table1.

TABLE I. ATTACK TYPES WHICH WILL BE DETECTED BY THE ID SYSTEM

Class	Known attack	Unknow attack
Probe	Ipsweep,nmap, portsweep,satan	Saint, scan
DoS	Back,land,Neptne,pod, smurf,teardrop	Apache2,processtable, udpstorm,mailbomb
U2R	Buffer_overflow,loadmodule, perl,rootkit	Xterm,ps,sqlattack
R2L	ftp_write,guess_passwd, imap,multihop,phf,spy, warezclient,warezmaster	Snmpgetattack,named, xlock,xsnoop,sendmail, httptunnel,worm, snmpguess

E. *Related Work*

Bharti et al (2010)[10] defined clustering as the best technique for intrusion detection, and k-mean clustering is one of the useful ID clustering technique because it gives efficient results in case of datasets. But sometimes k-mean clustering fails to give best result because of class dominance and no-class problems. The ID system is an effective approach to deal with the problems of networks using various neural network classifiers. Sapna et al (2011) [11] stated that network based intrusion detection are the best methods. IDS can be a piece of installed software or a physical appliance. The different types of attacks are normal, Probe attacks, u2R, Dos and R2l attacks. Attacks are generated randomly using a random function. The type of attack generated is classified to be a Probe, R2L, U2R or Dos attack [12].

Jianlinetal (2011) [13] worked on fuzzy clustering analysis. Fuzzy clustering is the most popular research currently. It is one of the most perfect and most widely used theories although the rear some drawbacks for classical algorithms. Aizhonget al (2010) [14] focused on pattern recognition as the best classifier selection to network ID and clustering based selection method. The multiple clusters are selected for a test sample. The purpose of selecting the multiple classifiers is to optimizing the pattern recognition. Ajitetal (2005) [15] explained Expectation-Maximization (EM) technique which used in point guesstimate. Given a set of noticeable variables X and unknown (latent) variables Z we want to estimate parameters q in a model. Sometimes the M-step is a constrained maximization, which means that there are constraints on legal solutions not encoded in the function itself. The method to arrange the set of objects into classes of similar (which are having same behavior) objects, is defined as clustering. Objects are being categorized into two categories, (1) Documents within a cluster should be similar (2) Documents from different clusters should be dissimilar.

III. PROPOSED ID SYSTEM FOR WSNS

The proposed ID system supposes that all nodes are equipped with sensor and radio system. This assumption enables all nodes to be eligible to be chosen as cluster head. Three steps of the methodology as follow: using the training data and its features, we train the system by clustering the four attacks to the cluster which representing the attacks. Another cluster will present the normal state in which there is no attack and all the detected intrusion is legal. Then it comes the role of SEP protocol which calculates the weighted election probabilities of each node to become CH according to the remaining energy in each node. The SEP protocol is shown in figure 1. Then the KNN classifier that is built with function in MATLAB with multiple values of K is used to find out the best detection rate as shown in figure 2. KNN works by choosing *k* cluster centers to coincide with *k* randomly chosen or *k* randomly defined points inside the hyper volume containing the pattern set. Then assign each pattern to the closest cluster center. The last step is to recompute the cluster centers using the current cluster memberships. If a convergence criterion is not met, move to step2 as shown in figure 2. Classic convergence criteria are used as no (or minimal) reassignment of patterns to new cluster midpoints, or minimal reduction in squared error.

IV. EXPERIMENT SETUP

The ID system for WSN is implemented using MATLAB. The network consists of 100 node distributed in area of 50*50 meter with all nodes start with same energy and are equipped with sensor and radio system as mentioned before. Many trails are done to determine some important parameters before running the experiment. First, we need to find out which data set will be used to train the system and detect attacks and also to test the system performance. Second, the features used for the best detection classification rate are determined. Finally, data inside the data set is normalized. Each step will be explained in details in the following subsections.

SEP Protocol Algorithm

1. Force each advanced node to be elected every sub-epoch of length (*1+a x m)/P /(1+a*) rounds
2. Probability of a normal node getting elected as cluster head is *P normal*

$$P\ normal = \frac{\mathrm{P}}{1 + a \times m}$$

$$T(i) = \begin{cases} \frac{P\ normal}{1-P\ normal\times\ (r\ mod\ \frac{1}{P\ normal})} & if\ i\ \epsilon\ G\ normal \\ 0 & otherwise \end{cases}$$

3. Probability of an advanced node getting elected as cluster-head is *P advanced*

$$P\ advanced = \frac{P}{1 + \mathrm{a} \times \mathrm{m}} \quad (1 + a)$$

$$T(i) = \begin{cases} \frac{P\ advanced}{1 - P\ advanecd\ \times\ (r\ mod\ \frac{1}{P\ advanced})} & if\ i\ \epsilon\ G\ advanced \\ 0 & otherwise \end{cases}$$

4. Average number of nodes elected per round = *nxP*

Fig. 1. SEP Protocol Algorithm.

Classification Algorithm

```
1. Data Feature selection
      Select the appropriate 19 features
2. Data pre-processing and normalization
   a. Select the nominal feature
   b. Calculate the probability using probability
      density function
         $Pr[a \le X \le b] = \int_{b}^{a} fX\,(x)dx$
   c. Replace the nominal with numerical value
3. Input : training data set , testing data set , group set , K-value
4. KNN classification
      Class = knnclassify(Sample, Training, Group, k)
5. Compute the detection rate
      Detection rate =
      # normal connections  misclassified as attack /
      total number of normal connections
```

Fig. 2. KNN Classifier Algorithm

A. KDD CUP'99 Intrusion Detection Data Set.

KDD cup '99 is the most widely used data set in network intrusion detection and evaluation [9]. MIT Lincoln Labs prepared and managed the 1998' DARPA Intrusion Detection Evaluation Program to survey and evaluate researches in intrusion detection. A typical set of data which includes a large diversity of intrusions simulated in a military network situation was provided. The 1999 KDD intrusion detection contest uses a version of this dataset. KDD training dataset consists of about 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type [16]. Attack types fall into four main categories: User to Root; Remote to Local; Denial of Service; and Probe.

B. KDD '99 Features

Features shown in table 2 are grouped into four groups as follows: **Basic Features:** can be derived from packet headers without inspecting the payload. Basic features are the first six features listed in table 2. **Content Features:** Domain knowledge is used to assess the payload of the TCP packets. This contains features such as the number of failed login attempts. **Time-based Traffic Features:** These features are designed to capture properties that mature over a 2 second time-based window. One example of such a feature could be the number of connections to the same host over the 2 second interval; **Host-based Traffic Features:** Utilize a historical window estimated over the number of connections – in this case 100 – as a substitute of time. Host based features are then designed to assess attacks, which distance intervals longer than 2 seconds [17].

TABLE II. LIST OF ATTRIBUTES

Total Attribute NSL_KDD		
Protocol_type	Service	Src_byte
Wrong_fragment	Flag	Num_failed_logins
Logged_in	Root_shell	count
Serror_rate	Srv_serror_rate	Rerror_rate
Same_srv_rate	Diff_srv_rate	Dst_host_srv_count
Dst_host_serror_rate	class	Srv_rerror_rate

C. Data Preprocessing and Normalization

Most classifiers in IDS range, particularly artificial intelligence like KNN, handle only numeric dataset and ignore the symbolic features. Therefore, in this section we present a simple version algorithm that transfers nominal features in KDD dataset into numeric value. Furthermore, after transformation, we normalize the dataset scale for all features into [0,1] to avoid dominance and feature impact.[18].

Step 1: Data Set Transformation:

There are three futures that have character values (protocol type, Service, Flag), which must be converted to numeric values by using Probability Density Function PDF as given by equation (4):

$$\Pr[a \le X \le b] = \int_a^b fx(x)\,dx \qquad (4)$$

Step 2: Data Set Normalization:

Normalization is essential to enhance the performance of intrusion detection system. Normalization phase must be applied on all features on KDD dataset. This paper has used MinMax function given by equation (5). To normalize numeric values to range between MinX and MaxX that are the minimum and maximum values for feature X, first [MinX, MaxX] is converted to new range [New MinX, New MaxX], According to equation (5) each value of V in the original range is converted to a new value.

$$new_v = \frac{v - \min x}{\max x - \min x} \qquad (5)$$

V. RESULTS AND DISCUSSION

The experiment is starting with creating a wireless sensor network using MATLAB, and clustering it using SEP protocol. At first, the energy for each node is calculated and based on calculated energy, we choose the cluster head which of course the nodes with the highest energy according to the SEP protocol. Secondly, the unlabeled patterns of nodes are grouped into clusters based on the distance between the cluster heads and nodes. The nodes join the cluster with closest cluster head. This minimizes the communication energy between the nodes and their cluster head and lead to preserve WSN energy and prolong the lifetime of WSN as a result. As we can see in figures 3 and 4, in each round we cluster the nodes and define a cluster head according to the sensor with the highest remaining energy.

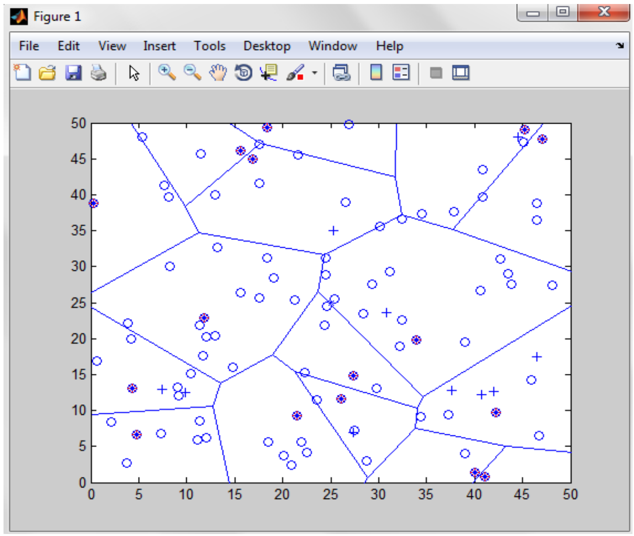


Fig. 3. Clustering 100 nodes using SEP protocol

For IDS, KNN classifier algorithm over KDD99' dataset is used to determine the optimum value of parameter k that reveals the best detection rate as shown in table 3. The experimental results are based on the standard evaluation metric for intrusion detection which is the detection rate.

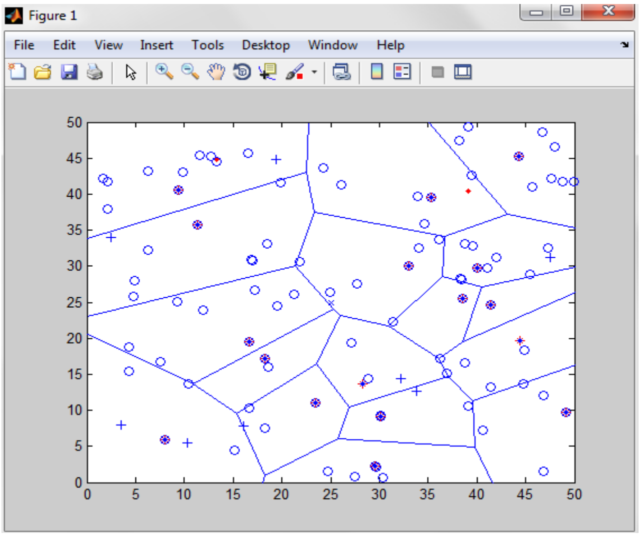


Fig. 4. Assigning new cluster head when the cluster head die

TABLE III. EXPERIMENTAL RESULT

K	DETECTION RATE
1	20.8333%
2	20.8333%
5	75%
10	70.8333%
20	50%
25	33. .8333%

The above table illustrates that, as the value of k increases, the detection rate will be increased until reach the optimal k-value with the highest detecting rate. Then, as the k-value increases, the detection rate will be decreased considerably.

From the table we conclude that the optimum value of k is 5 which results in the highest detection rate of 75%.

Comparing the results of the purposed experiment with other work which is not clustered before classification. The experimental results provide the highest detection rate up to 75%. Figure 5 shows the comparison results. The results also show that the KNN classification without clustering is working better in terms of recognition rate where k-value is less than 5. Although, with k = 5 or greater the KNN classifier with clustering provides the highest recognition rates.

The percentage of recognition rate is decreased with k-value increased for non clustered KNN. This percentage is decreased with increasing k-value for the clustered KNN.

VI. CONCLUSIONS

Intrusion Detection Systems are important tool to detect different types of attacks in WSN which help to monitor the activities and violations in WSN. It's important to consider the energy of the WSN during designing an intrusion detection system. In this paper we have designed an IDS for detecting four types of attacks which are Probe, DoS, U2R and R2L. We have focused on designing energy efficient IDS that preserve the energy of the WSN and prolong the lifetime of the nodes by using the SEP protocol which gives the best results comparing to non clustered network protocols. KDD CUP99' data set has been used for the intrusion detection to give more precise results. The system used KNN classification algorithm to determine the k-value that gives the maximum percentage recognition rate. Then SEP protocol is used for electing cluster head. The system can detect the intrusions with detection percentage rate of 75% at k =5.

As a future work we will consider to use different classification methods to compare with KNN classifier so that we can decide the best classification that works perfectly with the SEP protocol and to gain the maximum detection rate with the longest lifetime for the WSN.

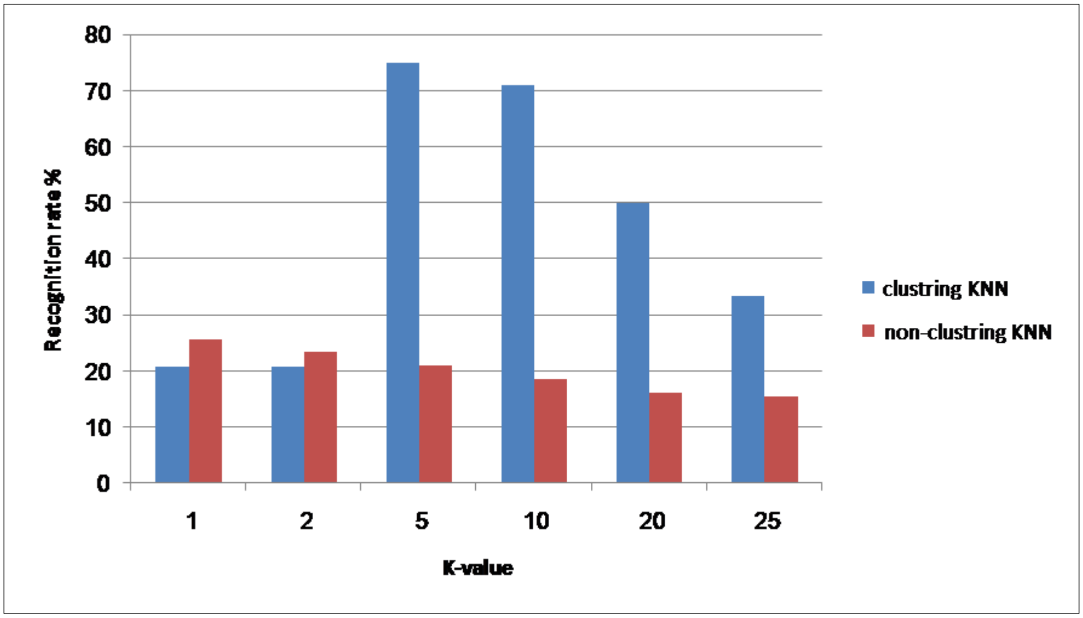


Fig. 5. KNN classification of clustered node compared with KNN classification with non-clustering node

REFERENCES

[1] Ismail Butun, Salvatore D. Morgera, and Ravi Sankar, A Survey of Intrusion Detection Systems inWireless Sensor Networks.

[2] A Detailed Analysis of the KDD CUP 99 Data SetTavallaee, Mahbod; Bagheri, Ebrahim; Lu, Wei; Ghorbani, Ali-A

[3] Qing Bian, Yan Zhang,"Research on Clustering Routing Algorithms in Wireless Sensor Networks," in 2010 International Conference on Intelligent Computation Technology and Automation.

[4] Jianguo SHAN, Lei DONG, Xiaozhong LIAO, Liwei SHAO, Zhigang GAO, Yang GAO Research on Improved LEACH Protocol of Wireless Sensor Networks.

[5] GeorgiosSmaragdakis, Ibrahim Matta, AzerBestavros, "SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks".

[6] Uvenir, H. A. &Akkus, A. "KNearest Neighbor Classification on Feature Projections". – ResearchGate

[7] Cover, T. M. &Hart,P. E. "Nearest Neighbor Pattern Classification", IEEE Trans. Inform. Theory, Vol. IT-13, pp 21-27

[8] Deokar, C. (2009). "Weighted K-Nearest Neighbor Algorithms (Solving the Curse of Dimensionality Problem)".

[9] Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I. (2005, October). Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets.

[10] Kusumbharti, SanyamShukla&Shweta Jain "Intrusion Detection using unsupervised learning".

[11] Sapna S. Kaushik, Dr. Prof.P.R.Deshmukh "Detection of Attacks in an Intrusion Detection System"Amravati India in 2011.

[12] Vladimir Golovko, PavelKachurka, LeanidVaitsekhovich "neural Network Ensembles for IntrusionDetection" Brest State Technical University in 2007.

[13] PengShanguo; Wang Xiwu; ZhongQigen; , "The study of EM algorithm based on forward sampling,"Electronics, Communications and Control (ICECC), 2011 International Conference on , vol., no., pp.4597-4600, 9-11 Sept. 2011 doi: 10.1109/ICECC.2011.6067693

[14] Maria Colmenares& Olaf WolkenHauer, "An Introduction into Fuzzy Clustering",http://www.csc.umist.ac.uk/computing/clustering.htm, July 1998, last update 03 July,2000

[15] http://home.deib.polim.it/matteucc/Clustering/tutorial_html/cmeans.html

[16] Hettich , S., & Bay, S. D. (1999, October 28). KDD Cup 1999 Data.Tavallaee , M., Bagheri, E., Lu, W., &Ghorbani, A. A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. Proceedings of the IEEE Symposium on computational Intelligence in Security and Defense application (CISDA 2009).

[17] Sathya, S. S., Ramani, R. G., &Sivaselvi, K. (2011). Discriminant analysis based feature selection in kdd intrusion dataset. International Journal of Computer Applications, 31(11).

[18] Salem, M. (2013, April 4). Preprocessing dataset in IDS. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/41129-preprocessing-dataset-in-ids.

[19] Ibrahim, L. M., Basheer, D. T., &Mahmod, M. S. (2013). A COMPARISON STUDY FOR INTRUSION DATABASE (KDD99, NSL-KDD) BASED ON SELF ORGANIZATION MAP (SOM).

A Study of Multi-Signal Monitoring System Establishment for Hemodynamic Energy Detection

YeonSoo Shin[1], DuckHee Lee[1]
[1]Korea Artificial Organ Center,
College of Medicine, Korea University.
Anam-Dong 5ga, Seonbuk-Gu,
Seoul 136-701, Korea

ChiBum Ahn[2]
[2]Department of Mechanical & Biomedical Engineering,
Engineering of College, Kangwon University
Kangwondaehak-gil, Chuncheon-si,
Gangwon-do, 200-701, Korea

SeungJoon Song[3]
[3]Department of Convergence Biomedical Engineering
Daelim University College
Imgoklo29 Dongan-gu
Anyang-si Gueonggi-do
431-715, Korea

Kyung Sun[1, 4]
[4]Department of Thoracic and Cardiovascular Surgery,
College of Medicine, Korea University
Anam-Dong 5ga, Seonbuk-Gu,
Seoul 136-701, Korea

***Abstract*—Deaths due to cardiovascular diseases are increasing worldwide, and multi-signal monitoring systems to diagnose such diseases are under development. However, only a few researches are underway for devices that monitor hemodynamic energy, which is a marker for pulsatile flow generated by the contraction and relaxation of the heart. Therefore, this study aimed to integrate multiple monitoring devices into a single device, while also incorporating hemodynamic energy monitoring. Blood pressure and flow were measured with two channels each, while electrocardiogram (ECG), photoplethysmography (PPG), and temperature were measured with one channel each. All signals were processed at hardware level, and then converted into analog voltage. The seven signals were then converted into digital signals with a data acquisition board (DAQ). The software was developed with Labview™ (National Instruments, U.S.A) to form a graphic user interface (GUI) on a tablet computer (ACER, U.S.A) through USB 2.0, to allow for monitoring and analysis of the signals obtained. Development of this system successfully formed a multi-signal monitoring system that integrates multiple signals into one device. Future directions include development of cardiovascular diagnosis algorithm and evaluation of the system via preclinical animal experiments.**

Keywords*—Cardiovascular disease; hemodynamic energy; monitoring system; Electrocardiogram; Photoplethysmography; Pressure; Flow

I. INTRODUCTION

According to the World Health Organization (WHO), the leading cause of death worldwide in 2011 was cardiovascular disease, which was responsible for 1.7 million deaths. WHO also predicted that two million people are expected to die from the same cause in 2015. Devices for diagnosing cardiovascular disease are under active development worldwide, and are differentiated into several types of sensors and devices. The heart maintains a pulsatile flow by its contractions and relaxations [1-3], and multiple researchers have analyzed the unique properties of pulsatile flow and compared them with those of nonpulsatile flow [4-11]. Multiple advantages of pulsatile flow during cardiovascular bypass have been reported, such as its similarities to the natural beating of the heart, as well as keeping the blood vessels open [12]. Also, a study utilizing the concept of hemodynamic energy characteristic of pulsatile flow suggests that pulsatile flow has greater hemodynamic energy, and has advantage in perfusion [13]. In 2002, Nitzan proposed a signal-processing algorithm that calculates the pulse transit time (PTT) from the R-wave and the photoplethysmography (PPG) recorded from the earlobes, fingers, and toes. The study reported the correlation between PTT and blood pressure in average adults [14]. Development of a system that integrates the conventional biological signals and hemodynamic energy will result in a system that can be used for both diagnosis of diseases as well as efficacies of treatments. However, such integrated system is currently difficult to find. Therefore, this study aimed to develop a multi-signal monitoring system that integrates the conventional signals including ECG, PPG, and temperature, as well as hemodynamic energy as Overall System.

II. OVERALL SYSTEM CONFIGURATION

The system has a total of seven channels. Pressure and flow are measured in two channels each, while ECG, PPG, and temperature are measured with one channel each. Figure 1 shows the overall configuration of the system, where multiple signals are relayed to a single input, processed by analog-to-digital converting (ADC), which then can be saved and analyzed.

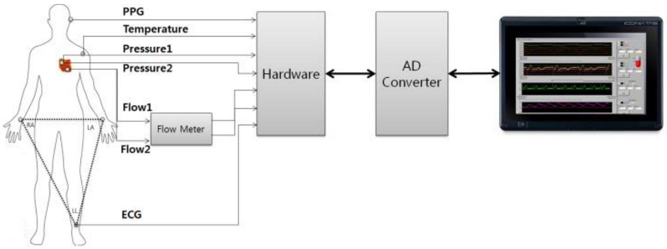


Fig. 1. System Blockdiagram

III. Hardware System Structure

Pressure is measured by deflection method with the ST AM100 amplifier (Senstech, Korea), a dynamic strain amplifier operating on AC 220 V. A low pass filter at 10 Hz was used, and an internal filter was used to amplify the gain by 200 times. The PS9030VY pressure sensor was used (Sensortechnics, Germany). Output is shown in voltage between 0~10 V, and can measure 0~4000 ustrain. Flow was measured with the TS410 flow meter (Transonic System, U.S.A), a tubing flow meter operating on AC 220 V. Flow detected with the ME-11PXL probe (Transonic System, U.S.A) was expressed in voltage between 0~1 V. Pulses were measured with EP520 (Laxtha, Korea), a photosensor that can be attached to the earlobes. The applied voltage of the EP520 was 5 V, and the LED Driving current control voltage was 0.38 V. The LED Driving voltage required approximately three seconds of response time before showing stable waveforms. To measure temperature, K-type thermocouple sensors were used with a K-type thermocouple amplifier chip, AD595 (Analog Device, Inc, U.S.A), which operates on 5 V. Change in temperature was detected as a voltage change of 10 mV per temperature change of 1 ˚C. ECG was measured with the Einthoven's method with standard leads I, II, and III. Snap electrodes (Laxtha, Korea) were used in conjunction with AD620 amplifier (Analog Device Inc, U.S.A) operating on 5 V. ECG signals were differentially amplified, and noise was eliminated with band-pass filtering and 60 Hz notch filtering by LM324 amplifier (Analog Device, Inc, U.S.A). The developed ECG hardware was processed into printed circuit boards (PCB), using Pads 9.4 software (Mentor Graphics, U.S.A). Because ECG, temperature, and pulse signals all operate on 5 V, the power input was supplied from an ADC board, NI USB-6009 (National Instruments, U.S.A). Because all signals are collected in analog form, they must be converted into digital signals before being processed by a computer. Therefore, NI USB-6009, which has 14 bit resolution, sample rate of 48 kS/s and accuracy range of 1.53 mV was used to convert the signals from analog to digital. These signals were transferred to a tablet PC (ACER, U.S.A) via USB 2.0 connection and measured using the LabView™ software.

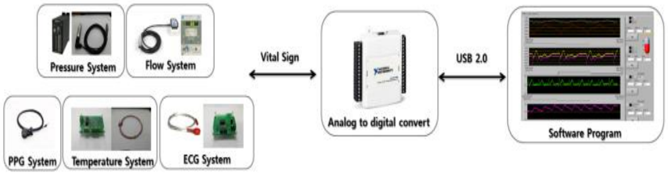


Fig. 2. Integrated System Configuration

IV. Concept Introduction of Hemodynamic Energy

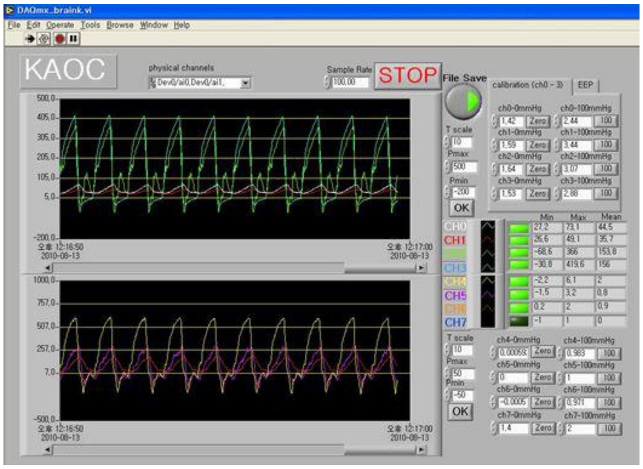
Shepard's equation, which calculates the hemodynamic energy in a pulsatile flow using pressure and flow rate was applied in this system [15].

$$\text{EEP(mm Hg)} = (\int fpdt)/(\int fdt) \qquad (1)$$

Energy equivalent pressure (EEP) is the amount of energy required to move a unit volume of blood across a unit distance, expressed in pressure units. f represents the flow rate (L/min), while p represents arterial pressure (mm Hg). EEP is determined by first calculating hemodynamic power curve, which is an integration of the product of flow rate and pressure during one cardiac cycle, and then dividing it by the pump flow-rate curve, which is an integration of the flow rate over time.

$$\text{SHE (ergs/}cm^3\text{)} = 1{,}332[((\int fpdt)\,/(fdt)) - MAP] \qquad (2)$$

Surplus hemodynamic energy (SHE) is the excess energy after subtracting the amount necessary for basic circulation [16]. Therefore, SHE is expected to be 0 in nonpulsatile flows where EEP equals mean arterial pressure (MAP), but positive in pulsatile flows where EEP is greater than MAP. Based on the hemodynamic energy equations, software was developed with the Labview™ software to monitor hemodynamic energy EEP and SHE in pulsatile flow generated by a Korean pulsatile biventricular pump, TPLS (Twin Pulse Life Support) [17].



physical channels: Dev0/ai0,Dev0/ai1,

Sample Rate: 100.00

T scale: 10
Pmax: 500
Pmin: -200

ch0-0mmHg: 1.42 Zero — ch0-100mmHg: 2.44 100
ch1-0mmHg: 1.59 Zero — ch1-100mmHg: 3.44 100
ch2-0mmHg: 1.64 Zero — ch2-100mmHg: 3.07 100
ch3-0mmHg: 1.53 Zero — ch3-100mmHg: 2.88 100

	Min	Max	Mean
CH0	27.2	73.1	44.5
CH1	26.6	49.1	35.7
CH2	-68.6	366	153.8
CH3	-30.8	419.6	156
CH4	-2.2	6.1	2
CH5	-1.5	3.2	0.8
CH6	0.2	2	0.9
CH7	-1	1	0

T scale: 10
Pmax: 50
Pmin: -50

ch4-0mmHg: 0.00059 Zero — ch4-100mmHg: 0.983 100
ch5-0mmHg: 0 Zero — ch5-100mmHg: 1 100
ch6-0mmHg: -0.0005 Zero — ch6-100mmHg: 0.971 100
ch7-0mmHg: 1.4 Zero — ch7-100mmHg: 2 100

Fig. 3. Measurement of Hemodynamic energy

V. Software for Vital Sign Measurement

One cycle of ECG is comprised of P, Q, R, S, and T-wave, out of which the QRS complex occurs during ventricular polarization and typically reside in higher frequency area. The P-wave is caused by atrial contractions, while the T-wave is caused by ventricular contractions, and both typically reside in the lower frequencies. The QRS complex plays a fundamental and critical role in ECG analysis; once the analysis of the QRS complex is complete, ST waves, PR waves, and RR intervals can also be analyzed. Therefore, ECG waveforms were displayed using secondary band-pass filters, 60 Hz notch filters, and FIR filtering at a software level to distinguish the QRS complex.

The pulse rate signal may be affected by the patient's respiration or skin condition, which affects the current and amplitude of the signals. Therefore, the patient must remain stationary to obtain stable waveforms. A baseline equation can be applied to calculate the pulse rate by comparing the occurrences of peak values.

$$\text{Pulse Rate} = \frac{60(sec)}{Duration\ Between\ Peaks} \qquad (3)$$

Because the AD595 chip displays 1˚C of temperature change per 10 mV of voltage change, equation 4 was applied to the software to show the temperature.

$$\text{Temperature} = \text{Measured voltage (mV) x } 100 \qquad (4)$$

As shown in figure 4, ECG and pulse signals were graphed in Labview™ in real time, and temperature was displayed by the thermometer icon.

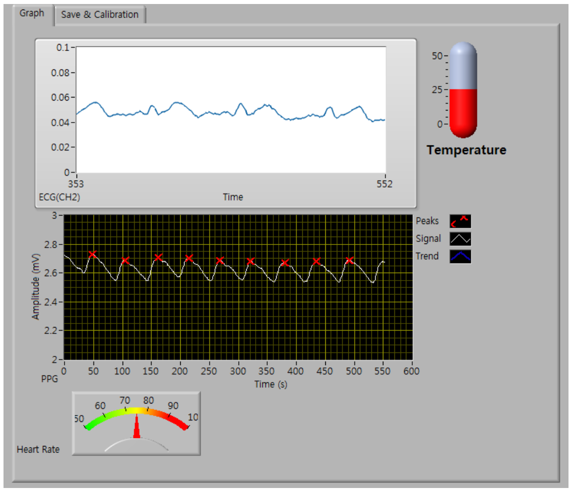


Fig. 4. Measurement of Complex Vital Signal

VI. Result of Integrated Hardware and Software

Hardware and software for multiple vital signals were integrated into a device and simplified. All seven analog signals can be converted into digital signals and transferred to the DAQ board and a portable tablet computer via USB 2.0, which can then be analyzed with the Labview™ software, as seen in figure 5. By combining multiple softwares, changes in individual signals can be graphed in real time onto a single screen. Sampling rate and duration can be adjusted as needed, and the raw data can be saved in text or excel format, which can then be analyzed statistically.

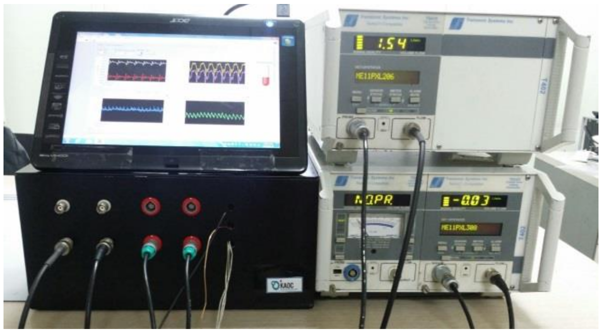


Fig. 5. Result of Integrated System

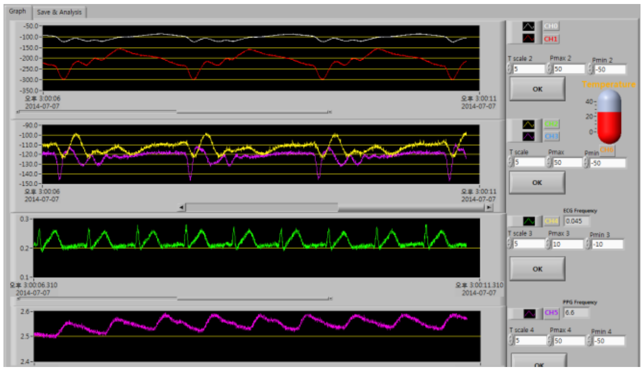


Fig. 6. Software User Interface

VII. Conclusion

In this study, a system was developed to combine measurements of signals necessary for calculation of hemodynamic energy, as well as other monitoring parameters, including pressure, flow, ECG, PPG, and temperature, for the purpose of the development of an integrated multi-signal monitoring system. Although the analog measurement of temperature change showed high precision at 10 mV = 1 ˚C, the converted digital value showed lower precision in comparison. This can be attributed to the noise from the power input, generated upon connection to the USB-6009 DAQ board. Efforts are being made to reduce the effect of this noise on measurement precision. Future directions include replacing cable connection with wireless connection in order to simplify the structure of the system. Furthermore, hemodynamic energy markers will be integrated with ECG and PPG to develop a vascular monitoring system. Once performance and safety of this newly developed system prove to be suitable for clinical use after *in vitro* and *in vivo* experiments, it may be possible to utilize this system in both monitoring and diagnosis for circulatory conditions.

Acknowledgment

This research was funded by the Ministry of Health & Welfare, Republic of Korea (grant number: A121157)

References

[1] Marroudis C, "To pulse or not to pulse", Ann Thorac Surg 1978. 25: 259-71.

[2] Hichey PR, Buchley MJ, Philbin DM, "Pulsatile and nonpulsatile cardiopulmonary bypass review of a counterproductive controversy", Ann thorac Surg 1983. 36: 720-37.

[3] Hornick P, Taylor K, "Pulsatile and nonpulsatile perfusion: the continuing controversy", J Cardiothorac Vasc Anesth 1997. 11: 310-5

[4] Drakos SG, Charitos CE, Ntalianis A, et al, "Comparison of pulsatile with nonpulsatile mechanical support in a porcine model of profound cardiogenic shock", ASAIO J 2005. 51: 26–9.

[5] Undar A, Rosenberg G, Myers JL, "Major factors in the controversy of pulsatile versus nonpulsatile flow during acute and chronic cardiac support", ASAIO J 2005. 51: 173–5.

[6] Brandes H, Albes JM, Conzelmann A, Wehrman M, Ziemer G, "Comparison of pulsatile and nonpulsatile perfusion of the lung in an extracorporeal large animal model", Eur Surg Res 2002. 34:321-9

[7] Undar A, Masai T, Frazier OH, Fraser CD. Pulsatile and nonpulsatile flow can be quantified in terms of energy equivalent pressure during cardiopulmonary bypass for direct comparisons, ASAIO J 1999. 610-4.

[8] Lodge AJ, Undar A, Daggett CW, Runge Tm, Calhoon JH, Ungerleiger RM, "Regional blood flow during cardiopulmonary bypass and after

circulatory arrest in an infant model", Ann Thorac Surg 1997. 63:1243-50.

[9] Ciadullo Rc, Schaff HV, Flaherty JT, Donahoo JS, Gott VL, "Comparison of regional myocardial blood flow and metabolism distal to critical coronary stenosis in the fibrillating heart during alternate periods of pulsatile and nonpulsatile perfusion", J Thorac Cardiovasc Surg 1978. 75:193-205.

[10] Maddoux G, Pappas G, Jenkins M, et al. "Effect of pulsatile and nonpulsatile flow during cardiopulmonary bypass on left ventricular ejection fraction early after aortocoronary bypass surgery", Am J Cardiol 1976. 37:1000-6.

[11] Murkin JM, Marzke JS, Buchan AM, Bentley C, Wong CJ. "A randomized study of the Influence of perfusion technique and PH management strategy in 316 patients undergoing coronary artery bypass Surgery", Mortality and cardiovascular mobidity. J Thhorac Cardiovasc Surg 1955. 110: 340-8.

[12] Shepard RB, Kirklin JW, "Relation of pulsatile flow to oxygen consumption and other variables during CPB", J Thorac Cardiovasc Surg 1969. 58: 694-702.

[13] Ji B, Undar A: "An evaluation of the benefits of pulsatile versus nonpulsatile perfusion during cardiopulmonary bypass procedures in pediatric and adult cardiac patients", ASAIO J 2006. 52: 357-61.

[14] M. Nitzan, B. Khanokhm and Y. Slovik, "The difference in pulse trasit time to the toe and finger measured by photoplethysmography", Physial. Meas, 2002. 23:85-93,

[15] Shepard RB, Simpson DC, Sharp JF, "Energy Equivalent Pressure", Arch Surg 1966. 96:93: 790-40

[16] Undar A, Zapanta CM, Reibson JD, et al. "Precise quantification of pressure flow waveforms of a pulsatile ventricular assist device", ASAIO J. 2005. 51(1): 56-9.

[17] Jung Joo Lee, Choon Hak Lim, Ho Sung Son, et al. "In vitro evaluation of the performance of Korean pulsatile ECLS(T-PLS) using precise quantification of pressure- flow waveforms", ASAIO J. 2005. 51(5): 604-8.

Numerical Simulation on Damage Mode Evolution in Composite Laminate

Jean-Luc Rebière
Laboratoire d'Acoustique de l'Université du Maine (UMR CNRS 6613)
Université du Maine
Av. Olivier Messiaen, 72085 Le Mans Cedex 9, France

***Abstract*—The present work follows numerous numerical simulation on the stress field analysis in a cracked cross-ply laminate. These results lead us to elaborate an energy criterion. This criterion is based on the computation of the partial strain energy release rate associated with all the three damage types: transverse cracking, longitudinal cracking and delamination. The related criterion, linear fracture based approach, is used to predict and describe the initiation of the different damage mechanisms. With this approach the influence of the nature of the material constituent on the damage mechanism is computed. We also give an assessment of the strain energy release rates associated with each damage mode. This criterion checked on glass-epoxy and graphite-epoxy composite materials will now be used in future research on new bio-based composite in the laboratory.**

Keywords—numerical simulation; damage mechanism; transverse cracking;longitudinal cracking; delamination; criterion

I. Introduction

During the last years, composite laminates are widely used in many structural applications thanks to their high strength to weight ratio. Their durability still needs to be carefully assessed. So, it is desirable to be able to rely on a suitable damage-growth criterion. Most of these type of composite laminates are composed with glass or graphite long fibers and polymer matrix. Experimentally, in this composite cross-ply laminate subjected to monotonic or fatigue tensile loading, the damage mechanisms sequence is as follows. The first observed damage is generally transverse cracking, causes by interlaminar stress concentration at the crack tips. High interlaminar stress levels may entail the debonding of layers at the interface of the plies with different orientations and/or they may also cause matrix cracking between fibres in the longitudinal layers. The composite structures damaged by incipient delamination or longitudinal cracking must be repaired. The main objective of this work is to evaluate the influence of the nature of the material system on the initiation and evolution of transverse crack damage, longitudinal crack damage and delamination.

For the study of transverse cracking damage, two particulars states are generally observed: the initiation of the first damage called "*first ply failure (FPF)*" and the limiting state of this damage, named "*characteristic damage state (CDS)*", when no more transverse crack can be created. The second type of damage observed is either longitudinal cracking or delamination. This second type of damage generally depends on following parameters: the laminate geometry, thicknesses of the *0°* or *90°* layers, the nature of the fibre (graphite, glass...)/matrix constituents, the loading history and the manufacturing cycle. For example, in [1] the initiation and growth of delamination was observed in a thick composite laminate. Ply separation is provoked by the increase of interlanimar normal and shearing stresses. In case of thin composite laminates, the damage mode succession can be different. In [1, 2] the second damage observed is longitudinal cracking which follows transverse cracking. In this case, local delamination appears between *0°* and *90°* layers. In every case, all the different damage modes causes fibre breaking in the *0°* layers. All fibre breaks entail, named "*splitting*", which appears just before the ultimate failure of the composite laminate.

In the literature, analytical and numerical approaches have been proposed for modelling the strain/stress relationship during damage growth mechanism. Some models are more suitable to describe the initiation of the first damage mode. They mainly rely on some stress field distribution and a relationship between loading and crack density is usually proposed. The simplest models like Steif [3], so called "*shear lag analyses*", usually involve elementary assumptions using the displacement and stress distributions. Other type of models like variational approaches [4] use the principle of minimum complementary energy [5-7]. Other types of studies use the finite element method [8]. We can also find some models based on phenomenological approaches [9], self-consistent analyses [10] or approaches based on specific aspects of the cracks [11-14]. The longitudinal cracking is equivalent to transverse cracking damage, but arise in the longitudinal plies (*0°* layers). Longitudinal cracks are not always continuous [15]. Generally longitudinal cracking occur very late in the laminate life. So, the investigation of longitudinal cracking is often ignored by many models in the literature. Relying on experimental observations, with this approach we suppose that the longitudinal cracks are continuous and that they span the whole length of the damage studied specimen. For the study of delaminated damage, a delaminated surface with a triangular shape at the crossing of longitudinal and transverse cracks was used for estimate the initiation of the interface debonding between orthogonal plies. In this article, we also study the initiation of delaminated surface with triangular shape.

In the literature, several approaches have been proposed to investigate the evolution of the different types of damage in composite cross-ply laminates and several kinds of criteria have been proposed [16], among them maximum stress based approaches. We can also find other types of criteria [15] rely on the energy release rates associated with each type of damage.

Our interest in damage mechanism evolution and succession lead us to bring out the respective contributions of the transverse crack damage, longitudinal crack damage or delamination damage mechanism development which can be found in the strain energy release rate for the two materials constituents. The strain energy release rate is expressed through an appropriate semi-analytical model [15] and decomposed into individual components related to damage mechanisms. Opposing to the simplest models of the literature, which only take into account only the normal stress in the loading direction, the general proposed model used here allows a thorough investigation of the strain energy release rate to be achieved. The present study is only restricted to damage growth in cross ply laminates. After numerous numerical simulations, it could be established that the influence of a given component of the stress field on some of the damage mechanisms can be neglected (some shear components). Concerning the initiation of transverse cracking, this kind of assumption gives good results. However, when the evolutions of transverse cracking or the transition to other types of damage are of interest, the simplest damage models cannot be used. In these cases, a more accurate description of the stress field is necessary like the proposed model. In the present article, we further develop the above analysis by first providing the numerical values of the parts of the decomposition of the strain energy release rate associated to each damage type.

II. Model

The studied specimen is a *[0$_m$, 90$_n$]$_s$* composite cross-ply laminate as represented in Fig. 1.

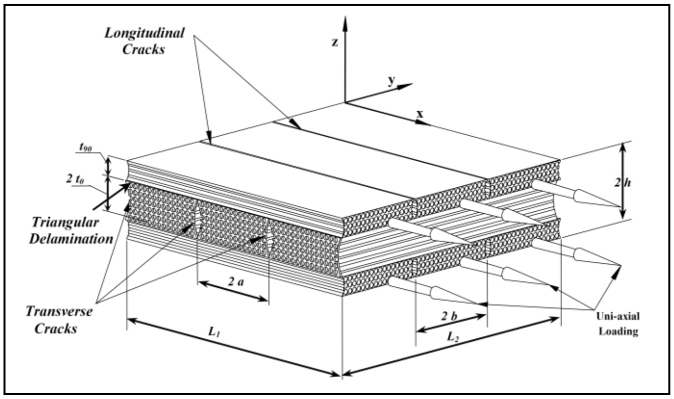


Fig. 1. Laminate damaged by transverse cracks, longitudinal cracks and triangular delamination

The geometric parameters used to describe the laminate architecture is the λ □coefficient *(λ = t_0/t_{90} where t_0 is the 0° ply thickness and t_{90} is the 90° ply thickness).* With the proposed approach longitudinal cracks are taken continuous by hypothesis [15]. Based on linear elastic fracture mechanics, the estimated values of the strain energy release rates are computed in a *"pre-damaged"* laminate, a method used in several damage models in the literature. Thus, there are already *"pre-existing"* transverse and longitudinal cracks and/or triangular delamination at the crack type.

Then, the evolution of transverse cracking damage is described in the following way. We consider a laminate with a periodic array of transverse cracks in the inner *90°* layer. Damage initiation of matrix cracking occurs when the spacing between two consecutive cracks is very important *(infinite)*. For studying longitudinal cracking a similar method can be used. The laminate is supposed to be *"pre-cracked"*. The initiation of the longitudinal damage is obtained for an infinite value of the damage parameter *(ratio of the spacing between two consecutive cracks to the central damaged layer thickness).*

The accepted assumptions for the crack geometries in the *0°* and *90°* layers of the laminate are as follows. The cracks surfaces are supposed to have a rectangular plane geometry. Each crack extends over the whole thickness and the whole width of the *90°* damaged ply. Similar assumptions are used for the longitudinal cracks in the two damaged *0°* layers. With these assumptions, it is sufficient to study the only *"unit damaged cell"*. This ''*unit damaged cell*'' thus lies between two consecutive transverse and longitudinal cracks. Triangular delaminated areas are located at the cross of transverse and longitudinal cracks at the interface of the *0°/90°* plies. For the initiation of delamination the size in the *x* direction *(d_x)* and in the *y* direction *(d_y)* is supposed to be equal and called *d_l*. In [15] the summary of the method is exposed to estimate the stress field distribution in the cracked laminate. In the damaged laminate, the stress field in the two layers has the following form:

$$\sigma_{ij}^{T(k)} = \sigma_{ij}^{0(k)} + \sigma_{ij}^{P(k)} \qquad (1)$$

In the undamaged laminate loaded in the *x* direction, the layers experience a uniform plane stress state $\sigma_{ij}^{0(k)}$ obtained by the laminate plate theory *(where k is the ply index, k = 0°, 90°).* The orthogonal cracks induce stress perturbations in the *0°* and *90°* layers which are denoted $\sigma_{ij}^{T(k)}$ [15].

III. Strain Energy Release Rate

The laminate is supposed to be *"pre-damaged"* by *"pre-existing"* transverse and longitudinal cracks. The size of the unit damaged cell depends on the transverse and longitudinal damage levels in the *90°* and *0°* layers. The strain energy release rate *G* associated with the initiation and development of ply cracking for a given stress state is defined by the following expression:

$$G = \frac{d}{dA}\widetilde{U}_d(\sigma, A) \qquad (2)$$

With:

$$\widetilde{U_d} = N.M.U_{cel} \qquad (3)$$

where $\widetilde{U}_d$ is the strain energy of the whole laminate and *A* is the cracked area. Let L_1 denote the laminate length in the *x* direction and L_2 its width in the *y* direction. The strain energy in the damaged unit cell is denoted by U_{cel}. N ($N = L_1/2\bar{a}t_{90}$) is the number of transverse cracks and M ($M = L_2/2\bar{b}t_{90}$) is the number of longitudinal cracks. Dimensionless quantities are defined by, $\bar{a} = a/t_{90}$, $\bar{b} = b/t_{90}$.

The crack area is $A = L_1 L_2 \left(1/\bar{a} + \lambda/\bar{b}\right)$. The strain energy release rates associated with transverse and longitudinal cracking are denoted G_{FT} and G_{FL} respectively. The transverse *(resp. longitudinal)* cracking growth is characterized by the increase of the transverse (*resp. longitudinal)* crack surface initiated in the *90° (resp. 0°)* layers. All details are given in [15]. Then:

$$G_{FT} = \frac{d\tilde{U}_d}{dA} = \frac{d\tilde{U}_d}{d\bar{a}} \frac{d\bar{a}}{dA} \qquad (4)$$

$$G_{FL} = \frac{d\tilde{U}_d}{dA} = \frac{d\tilde{U}_d}{d\bar{b}} \frac{d\bar{b}}{dA}$$

The strain energy release rates associated with delamination is G_{del}, we get :

$$G_{del} = \frac{d\tilde{U}_d}{d\bar{a_l}} \frac{d\bar{a_l}}{dA_d} \qquad (5)$$

For the analysis of the delamination evolution, only isosceles triangular geometries of the debonded area are studied. In [1], the authors have experimentally observed similar triangular areas for the initiation of delamination whereas during its propagation, damage can grow along the longitudinal and/or transverse cracks.

IV. Results

All the numerical simulations are carried out for a prescribed uni-axial loading of *150MPa*. The glass/epoxy and *T300-934* graphite/epoxy material system are studied in the following numerical computations of the strain energy release rate.

The energetical criterion proposed is elastic linear fracture based approach. The parameters involved in the study are the constraining parameter, the thickness of the two *0°* and *90°* layers and the material constituent system. All the partial strain energy release rates, associated with the initiation of transverse cracking, longitudinal cracking or delamination are normalized by the critical strain energy release rate. For this *8* ply cross-ply laminate, transverse cracking is thus the first observed damage. We can also observe that the strain energy release rates, G_{FT}/G_{cr}, G_{FL}/G_{cr} and G_{del}/G_{cr} have similar variation laws for all the computed materials.

In Fig. 2-4 the results of only two types of materials are exposed. We can observe with all the G_{FT}/G_{cr}, G_{FL}/G_{cr} and G_{del}/G_{cr} values that in graphite/epoxy laminate, it is more difficult to initiate the three damage modes than in the glass/epoxy laminate. All the strain energy release rates are decreasing functions of the constraining parameter λ. For instance, in a *8* ply laminate, when the value of the constraining parameter λ is increased, the thickness of the *0°* plies becomes greater. In this case, the fibers in the *0°* plies carry most of the tensile loading and the initiation of the three different damage modes is delayed. Although no experimental data are reported on Fig. *2-4*, the results of the numerical simulations confirm the main points: the proposed approach agrees with experimental data for the influence of the material constituent and the initiation of transverse cracking is the first damage mode. It also predicts the readiness to initiate the three types of damage in the case of a *8* ply laminate containing a thick *90°* layer. Experimentally, it was observed that, after a certain loading level, the number of longitudinal cracks can become more and more important [1]. At this loading level, delamination can appear along the longitudinal cracks; moreover, some small induced transverse cracks can appear along the longitudinal cracks with delaminated areas.

TABLE I. Mechanical Properites and Ply Thickness for Glass Epoxy And T300/934 Graphite Epoxy Systems

	Glass/Epoxy	T300/934 Graphite / Epoxy
E_{LT}(GPa)	*41.7*	*144.8*
$E_{TT'}$(GPa)	*13*	*11.7*
G_{LT}(GPa)	*3.4*	*6.5*
$G_{TT'}$(GPa)	*4.58*	*3.5*
ν_{LT}	*0.3*	*0.3*
ν_{TT}	*0.42*	*0.54*
Ply thickness (mm)	*0.203*	*0.132*

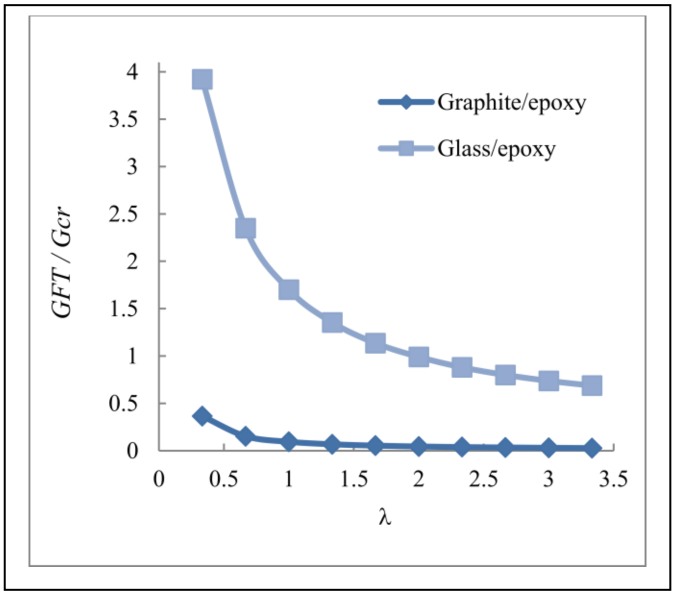


Fig. 2. Strain energy release rate G_{FT} /G_{cr} in laminate *T300/934* graphite epoxy and glass epoxy versus constraining parameter λ.

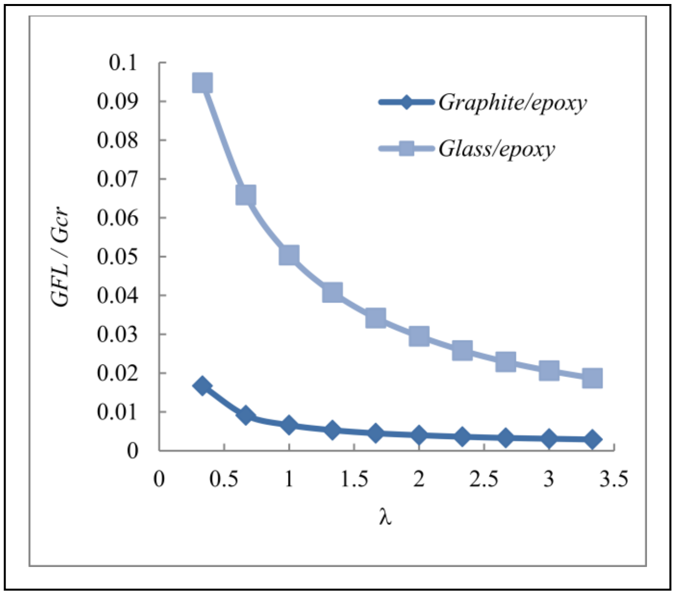


Fig. 3. Strain energy release rate G_{FL}/G_{cr} in laminate *T300/934* graphite epoxy and glass epoxy versus constraining parameter λ.

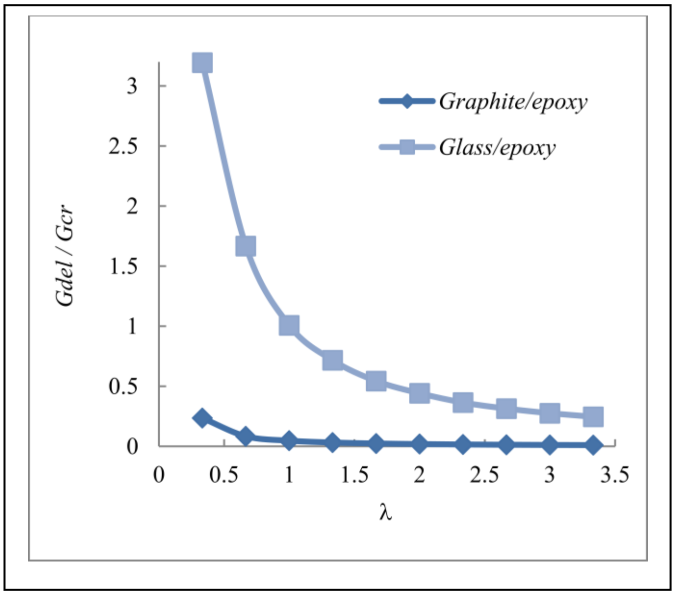


Fig. 4. Strain energy release rate G_{del}/G_{cr} in laminate *T300/934* graphite epoxy and glass epoxy versus constraining parameter λ.

V. Conclusion

The energetical criterion is proposed to predict and describe the initiation of the different damage mechanisms occurring in symmetrical composite cross-ply laminates under uniaxial loading. The influence of the material constituent is exposed. All the numerical simulation that the variation of the strain energy release rates as function of the constraining parameter λ is similar for the two types of material constituent system. The difference lies in the numerical values of the strain energy release rates. This result shows that it is more difficult to initiate damage in graphite/epoxy laminate. The strain energy release rates are always computed in a *“pre damaged”* state, with *"pre-existing"* transverse and longitudinal cracks. The curves displayed confirm that transverse cracking first occurs in the *90°* layers and longitudinal cracking arrive at the end of the laminate life for important value of the crack density. In a *8* ply laminate, when the value of the constraining parameter is increased, the thickness of the *0°* plies becomes greater and carry most of the loading and the initiation of the three damage is delayed.

References

[1] E. Urwald, ’’Influence de la géométrie et de la stratification sur l'endommagement par fatigue de plaques composites carbone/époxyde’’, Ph.D. dissertation. Université de Poitiers; 1992.

[2] R.-D. Jamison, K. Schulte, K.-L. Reifsnider, W.-W. Stinchcomb, ‘’Characterization and analysis of damage mechanisms in tension-tension fatigue of Graphite/Epoxy laminates. Effects of Defects in Composites Materials’’, ASTM STP 836, American Society for Testing and Materials, 1984, pp. 21-55.

[3] P.S. Steif ‘’Parabolic shear-lag analyses of a $[0/90]_s$ laminate. In: Transverse crack growth and associated stiffness reduction during the fatigue of a simple crossply laminate’’, eds.: Ogin S.L., Smith P.A., Beaumont P.W.R. Report CUED/C/MATS/TR 105, Cambridge University, 1984, pp. 40-41.

[4] J. Lemaître, J.-L. Chaboche, ’’Mechanics of Solid Materials’’, Cambridge University Press, 1994.

[5] V.V. Vasil'ev, A.A. Duchenco, ’’Analysis of the tensile deformation of glass-reinforced plastics’’, Translated from Mekhanica Polimerov.; vol 1, 1970, pp.144-147.

[6] A. Hosoi, H. Kawada, ‘’Stress analysis of carbon fiber reinforced plastics, containing transverse cracks, considering free-edge effect and residual thermal stress’’, Mater Sci Eng A, vol. 4498, No 1-2, 2008, pp. 69-75. Doi:10.1016/j.msea.2007.11.153

[7] J.-L. Rebière,’’Modélisation du champ des contraintes créé par des fissures de fatigue dans un composite stratifié carbone/polymère’’, Ph.D. dissertation, Université de Poitiers, 1992.

[8] C.T. Herakovich, J. Aboudi, S. W. Lee, E. A. Strauss, ’’Damage in composite laminates: effects of transverse cracks’’, Mech Mater., vol. 7, No 2, 1988, pp. 91-107.

[9] G. Lubineau, P. Ladevèze, D. Violeau, “Durability of cfrp laminates under thermomechanical loading: a micro-meso damage model”, Compos Sci Technol, vol. 66, No 7-8, 2006, pp. 983–92.

[10] S. Adali, R. K. Markins, ’’Effect of transverse matrix cracks on the frequencies of unsymmetrical, cross-ply laminates’’, J the Franklin Institute.; vol. 329, No 4, 1992, pp. 655-665.

[11] Barbero E. J., Cortes D. H., ’’ A Mechanistic Model for Transverse Damage Initiation, Evolution, and Stiffness Reduction in Laminated Composites. ’’ Composites Part B; vol. 41/2, 2010, pp.124-132.

[12] Yokozeki T., Aoki T., Ishikawa T., ’’Consecutive matrix cracking in contiguous plies of composite laminates’’, Inter J Solids and Struct.; vol. 42, 2005, pp.2785–2802.

[13] M.M. Moure, S. Sanchez-Saez, E. Barbero, E.J. Barbero, ’’Analysis of damage localization in composite laminates using a discrete damage mode’’, Composites Part B: Engineering, In Press, Accepted Manuscript, Available online 23 May 2014.

[14] G. Sadeghi, H. Hosseini-Toudeshky, B. Mohammadi, ’’An investigation of matrix cracking damage evolution in composite laminates – Development of an advanced numerical tool’’, Compos Struct, vol. 108, 2014, pp. 937-950.

[15] J.-L. Rebière, D. Gamby, ’’A criterion for modelling initiation and propagation of matrix cracking and delamination in cross-ply laminates’’, Compos Sci Technol, vol. 64, No 13-14, 2004, pp. 2239-2250. Doi: 10.1016/j.compscitech.2004.03.008

[16] N.-V. Akshantala, R. A. Talreja, ’’micromechanics based model for predicting fatigue life of composite laminate’’, Mater Sci Eng A, vol. 285, No 1-2, 2000, pp. 303-313.

A Study on Relationship between Modularity and Diffusion Dynamics in Networks from Spectral Analysis Perspective

Kiyotaka Ide, Akira Namatame
Department of Computer Science
National Defense Academy of Japan
Yokosuka, Japan

Loganathan Ponnambalam , Fu Xiuju,
Rick Siow Mong Goh
Computing Science, Institute of High Performance Computing, A*STAR
Singapore, Singapore

***Abstract*—Modular structure is a typical structure that is observed in most of real networks. Diffusion dynamics in network is getting much attention because of dramatic increasing of the data flows via the www. The diffusion dynamics in network have been well analysed as probabilistic process, but the proposed frameworks still shows the difference from the real observations. In this paper, we analysed spectral properties of the networks and diffusion dynamics. Especially, we focus on studying the relationship between modularity and diffusion dynamics. Our analysis as well as simulation results show that the relative influences from the non-largest eigenvalues and the corresponding eigenvectors increase when modularity of network increases. These results have the implication that, although network dynamics have been often analysed with the approximation manner utilizing only the largest eigenvalue, the consideration of the other eigenvalues is necessary for the analysis of the network dynamics on real networks. We also investigated Node-level Eigenvalue Influence Index (NEII) which can quantify the relative influence from each eigenvalues on each node. This investigation indicates that the influence from each eigenvalue is confined within the modular structures in the network. These findings should be made consideration by researchers interested in diffusion dynamics analysis on real networks for deeper analysis.**

***Keywords*—complex network; modularity; diffusion; SIS model; graph spectra; eigenvalue and eigenvector**

I. Introduction

Diffusion phenomena ongoing on today’s well-networked society can be often analyzed as probabilistic diffusion processes in complex networks. And, because the social systems have been keeping glowing up more dynamically and complexly, studying the probabilistic diffusion process in complex networks has been gathering a lot of attentions. Also, the probabilistic diffusion process has been well-applied to various fields, such as information spreads, dissemination of new products, computer virus spread, and epidemics.

Modular structure is a ubiquitous characteristic found in many real networks [e.g. 1]. Identifying hidden modular structure in real networks has been studied by many researchers in the scope of social network analysis [e.g., 2–8]. For instance, Newman’s community detecting algorithm using betweenness centrality [4] is the pioneer work that triggered the development of community detecting algorithms. But, their algorithm have two problems; 1) the number of communities is needed to be estimated in advance, even if the user might want to know the most optimized number of partitioned communities as the result of optimization. 2) computation time is too long. The first problem was solved by the introduction of modularity Q that is an index that can quantify the accuracy of the partitioning [5]. Then the users can identify the most proper partitioning instead of deciding the number of communities in the network in advance. The second problem was solved using the greedy algorithm that the completely separated graphs connect to be higher modularity Q [6]. After that, many researches have proposed various methods and especially apply them to the social network analysis. In addition to that, diffusion properties on modular networks have been studies by many researchers. For instance, Gao et al. [9] investigated the relationship between the number of modules and the properties of percolation on the randomly modularized network, which results that modularized networks are more destructible than a single independent network. In terms of the probabilistic diffusion dynamics on modular networks, it is reported that resonance-like phenomena can be seen in the probabilistic diffusion processes on modular networks [10]. Also, Saumuell-Mendiola [11] studied the SIS diffusion model on a coupled network which consists of two independent networks combined each other, and they reported that epidemics are prone to arise on the interconnected networks comparing to the single independent networks. Furthermore, Sahneh et al. [12] and Wang et al. [13] proposed the theories that the epidemic threshold for a coupled network can be calculated by the adjacency matrix of the coupled network. Also, for the Susceptible-Infected-Recovered (SIR) diffusion model, analyzing the interconnection between communities is also important [14, 15].

Analyzing the diffusion process in networks as Susceptible-Infected-Susceptible (SIS) model has been one of the conventional approaches that can be well-applied to the study of information diffusion as well as epidemics [16-21]. In the SIS model, each node in the network is probable to be assigned to two states, susceptible state and infected state. In the epidemics context, the susceptible state nodes represent the healthy individuals that are probable to be infected. On the other hand, the infected state nodes represent patients that are probable to influence its neighbors at a certain infection rate in the epidemics context. Then, the infected state nodes are possible to automatically return to the susceptible state at a certain recovery rate. One of the important insights from the studies on the SIS model in complex network is that critical phenomenon, which the steady-state fraction of infected nodes suddenly jumping up at the certain condition, can be observed. Then, finding the critical point (i.e. threshold or tipping point) has been attracting many researchers' interests, and many analytical approaches as well as simulative approaches have been proposed so far [19,22-24]. For instance, Kephart and White [22] firstly analyzed SIS model and formulated time-evolution of the steady-state fraction of infected nodes in homogeneous network. Wang et al. [23] proposed a more advanced analytical framework for general networks from the spectral point of view. They reported that the critical point can be approximately calculated by the inverse of the largest eigenvalue of the adjacency matrix of the underlying network. Mieghem et al. [24] developed "the N-intertwined mean field approximation model". And our work is based on their analytical frameworks. In addition, Mieghem et al. [25] also proposed another approach from spectral analysis perspective.

Although many theories have been proposed, as mentioned above, Pastor-Satorras and Vespignani [26] reported that, in scale-free network, the critical phenomena cannot be observed from their analysis of the empirical survey results of computer virus spread. They also found that the infections are localized within very small areas before the critical point. Furthermore, they reported that the steady-state fraction of infection saturates to very lower than analytically expected. These empirical facts differ from the analytical results introduced above. Recently, this contradiction is elucidated by the localization-delocalization phenomenon reported by Goltsev et al. [27]. The In their paper, the inverse participation ratio (IPR) is applied to the network diffusion analysis from the spectral point of view, and concludes that hubs, edges with large weight and dense sub-graphs in networks are probable to be the centers of localization.

Because of the limitation of computational performance and data availability, previous analysis on the diffusion model had tried to find out better approximation approaches to figure out the dynamics. These analytical results based on the conventional linear algebra based analysis indicate that the largest eigenvalue and the principal eigenvector of the adjacency matrix can approximate the diffusion dynamics on general networks [23, 24]. However, according to the results of our analysis from spectral point of view and numerical simulations, the accuracy of this approximation method varies depending on the modularity of the network. In our previous works [28], we quantified the accuracy of the approximation method utilizing only the largest eigenvalue of the adjacency matrix and found that the accuracy is low in some real networks. In this paper, we insist that the accuracy of the approximation method depends on the modularity of networks, which verifies numerical simulations.

Also, our proposed measure, Node-level Eigenvector Influence Index (NEII) [28] which can quantify and visualize the influences from an arbitrary eigenvalue to each node, captures the insight that only considering the largest eigenvalue cannot appropriately approximate the dynamics on the highly modularized networks.

In the second section, we review some existing analytical frameworks and provide our proposed analytical frameworks that will be fundamentals for the later discussions. In the third section, we examine the spectral properties of some artificial complex networks and real networks, which indicates that the importance of non-largest eigenvalue for the diffusion properties on the modular networks and verify the hypothesis in the previous section by numerical simulations. In the fourth section, we develop the parameterized modular network formation algorithm to verify the hypothesis. In the fifth section, we introduce the Node-level Eigenvector Influence Index (NEII). Then the investigation of real modular networks conducted in the sixth section. Finally, we conclude this paper in the seventh section.

II. Analysis of Probabilistic Diffusion on Networks

A. N-Intertwined mean field approximation model

Mieghem et al. [24] developed the N-intertwined mean field approximation model, then an important results of their analysis is the following Markov differential equation,

$$\begin{aligned}\frac{d\mathbf{V}(t)}{dt} &= \beta\mathbf{A}\mathbf{V}(t) - \mathrm{diag}\big(v_i(t)\big)(\beta\mathbf{A}\mathbf{V}(t) + \delta\mathbf{e}) \\ &= (\beta\mathbf{A} - \delta\boldsymbol{I})\mathbf{V}(t) - \beta\mathrm{diag}\big(v_i(t)\big)\mathbf{A}\mathbf{V}(t),\end{aligned} \tag{1}$$

where $v_i(t)$ denotes the probability that the node i is infected at time t, β is infection rate, δ is recovery rate, $\mathbf{V}(t) = \big(v_1(t), v_2(t), v_3(t), \cdots, v_N(t)\big)^T$, $\mathbf{e}$ is the all-one vector, and $\mathrm{diag}\big(v_i(t)\big)$ is the diagonal matrix where the diagonal elements consists of $v_1(t), v_2(t), v_3(t), \cdots, v_N(t)$. According to the comparison results with the numerical simulation results in small networks, the accuracy of this model is good enough except the region around threshold. In the studies of Susceptible-Infected-Susceptible (SIS) model [16-21], researchers have been making efforts to identify the threshold τ_c of the effective infection ratio $\tau \equiv \frac{\beta}{\delta}$. Then, several approaches have been taken to analytically calculate the threshold [22-26]. One of the most prominent achievements is that the threshold can be derived by the inverse of the largest eigenvalue of the adjacency matrix as follows,

$$\tau_c = \frac{1}{\lambda_1(\mathbf{A})}, \tag{2}$$

where $\lambda_1(\mathbf{A})$ denotes the largest eigenvalue of the adjacency matrix $\mathbf{A}$.

B. Spectral analysis

To solve the differential equation (1), we assumes that the fraction of infection on each node $v_i(t)$ is sufficiently small and ignoring the term, the equation (1) can be solved as the expression (3) using the eigenvalue decomposition,

$$\mathbf{V}(t) = \mathbf{U}\mathrm{diag}(\exp((\beta\lambda_k - \delta))t)\mathbf{U}^{\mathrm{T}}\mathbf{V}(0), \qquad (3)$$

where $\lambda_k(\mathbf{A})$ is *k*th eigenvalue of the adjacency matrix **A** and **U** denotes the orthonormal matrix which the *k*th column consists of the eigenvector of the *k*th eigenvalue. Then, the equation (3) can be rewritten as,

$$\mathbf{V}(t) = \sum\nolimits_k \exp\big((\beta\lambda_k(\mathbf{A}) - \delta)t\big)\mathbf{u}_k\mathbf{u}_k^{\mathrm{T}}\,\mathbf{V}(0), \qquad (4)$$

where $\mathbf{u}_k$ is eigenvector of the *k*th eigenvalue of the adjacency matrix **A**.

Assuming that the initial infection is randomly assigned to each node *i* at the probability $v_i(0) = 1/N$, the probability of infection on the node *i* at time *t* can be obtained as below,

$$\begin{aligned} v_i(t) = \frac{1}{N}\big(&\exp\big((\beta\lambda_1 - \delta)t\big)u_{1i}\|\mathbf{u}_1\| \\ &+ \exp\big((\beta\lambda_2 - \delta)t\big)u_{2i}\|\mathbf{u}_2\| + \cdots \\ &+ \exp\big((\beta\lambda_n - \delta)t\big)u_{ni}\|\mathbf{u}_n\|\big), \end{aligned} \qquad (5)$$

where the norm $\|\mathbf{u}_k\|$ stands for the sum of all elements of the eigenvector corresponding *k*th eigenvalue, that is $\|\mathbf{u}_k\| = u_{k1} + u_{k2} + u_{k3} + \cdots + u_{kn}$. Then, each term in the formula (5) implies that the influence from the *k*th eigenvalue λ_k toward a given node *i* is governed by the product of the *i*th component u_{ik} and the norm $\|\mathbf{u}_k\|$. Furthermore, the fraction of infected nodes over the whole network $\mathrm{y}(t)$ can be calculated by taking the average of $v_i(t)$ as follows,

$$\begin{aligned} \mathrm{y}(t) &= \frac{1}{N}\sum_{i=1}^{N} v_i(t) \\ &= \frac{1}{N^2}\sum_{k=1}^{N} \exp\big((\beta\boldsymbol{\lambda}_k - \delta)t\big)\|\mathbf{u}_k\|^2. \end{aligned} \qquad (6)$$

III. Spectral Property and Diffusion

A. Influence from the Non-Largest Eigenvalues

In the previous literatures, accuracy of the approximation method only utilizing the largest eigenvalue have not been discussed extensively and believed that is applicable to any types of networks. However, our analytical framework from the spectral point of view shows that influences from not only the largest eigenvalue of the adjacency matrix but also the other non-largest eigenvalues are important to express diffusion processes more accurately, which is validated by numerical simulation.

Then our investigation of the real networks shows that the modular networks with high modularity tend to show the property that the influences from the non-largest eigenvalues and the corresponding eigenvectors are significant.

In our previous work [28], we investigated the values of $\|\mathbf{u}_k\|^2$ in equation (6) in several artificial complex networks and real networks. The equation (6) can be expanded,

$$\begin{aligned} \mathrm{y}(t) = \frac{\exp\big((\beta\boldsymbol{\lambda}_1 - \delta)t\big)\|\mathbf{u}_1\|^2}{N^2}\Bigg(&1 \\ &+ \exp(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\frac{\|\mathbf{u}_2\|^2}{\|\mathbf{u}_1\|^{\mathbf{2}}} + \cdots \\ &+ \exp(\boldsymbol{\lambda}_n - \boldsymbol{\lambda}_1)\frac{\|\mathbf{u}_n\|^2}{\|\mathbf{u}_1\|^2}\Bigg). \end{aligned} \qquad (7)$$

As the equation (7) indicates, when the absolute value of $|\lambda_1 - \lambda_{k,\{k\neq 1\}}|$ is large and the dominance index ρ, which we defined as $\|\mathbf{u}_k\|^2/\|\mathbf{u}_1\|^2$, is small, $\mathrm{y}(t)$ can be significantly governed by the largest eigenvalue and its corresponding eigenvector. In contrast, when $|\lambda_1 - \lambda_{k,\{k\neq 1\}}|$ is small and the ρ is large, the term including the *k*th non-largest eigenvalue should be considered, that is, the approximation method only using the largest eigenvalue and corresponding eigenvectors, such as equation (2), is not applicable in this case.

Billen et al. [29] show that, as the number of triangles (i.e. clusters or three cycles) in a network increases, the spectrum of the network is positively skewed. This insight indicates that the value of $|\lambda_1 - \lambda_{k,\{k\neq 1\}}|$ increases when the clustering coefficient of the network increases, which implies $\mathrm{y}(t)$ tends to be governed by the largest eigenvalue when the clustering coefficient of the network is large. However, several studies on real data analysis and numerical simulation results [20] that scale-free network which has comparably larger clustering coefficient shows small steady-state fraction of infections, $\mathrm{y}(\infty)$, which indicates that consideration of the absolute value of $|\lambda_1 - \lambda_{k,\{k\neq 1\}}|$ is not important, and consideration of the value of ρ is more inevitable to measure the importance from the eigenvalues in each term of the equation (7).

Then, we investigated distribution of ρ in some artificial complex networks and real networks. Figure 1 shows the comparison of the distribution of ρ among the several networks, such as Barabasi-Albert scale-free network (BA), Erdos-Renyi random network (RND), random regular network (RR), Co-authorship Network of Network Scientists (CNNS) [30, 31] and UK members of parliament on Twitter network (UKMPTN) [32, 33]. European road network (EuroRoad) [22, 25], dolphin (Dolphin) network [31, 34], Email network (Email) [31, 35], and Jazz musicians' collaborating network (JazzNet) [36, 37]. Table 1 provides with the detailed network information including the optimal modularity Q that is an index to quantify the goodness of the partitioning and explained in the next section.

As can be seen in the figure, in RND and RR, the relative influence from the largest eigenvalue is apparently dominant and the relative influences from the non-largest eigenvalues are almost negligible. However, in the real networks, the relative importance from the non-largest eigenvalues increases. Especially CNNS and EuroRoad which show apparently high modularity and are apparently influenced from the non-largest eigenvalue. Especially, the 4th eigenvalue for CNNS and the 2nd eigenvalue for EuroRoad, is more dominant than those of their largest eigenvalues. This fact implies that the non-largest eigenvalues are more influential in the networks with the higher optimal modularity Q and the approximation only utilizing the largest eigenvalue and primary eigenvector is not appropriate to analyse these networks.

TABLE I. BASIC INFORMATION FOR THE INVESTIGATED NETWORKS

Network	N	Average Degree	Clustering Coefficient	The Largest Eigenvalue	Optimal # of Community	Optimal Modularity Q
BA	500	3.99	0.112	12.76	14	0.4911
RND	500	4	0.002	5.19	13	0.4901
RR	500	4	0.005	4	15	0.5103
CNNS	379	4.82	0.741	10.38	19	0.8386
UKMPTN	419	9.09	0.281	13.83	5	0.6270
Dolphin	62	5.13	0.259	7.19	4	0.4955
Email	1133	9.62	0.398	20.75	12	0.5070
EuroRoad	1039	2.51	0.0189	4.01	24	0.8649
JazzNet	198	27.70	0.617	40.03	4	0.4389

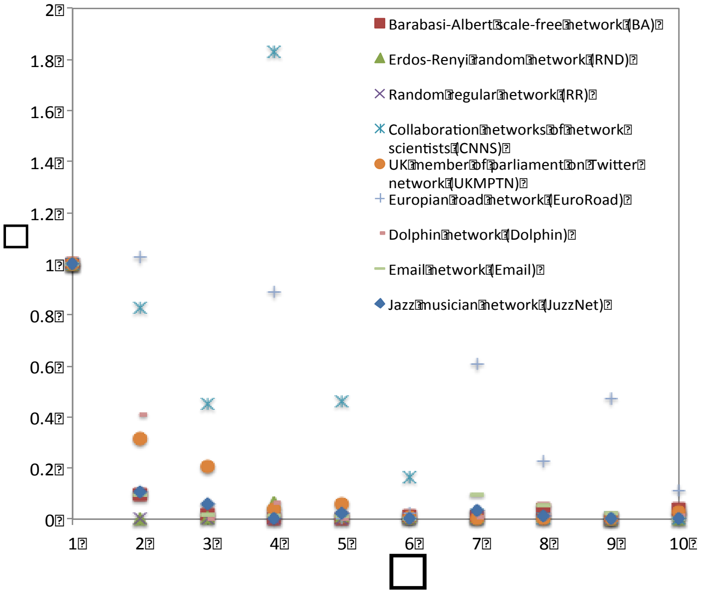


Fig. 1. Comparison of distribution of ρ, which we define at the dominance index of the principal vector, of Barabasi-Albert scale-free network (BA), Erdos-Renyi random network (RND), random regular network (RR), Co-authorship Network of Network Scientists (CNNS), UK member of parliament on Twitter network (UKMPTN), Euro road network (EuroRoad), dolphin (Dolphin) network, Email network (Email), and Jazz musicians' collaboration network (JazzNet).

B. Verification Simulation

Based on the results in the previous section, we hypothesize that, when we analyze the diffusion dynamics in real networks, we must consider not only the largest eigenvalue and the principal eigenvector of the adjacency matrix, but also the other eigenvalues and their corresponding eigenvectors.

As indicated in formula (2), the critical point is approximately calculated as the inverse of the largest eigenvalue of the underlying network's adjacency matrix. To verify if this approximation method, utilizing only the largest eigenvalue, is appropriate for every network, we simply compare analytically derived approximated thresholds $\tau_{\mathrm{c,AD}}$ with numerically calculated thresholds $\tau_{\mathrm{c,Sim}}$ in the networks introduced in the previous section. In this series of numerical simulations, we change the effective infection ratio by 0.001 (recovery rate δ is a constant = 1), and the fraction of infected nodes at 100 time-steps is assumed to equal the steady-state fraction of infected nodes, y_∞. At the constant effective infection rate, the simulations repeated 100 trials and the obtained results were averaged. 2% of the nodes were randomly selected as initial infected nodes in each trial. In Figure 2, blue plots display the evolution of y_∞ as the function of τ normalized by $\tau_{\mathrm{c,AD}}$ of each network. If the difference between $\tau_{\mathrm{c,AD}}$ and $\tau_{\mathrm{c,Sim}}$ is minimal, the blue plots begin to increase around 1 on the horizontal axis. Conversely, if the difference between $\tau_{\mathrm{c,AD}}$ and $\tau_{\mathrm{c,Sim}}$ is significant, the blue plots begin to increase much farther along the horizontal axis. As displayed in this figure, the difference between $\tau_{\mathrm{c,AD}}$ and $\tau_{\mathrm{c,Sim}}$ in RND and RR, in which the largest eigenvalue is prominently dominant, are almost negligible. In contrast, the differences are significant in CNNS and EuroRoad, which possess a comparatively large ρ for the non-largest eigenvalues as displayed in Figure 1 and large modularity Q as displayed in Table 1. These results demonstrate that an approximation method only considering the largest eigenvalue and the principal eigenvector is not appropriate for a network having comparatively large ρ_k values for its non-largest eigenvalues and large modularity Q.

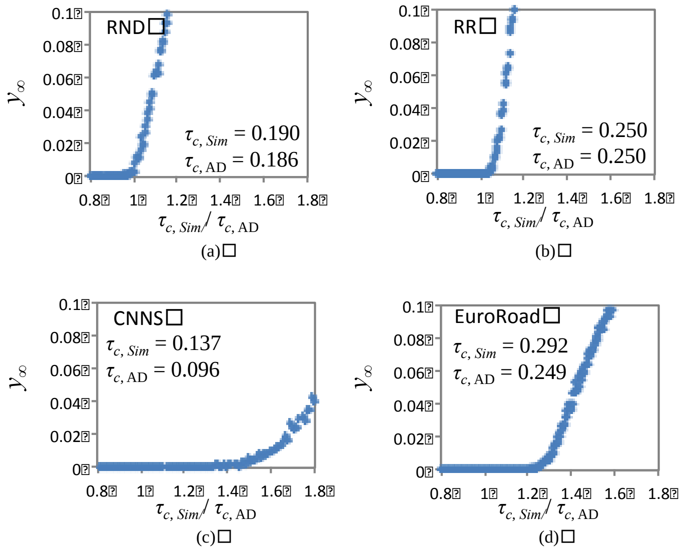


Fig. 2. Simulation results around the threshold. The steady-state fractions of infected nodes, y_∞, for (a) the Erdos-Renyi random network, (b) the random regular network (RR), (c) the Co-author Networks of Network Scientists

(CNNS) and (d) Euro road network (EuroRoad) are plotted as the function of effective infection ratio normalized by the analytically derived threshold of each network. The simulated threshold $\tau_{c,\mathrm{Sim}}$ for each network corresponds to the tipping point of blue plots in each figure.

IV. Relationship Between Modularity and Importance of the Non-Largest Eigenvalue

According to the results in the previous section, it can be hypothesized that not only the largest eigenvalue but also the other non-largest eigenvalues are also important in networks with the large optimal modularity Q. The comparison results in the previous section show that the difference between $\tau_{c,\mathrm{AD}}$ and $\tau_{c,\mathrm{Sim}}$ in RND and RR, in which the largest eigenvalue is prominently dominant, are almost negligible. In contrast, the differences are significant in the real networks that show high optimal modularity Q, such as CNNS (Q = 0.84) and EuroRoad (Q = 0.86). Also, in our previous work [28], we proposed an index “the diffusion power” that can quantify the ease of diffusion on an arbitrarily network. An investigation of the diffusion power indicates that the networks with high optimal modularity Q shows that there are significant differences between the diffusion power when considering the all eigenvalues and eigenvectors and the diffusion power only considering the largest eigenvalue and principal eigenvector. Therefore, we hypothesized that the modularity of networks relates the importance of the non-largest eigenvalues and eigenvectors for the analysis of their diffusion dynamics. Therefore, we investigate the relationship between the optimal modularity Q and the importance of the non-largest eigenvalues and eigenvectors.

$$\mathrm{Index} \equiv \frac{\max(\|\mathbf{u}_{2\sim 10}\|^2)}{\|\mathbf{u}_1\|^2}. \qquad (8)$$

To show the relationship between the optimal modularity Q and the importance of the non-largest eigenvalues, we firstly develop the network formation algorithm that can change optimal modularity Q of the network by changing the network modularity control parameter (NMCP), *p*. The NMCP determines the ratio of the number of total links connecting the nodes inside the modules to the number of total links interconnect between modules. The step-by-step procedures for this parameterized network is as follows,

Step 1: Determine the number of module, the size of the modules (i.e. the number of nodes in each module), and the number of total links, *L*, for entire network in advance.

Step 2: Determine the NMCP, then caluculate the number of links interconnecting between modules.

Step 3: Calculate the number of links connecting the nodes inside each module, which is calculated by multiplying the NMCP and the number of total links (i.e. *p*∗*L*).

Step 4: Randomly connect the links within each module by the links calculated in Step 3.

Step 5: Calculate the number of links inter-connecting each module, which is calculated by (1–*p*)*L*.

Step 6: Randomly connect each module by the links calculated in Step 5.

As shown in Figure 3, when the value of *p* increases, densities inside each module are increase while the connections between each module become sparse, then the optimal modularity Q also increases.

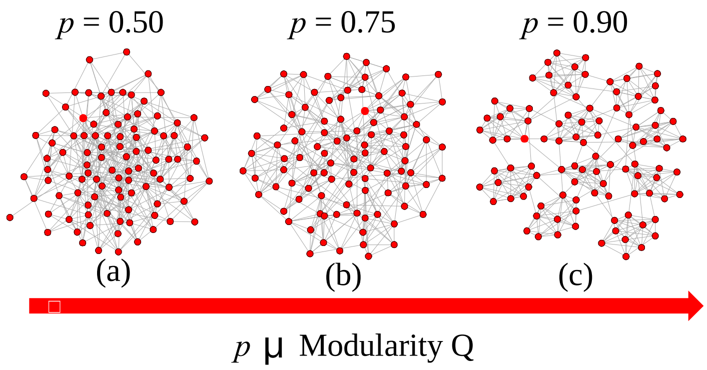
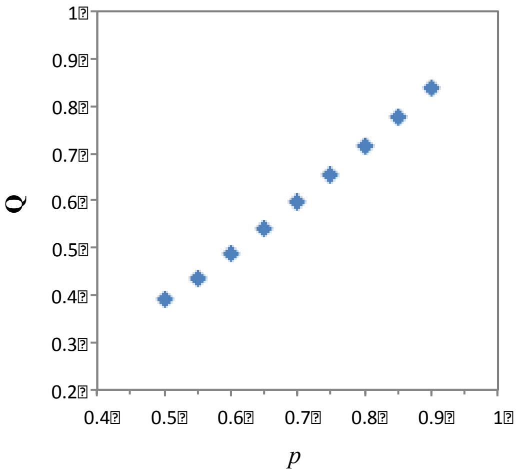


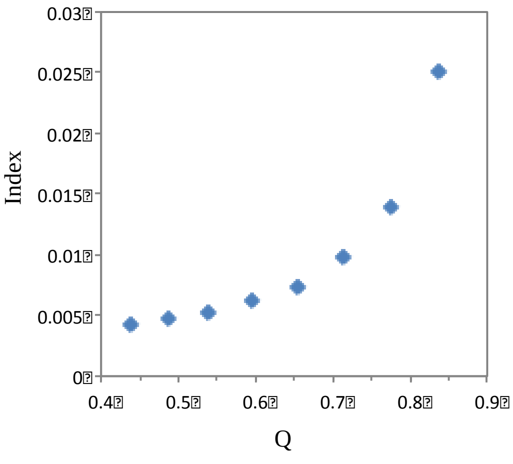
Fig. 3. Examples of modular network created by the parameterized network formation algorithm. The parameterized modular network formation algorithm can change optimal modularity Q of the network by changing network modularity control parameter (NMCP), *p*. For these three modular networks, the number of modules is 10, the size of each node is 10 nodes and the total number of links is 400. (a) Modular network when *p*=0.50, 200 links for inner-module links and 200 links for the inter-module links. (b) Modular network when *p*=0.75, 300 links for inner-module links, and 100 links for the inter-module links. (c) Modular network when *p*=0.90, 360 links for inner-module links and 40 links for the inter-module links.

According from the equation (7), an index to quantify the importance from the non-largest eigenvalues is defined as follows,

Then, using the proposed network formation method, we constructed modular networks as changing the value of *p* from 0.5 to 0.9 by 0.05. 10,000 modular networks for a given *p* are constructed and calculated the optimal modularity Q and the proposed index. Then, the average values of these values are plotted in figure 4.



(a)



(b)

Fig. 4. (a) Relationships between the parameter p in the proposed network formation algorithm and the average value of optimal modularity Q, (b) Relationships between averaged optimal modularity Q and the average value of Index

The figure 4 shows results for the modular networks with 400 nodes and 2,000 links created by the proposed network formation method. Figure 4-(a) shows the relationship between the parameter p and the averaged value of the optimal modularity Q. Also, figure 4-(b) shows the relationship between the averaged optimal modularity Q and the average value of the index. As can be seen these figure, optimal modularity Q linearly increase as the value of p in the proposed network formation method increase. Also, the average value of the Index exponentially increases as the average value of optimal modularity Q increase, which verifies the fact that the importance from the non-largest eigenvalues in diffusion dynamics increase as the optimal modularity Q of the network increase.

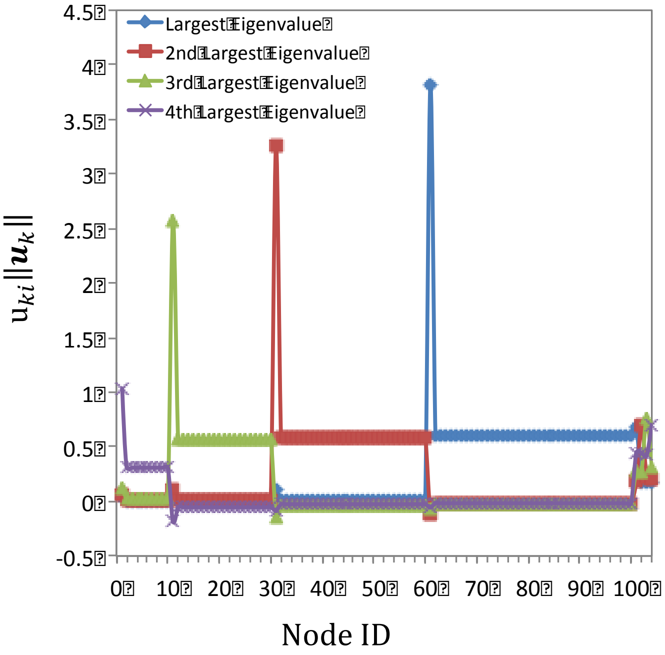
V. Node-Level Eigenvalue Influence Index

In this section, we investigate how does modular structure affects the spectral properties in the networks. Our proposed an index the Node-level Eigenvalue Influence Index (NEII), ω, that can quantify the influences from an arbitrary eigenvalue, λ_k, to the dynamical process on each node.

The equation (5) indicates that the significance of the contributions from an arbitrary kth eigenvalue to increase the infection probability on a node i is governed by the value of $u_{ki}\|\mathbf{u}_k\|$. Therefore, we defined the Node-level Eigenvalue Influence Index (NEII) $\omega_{ki} \equiv u_{ki}\|\mathbf{u}_k\|$ and investigate the ω_{ki} on each node. In the previous literatures [28], the localized-delocalized phenomenon of eigenvalues is measured by the inverse participation ratio (IPR). If IPR is large, the infections diffuse only within the small confined area, and vice versa. However, the IPR does not distinguish positive or negative, so that the IPR is only applicable for the largest eigenvalue. On the other hand, NEII can distinguish positive or negative influence from all eigenvalues and can be applied to the analysis of the impacts from the all eigenvalues. According to Perron-Frobenius theory, the only eigenvalue in which all elements in the corresponding eigenvector are non-negative is the largest eigenvalue λ_1 . In other words, the other eigenvectors corresponding to the other non-largest eigenvalues have negative elements, which means that the corresponding ω_{ki} contributes to decrease the probability of infection on the node i if ω_{ki} is negative.

Fig. 5 shows the shows the distribution of ω_{ki} on each node in the benchmark toy network. The benchmark toy network consists of four different size star networks connecting each other via the four-nodes complete graph at the center, as shown in Fig. 6 As can be seen in Fig. 5 and Fig. 6, ω_{1i} for the largest eigenvalue is always positive because of the Perron-Frobenius theory. Also, the value of ω_{1i} for the largest eigenvalue and its corresponding eigenvector are positively maximized on node #61 that is the hub node in the largest star graph. The influences from the second largest eigenvalue and the corresponding eigenvector are the positively maximized on the node #31 the hub node in the second largest star graph, but, as can be seen in Fig. 5-(b) that is the enlarged view of Fig. 5-(a), the second largest eigenvalue negatively affect to the node #61 that is the hub node in the largest star graph. In addition to that the third largest eigenvalue positively influence on the node #11 that is the hub of third largest star graph, but negatively affect the hubs of the largest and the second largest star graphs.

Fig.6 visualizes the significance of the value of ω_{ki} for the largest eigenvalue to the fourth eigenvalue (k = 1 to 4) by color gradient on each node on network. In these figures, the maximum positive value of ω_{ki} is coloured by the deepest red and gradually changes to green as the relative significance approaches to zero. The minimum negative value of ω_{ki} is coloured by the deepest blue and gradually changes to green as the relative significance approaches to zero. The size of each node is proportional to the absolute value of ω_{ki}.



(a)

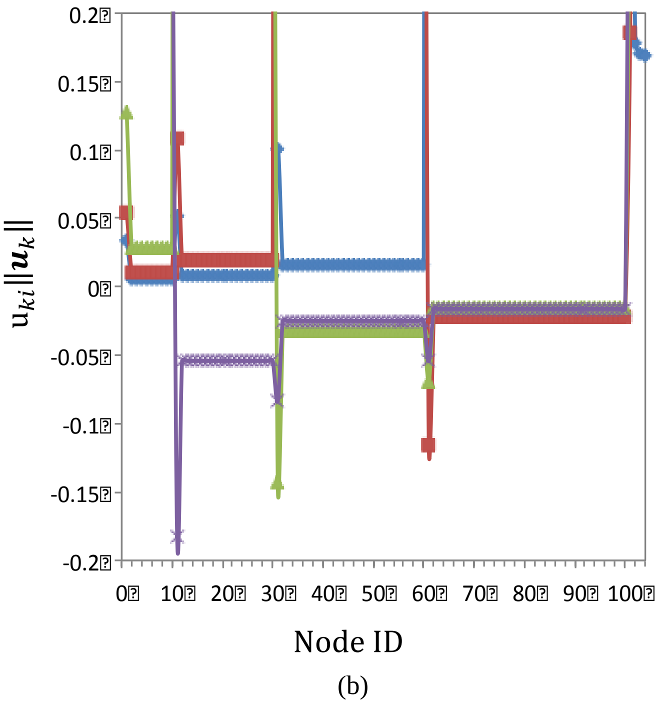


Fig. 5. (a) Distribution of ω_{ki} on each node in the benchmark toy network. (b) Enlarged view

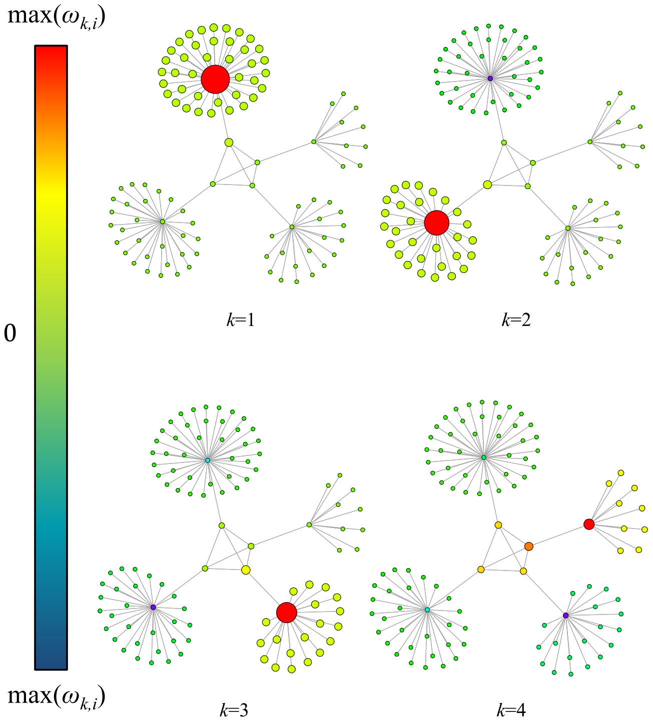


Fig. 6. Visualization the significance of the value of the ω_{ki} by color gradient on each node on network

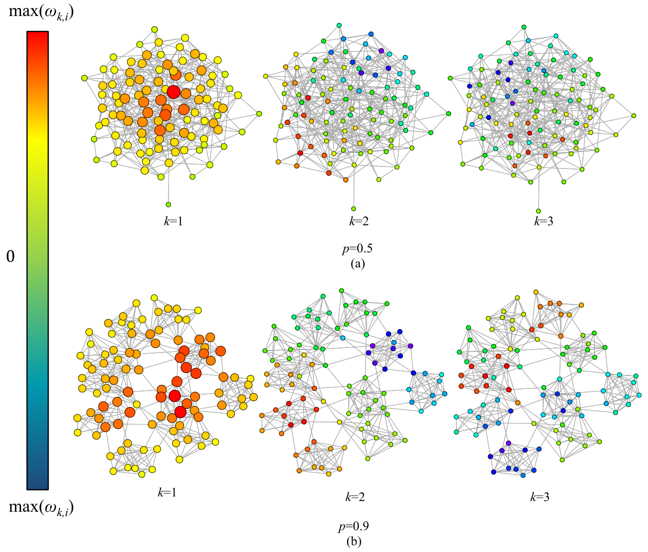


Fig. 7. Distribution and significance of the value of ω_{ki} when p for the proposed parameterized modular network formation algorithm is changed (a) p = 0.5 and (b) p = 0.9.

Appling this visualization technique to the artificial modular networks crated by the parameterized modular network formation algorithm proposed in the previous section, we observed how the distribution and significance of the value of ω_{ki} varies as modularity of networks increase. Fig. 7 shows the 100-nodes modular networks (the number of modules is 10 and each module size is 10 nodes) created by the proposed network formation algorithm in which each node is colored by the significance of the value of ω_{ki} for k = 1 to 4. Figure 7-(a) and (b) correspond with the modular network for P = 0.5 and 0.9 respectively. As can be seen in the figures for k=1, the color of all nodes is always red because of the positive values of the elements of principal eigenvector due to the Perron-Frobenius theorem. Also, it can be observed that, when the modularity of the network increases, the influences from the non-largest eigenvalues are localized within some modules whether the influences are positive or negative.

VI. Real Modular Networks

In this section, we investigate the NEII, ω_{ki}, in real modular networks that show comparatively high optimal modality Q, such as CNNS and EuroRoad. The results highlights that the needs of consideration of non-largest eigenvalues and corresponds eigenvectors when analyze the diffusion process on the real modular networks. Fig. 8 indicates the colored network by the significance of ω_{ki} for CNNS of which the optimal modularity Q is about 0.84. As shown, the maximum impact is provided by the fourth largest eigenvalue, which fit with the insight in the Fig. 1. Also, in Fig. 9 for EuroRoad of which the optimal modularity Q is about 0.86, the impacts from the second and fourth eigenvalues are significant, which fit with the insight in the figure 1, too.

These results indicate that we need to consider the non-largest eigenvalues and eigenvectors to capture the diffusion dynamics and the well-used approximation method only utilizing the largest eigenvalue is not applicable for the real networks with high optimal modularity.

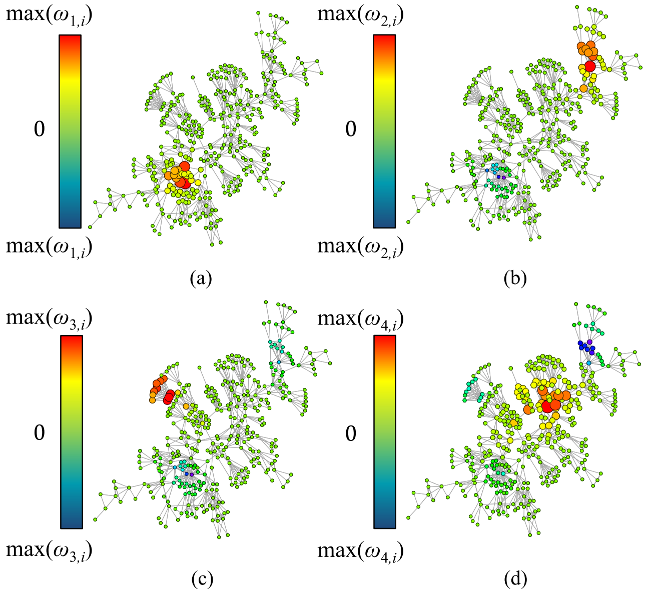


Fig. 8. Visualization of ω_{ki} on CNNS The color gradient and the size of each node indicate the relative significance of the value of ω_{ki} for (a) the largest eigenvalue λ_1, (b) the second largest eigenvalue λ_2, (c) the third eigenvalue λ_3, and (d) the fourth largest eigenvalue λ_4.

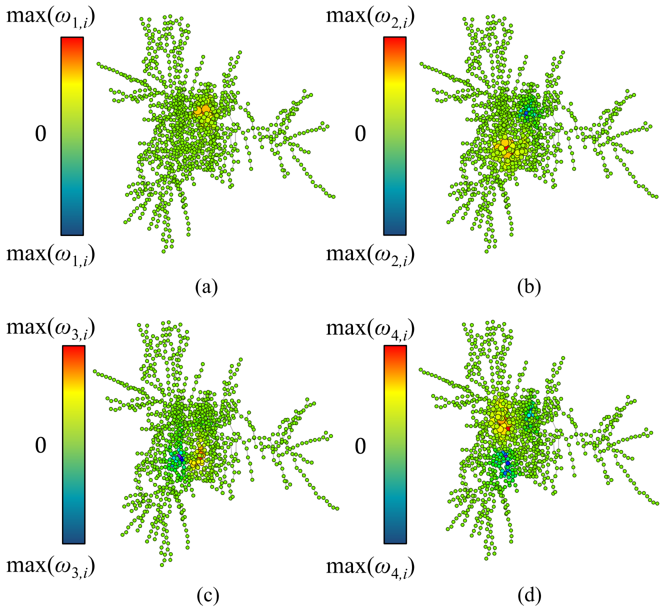


Fig. 9. Visualization of ω_{ki} on EuroRoad The color gradient and the size of each node indicate the relative significance of the value of ω_{ki} for (a) the largest eigenvalue λ_1, (b) the second largest eigenvalue λ_2, (c) the third eigenvalue λ_3, and (d) the fourth largest eigenvalue λ_4.

VII. Conclusion

Several studies reported that there exist modular structures in real networks. In this paper, we investigate spectral property of several networks. Also, diffusion phenomena in society have been studied as probabilistic diffusion dynamics on networks. So far probabilistic diffusion dynamics have been analysed in approximated manner only using the largest eigenvalue of the adjacency matrix. But, our investigation of spectral property of modular networks shows that that not only the largest eigenvalue but also the other eigenvalues are critical when analyse the network with high modularity, which verifies by the parameterized modular network formation method and numerical simulations. Furthermore, we investigated the node-level eigenvalue influence index that measures the relative dominance from each eigenvalues and their corresponding eigenvectors on each node, which indicates that the influences from each eigenvalue and corresponding eigenvectors to diffusion dynamics are localized within the modular structures. For our future works, we will extend our spectral analysis to the investigation of the relationship between Laplacian matrix of the networks and probabilistic diffusion dynamics.

Acknowledgment

The authors would like to thank anonymous reviewers for their valuable comments. This research is supported in part by Grant-in-Aide for Scientific Research in Japan No.25330277 & No.26282089.

Reference

[1] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney, “Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters”, Internet Math, Vol. 6, No. 1, pp.1-123, 2009

[2] Fortunato, S. (2010), Community detection in graphs, Physics Reports, Vol.486, No.3-5:75-174

[3] Reichardt, J. (2009) Structure in complex networks, Lecture Notes in Physics 766, Springer

[4] Girvan, M .,and M.E.J. Newman (2002) Community structure in social and biological networks, Proceeding of the National Academy of Sciences of the United States of America, Vol99 :7821-7826

[5] Newman, M.E.J, and M. Girvan (2004) Finding and evaluating community structure in networks, Physical Review E, Vol.69, No. 026113

[6] Newman, M.E.J (2004) Fast algorithm for detecting community structure in networks, Physical Review E, Vol.69, No. 066133

[7] Newman, M.E.J. (2006) Finding community structure in networks using the eigenvectors of matrices, Physical Review E, Vol.74, No.3 :036104

[8] Newman, M.E.J. (2006) Modularity and community structure in networks, Proceeding of the National Academy of Sciences of the United States of America, Vol103 :8577-8582

[9] J. Gao, S. V. Buldyrev, H. E. Stanley, and S. Havlin, “Networks formed from interdependent networks,” Nature Physics, vol. 8, no. 1, pp. 40–48, January 2012.

[10] L. Huang, K. Park, and Y-C Lai, “Information propagation on modular networks”, Physical Review E, Vol.73, 035103, 2006

[11] A. Saumell-Mendiola, M. A. Serrano, and M. Boguna, “Epidemic spread on interconnected networks”, Physical Review E, Vol. 86, 036106, 2012

[12] F. D. Sahneh, C. Scoglio, and F. N. Chowdhury, “Effect of coupling on the Epidemic threshold in intercoonected complex networks: A spectral analysis”, arXiv:1212.4194 [physics.soc-ph], 2012

[13] H. Wang et al., “Effect of interconnection network structure on the epidemic threshold”, Physical Review E, Vol.88, 022801, 2013

[14] J. Hindes, S. Singh, C. R. Myers, and D. J. Schneider "Epidemic fronts in complex networks with metapopulation structure", arXiv:1304.4310 [physics.soc-ph]

[15] M. Dickison, S. Havlin, and H. E. Stanley "Epidemics on interconnected networks", Phys. Rev. E Vol.85, 066109, 2012

[16] R. Albert, A. L. Barabasi, "*Statistical mechanics of complex networks*". Reviews of Modern Physics, Vol.74, pp.47-94, 2002.

[17] S. N. Dorogovtsev and J. F. F. Mendes, "*Evolution of networks*", Advances in Physics, Vol.51, pp.1079-1187, 2002.

[18] M. E. J. Newman, "*The structure and function of complex networks*", SIAM Review, Vol.45, pp.167-256, 2003.

[19] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "*Critical phenomena in complex networks*", Reviews of Modern Physics, Vol.80, pp.1275- , 2008.

[20] R. Pastor-Satorras and A. Vespignani, "*Epidemic dynamics and endemic states in complex networks*", Physical Reviews E, Vol.63, 066117, 2001.

[21] M. Bogu ñ a, R. Pastor-Satorras, and A. Vespignani, "*Absence of Epidemic Threshold in Scale-Free Networks with Degree Correlations*", Vol. 90, 028701, 2003.

[22] J. O. Kephart, S. R. White, "*Directed-graph epidemiological models of computer viruses*", Proceedings of Research in Security and Privacy, pp.343-359 1991.

[23] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "*Epidemic spreading in real networks: An eigenvalue viewpoint*", Proceeding of International Symposium on Reliable Distributed Systems (SRDS'03) 2003, pp.25-34, 2003.

[24] P. V. Mieghem, J. Omic, and R. Kooij, "*Virus spread in networks*", IEEE/ACM Trans. Netw. Vol 17, No. 1, pp.1-14, 2009.

[25] P. V. Mieghem, "*Epidemic phase transition of the SIS type in networks*", Europysics Letters, Vol.97, 48004, 2012.

[26] R. Pastor-Satorras and A. Vespignani, "*Epidemic Spreading in Scale-Free Networks*", Physical Review Letters, Vol.86, 3200-3203, 2001.

[27] A.V. Goltsev, et al., "*Localization and Spreading of Diseases in Complex Networks*", Physical Review Letters. Vol.109, 128702, 2012.

[28] K. Ide, A. Namatame, L. Ponnambalam, F. Xiuju, and R. S. M. Goh, "Spectral Approach for Information Diffusion", Proceeding of ASE SocialCom 2014, In press

[29] J. Billen, M. Wilson, and A. Baljon, "*Eigenvalue spectra of spatial-dependent networks*", Physical Review E, Vol.80, 046116, 2009

[30] M. E. J. Newman, "*The structure of scientific collaboration networks*", Proc. Natl. Acad. Sci. USA, Vol.98, pp.404-409, 2001.

[31] Open dataset is available at: http://www-personal.umich.edu/~mejn/netdata/

[32] D. Greene, and P. Cunningham, "*Producing a Unified Graph Representation from Multiple Social Network Views*", arXiv preprint arXiv:1301.5809, 2013.

[33] Open dataset is available at: http://mlg.ucd.ie/networks/politics-uk.html

[34] L. Šubelj and M. Bajec, "*Robust network community detection using balanced propagation*", European Physical Journal B,, Vol.81, No.3, pp.353-362, 2011.

[35] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson,"*Behavioral Ecology and Sociobiology*", Vol.54, pp.396-405, 2003.

[36] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt and A. Arenas, "*Self-similar community structure in a network of human interactions*", Physical Review E , vol. 68, 065103(R), 2003.

[37] P.Gleiser and L. Danon , Adv. Complex Syst.6, 565, 2003.

Hardware Segmentation on Digital Microscope Images for Acute Lymphoblastic Leukemia Diagnosis Using Xilinx System Generator

Prof. Kamal A. ElDahshan
Professor of Computer Science, Dept. of Mathematics,
Faculty of Science,
AL-AZHAR University, Cairo, Egypt

Prof. Mohammed I. Youssef
Dept. of Electronic Engineering, Faculty of Engineering,
AL-AZHAR University, Cairo, Egypt

Dr. Emad H. Masameer
Assistant Professor of Computer Science, Dept. of
Mathematics, Faculty of Science,
AL-AZHAR University, Cairo, Egypt

Mohammed A. Mustafa
Lecturer assistant, Dept. of Management Information
Systems, Modern Academy for Computer Science and
Information Technology, Cairo, Egypt

***Abstract*—Image segmentation is considered the most critical step in image processing and helps to analyze, infer and make decisions especially in the medical field. Analyzing digital microscope images for earlier acute lymphoblastic leukemia diagnosis and treatment require sophisticated software and hardware systems. These systems must provide both highly accurate and extremely fast processing of large amounts of image data. In this work, the hardware segmentation framework for Acute Lymphoblastic Leukemia (ALL) images based color histogram of Hue channel of HSV color space is proposed to segment each leukemia image into blasts and background using Field Programmable Gate Array (FPGA). The main purpose of this work is to implement image segmentation framework in a FPGA with minimum hardware resources and low execution time to be suitable enough for medical applications. Hardware framework of segmentation is designed using Xilinx System Generator (XSG) as DSP design tool that enables the use of Simulink models, implemented in VHDL and synthesized for Xilinx SPARTAN-3E Starter kit (XC3S500E-FG320) FPGA.**

Keywords—Medical Image Processing; FPGA; Image Segmentation; Xilinx System Generator

I. Introduction

Leukemia is a type of cancer caused by abnormal increase of the white blood cells. According to [1] leukemia can be classified into acute and chronic. Acute leukemia spreads very rapidly and has to be treated promptly rather than chronic leukemia that does not have to be treated promptly. Acute leukemia can be either lymphoblastic (ALL) or myelogenous (AML), based on affected cell type. Chronic leukemia can be either lymphoblastic (CLL) or myelogenous (CML). Acute lymphoblastic leukemia (ALL) is considered the prime focus of this work, which has a higher expectation of survival rate compared to AML.

Image segmentation is a process of partitioning the image into multiple segments. For biomedical imaging applications, image segmentation is a founding step in image analysis as it will directly affect the post-processing. It is a crucial component in diagnosis [2] and treatment [3]. The main goal of acute leukemia blood cell segmentation is to extract components such as blast from its complicated blood cells background. There are many techniques that have been developed for image segmentation such as threshold techniques [4], edge detection [5] and watershed techniques [6]. Due to the complex nature of blood cells and overlapping between these cells, segmenting them remains a challenging task [7]. Many algorithms for segmentation have been developed for grayscale images rather than color images which require more information to be processed [8].

Image processing algorithms implemented in FPGA hardware have emerged as the most viable solution for improving the performance of image processing systems. It offers a compromise between the flexibility of general purpose processors and ASICs. FPGAs are recently used in many image processing applications such as image compression [9] [10] [11], image filtering [12] [13] and wireless communication [14] [15].

Xilinx System Generator is a DSP design tool [16] [17] that deal with many images processing application. XSG is a part of the ISE design suite that provides Xilinx DSP Blockset for application specific design. The main advantage of using Xilinx system generator for FPGA implementation is that Xilinx Blockset provides close integration with MATLAB Simulink that helps in co-simulating the FPGA module with pixel vector provided by MATLAB Simulink Blocks [18].

In this paper, segmentation based color histogram of Hue channel of HSV color space is used [19]. Before hardware segmentation, pre-processing of the acute lymphoblastic leukemia image to convert an image from RGB color space to HSV color space is required. This work focuses on implementing multilevel thersholding segmentation based on color histogram of H channel of HSV color space in hardware.

All algorithms are initially implemented in MATLAB to realize the segmentation results. The pipelined framework of multilevel thersholding image segmentation is implemented in a FPGA. This work presents segmentation framework using Xilinx System Generator and also implemented in low cost basic FPGA device Spartan-3E.

II. Leukemia Segmentation

The ultimate goal of ALL segmentation is to extract components such as blast from its complicated blood cells background. There are 6 steps involved in applying image segmentation process:

Step 1: transforming the source RGB color space to HSV color space.

Step 2: extracting H channel from HSV color space.

Step 3: Selecting color range of nucleus and cytoplasm by using color histogram of H channel. Two angle values A1, A2 are obtained from color histogram for multilevel thersholding segmentation.

Step 4: Implementing the median filter N X N (N = 7) to the resulted images.

Step 5: Synthesizing the HSV image.

Step 6: Converting the HSV image to RGB to display.

Fig. 1 illustrates the block diagram of ALL segmentation.

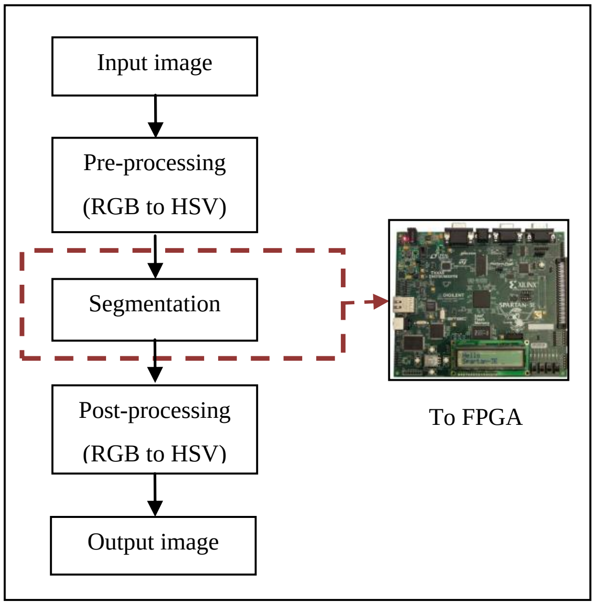


Fig. 1. Block diagram of ALL segmentation

The pre-processing and post-processing steps for ALL segmentation are proposed using Simulink blocks. Fig. 2 presents color space conversion block from RGB to HSV image and applying median filter to H channel for further processing.

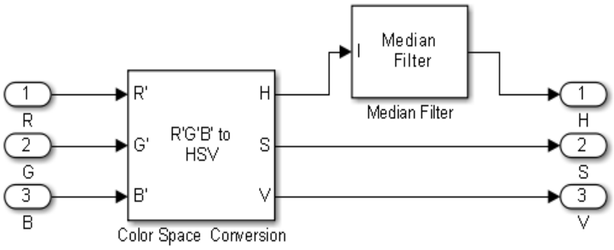


Fig. 2. Pre-processing for ALL segmentation

III. Proposed Framework

A. Hardware Design

To accomplishing Image processing task using Xilinx System Generator, two Software tools are needed to be installed. This work uses MATLAB version R2012b and Xilinx ISE 14.5. The model is built for image segmentation using library provided by Xilinx Blockset. According to the design of segmentation to meet hardware requirements, pre-processing the HSV image prior to the main hardware architecture is needed due to the nature of hardware that deals with an image as a vector. Also, image post-processing is required. There are three stages involved in ALL hardware segmentation process using Simulink and Xilinx blocks:

- Hardware pre-processing
- Xilinx models for HW segmentation
- Hardware post-processing

Fig. 3 represents the main block diagram of proposed framework.

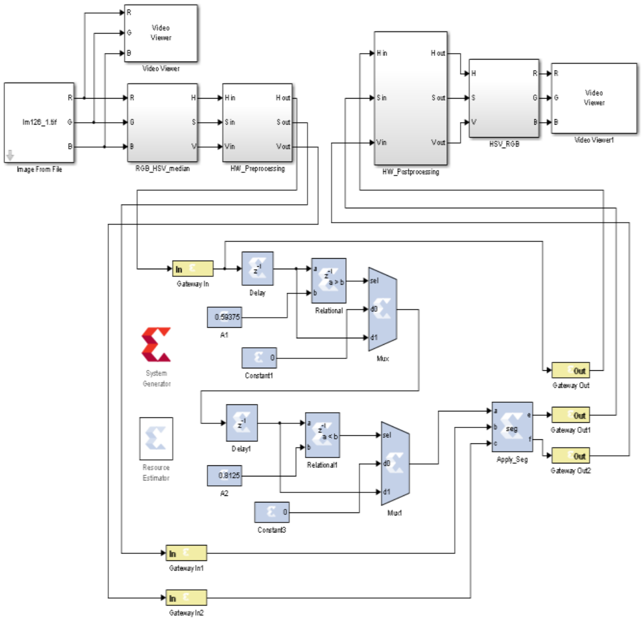


Fig. 3. Block diagram of proposed framework

The image pixels are provided to Xilinx models in the form of multidimensional H|S|V separate color signals in the form of vector in Xilinx fixed point format. The reflected results can be seen on a video viewer. Once the expected results are obtained, XSG is configured to be suitable for SPARTAN-3E XC3S500E-FG320.

a) Hardware Pre-Processing

Pre-processing blocks provide an input image suitable for FPGA as vector array. Reshape blocks convert the HSV image channels into single array of pixels. The process of setting sampling mode is obtained using frame conversion. Unbuffer blocks convert this frame to scalar samples output at a higher sampling rate. The model based design used for image pre-processing for FPGA is shown in Fig. 4.

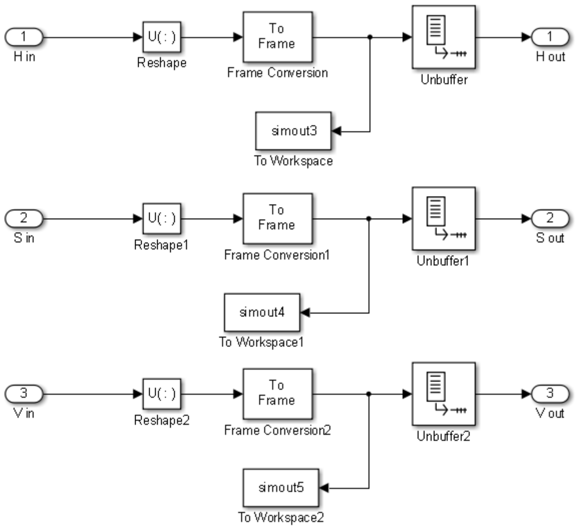


Fig. 4. Hardware Pre-processing

b) Xilinx Models for HW Segmentation

Hardware segmentation process is modeled using Xilinx blocks. Once the FPGA boundaries have been established using the Gateway In and Gateway Out blocks, the DSP design can be constructed using Xilinx DSP blocks. Within the Gateway In and Gateway Out blocks, Simulink blocks are not supported for use.

Xilinx fixed point type conversion is made by Gateway In blocks. Image Segmentation process is achieved based on two angle values that obtained from color histogram of H channel. These two values are represented using two constant blocks. Multilevel thersholding operated using Relational and Mux blocks. This is followed by certain blocks to merge all the processed data. Fig. 5 shows the ALL segmentation using Xilinx blocks.

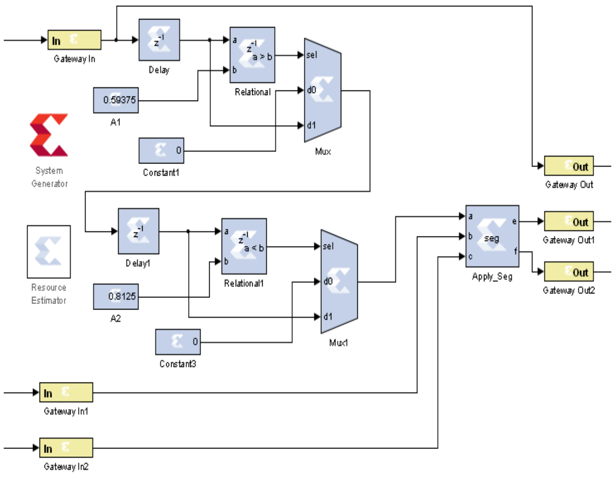


Fig. 5. ALL Image Segmentation Using Xilinx blocks

c) Hardware Post-Processing

Post-processing blocks converts an image from vector to 2D matrix as shown in fig. 6. Buffer blocks are used to convert scalar samples to frame output at lower sampling rate. The process of converting 1D image to 2D image is obtained using reshape blocks.

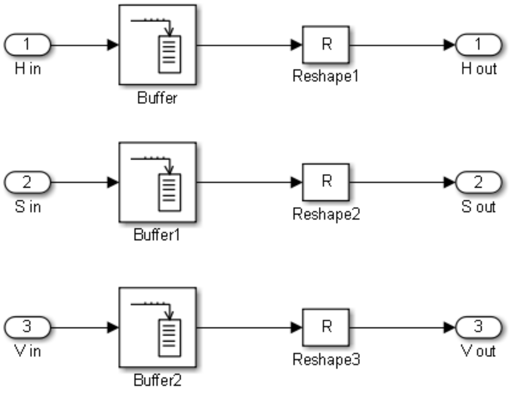


Fig. 6. Hardware Post processing

B. Hardware Co-Simulation

Once the results are obtained from hardware design; the model is implemented for JTAG hardware co-simulation. The System generator parameters are set and generated. On compilation, programming file in VHDL is created to be accessed by Xilinx ISE. The module is synthesized and implemented on FPGA. Fig. 7 illustrates the hardware co-simulation block.

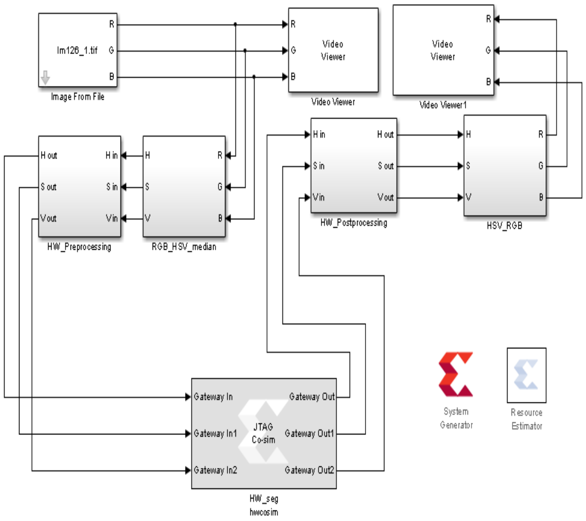


Fig. 7. Hardware co-simulation

Fig. 8 presents the RTL schematic of the resulting circuit for hardware segmentation.

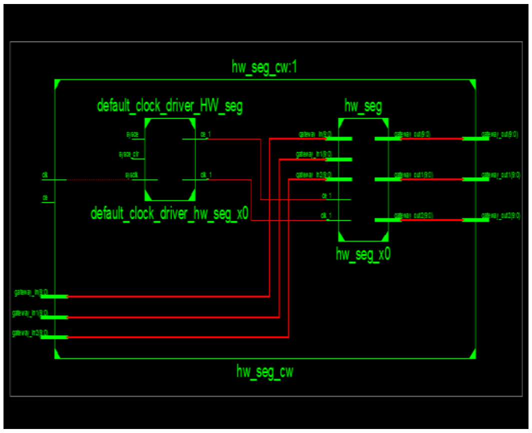


Fig. 8. RTL Schematic for Hardware Segmentation

IV. Results and Discussions

Microscope Images of ALL are taken from ALL-IDB database [20]. The images of the database have resolution 256 × 256. Figure 9 shows the sample of ALL images.

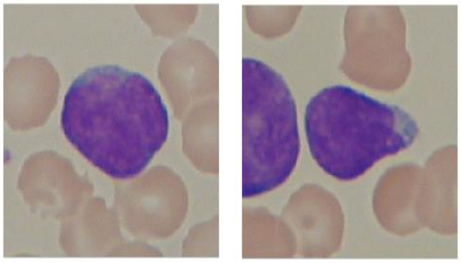


Fig. 9. Sample of ALL images

The SPARTAN-3E (XC3S500E-FG320) resource usage is estimated for proposed framework as shown in table 1. The VHDL code for the proposed hardware segmentation has 2381 lines of VHDL code. This is due to the huge amount of floating point-fixed point conversions.

TABLE I. Device Utilization Summary

Resource	Used	Available	Utilization
Flip Flop	22	9312	1%
Slices	30	4656	1%
LUTs	39	9312	1%
IOBs	61	232	26%
Minimum period: 3.524ns (Maximum Frequency: 283.768MHz) Minimum input arrival time before clock: 1.973ns Maximum output required time after clock: 8.766ns Maximum combinational path delay: 6.113ns			

The original image of ALL is shown in Fig. 10 while the resulted image after hardware segmentation is represented in fig. 11.

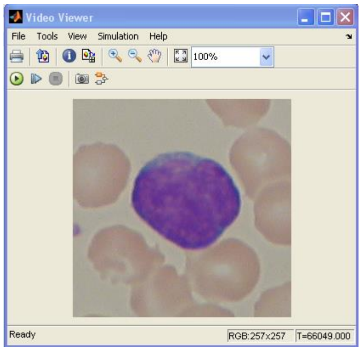


Fig. 10. Original ALL image

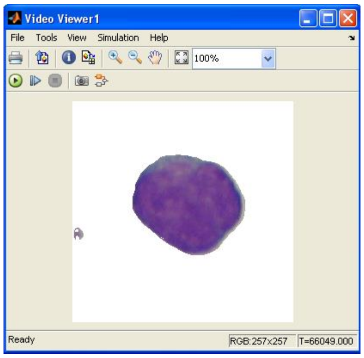


Fig. 11. Resulted ALL image

V. Conclusion

In this work, the hardware segmentation framework based color histogram of Hue channel of HSV color space is proposed. The results obtained from the Xilinx and Simulink model showed that the proposed framework achieved a superior performance and quality. In term of the device utilization, the implementation occupies around 1% of the used SPARTAN-3E XC3S500E-FG320 FPGA. In the proposed framework, the interfacing of Matlab and XSG is done. The ALL Image segmentation is performed on Matlab as well as Simulink Model and it has been verified using SPARTAN-3E XC3S500E-FG320 FPGA.

In the future work, the design framework used in this work will be optimized and implemented on other Xilinx FPGA Kits such as Virtex-7.

REFERENCES

[1] G .. C.C.Lim, "Overview of Cancer in Malaysia," Japanese Jomal of Clinical Oncology, Department of Radiotherapy and Oncology, Hospital Kuala Lumpur, 2002.

[2] P. Taylor, "Invited review: computer aids for decision-making in diagnostic radiology— a literature Review," Brit. J. Radiol.., 68:945–957, 1995.

[3] V.S. Khoo, D.P. Dearnaley, D.J. Finnigan, A. Padhani, S.F. Tanner, and M.O. Leach, "Magnetic resonance imaging (MRI): considerations and applications in radiotheraphy treatment planning," Radiother. Oncol., 42:1–15, 1997.

[4] S.U. Lee and S.Y. Chung, "A comparative performance study of several global thresholding techniques for segmentation," in Computer Vision, Graphics and Image Processing, no. 52, 1990, pp. 171-190.

[5] D. Ziou and S. Tabbone, "Edge detection techniques - An overview," in Technical Report, no. 195, Departement de math et informatique, Universite de Sherbrooke, 1997.

[6] L. Vincent and P. Soille, "Watersheds in digital spaces; an efficient algorithm based on immersion simulations," in IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 13, 1991, pp. 583-597.

[7] S. Mao-jun, W. Zhao-bin, Z. Hong-juan, and M. Yi-de, "A New Method for Blood Cell Image Segmentation and Counting Based on PCNN and Autowave," in ISCCSP 2008 Malta, 2008.

[8] W. Skarbek and A. Koschan, "Colour Image Segmentation," Institute for Technical Informatics, Technical University of Berlin, Berlin 1994.

[9] J. Rosenthal, "JPEG Image Compression Using an FPGA," MS thesis, Dec 2006.

[10] Y. Kim, K. Jun and K. Rhee , "FPGA Implementation of Subband Image Encoder Using Discrete Wavelet Transform," 1999 IEEE TENCON.

[11] S. K.Shah, R. K.Soni and B. Shah, "FPGA Implementation of Image Compression using bottom- up approach of Quad tree technique," IETE Journal of research , Vol 57, Issue 2, Mar-Apr 2011.

[12] D. Rao, S. Patil, N. Babu and V. Muthukumar , "Implementation and Evaluation of Image Processing Algorithms on Reconfigurable Architecture using C-based Hardware Descriptive Languages," International Journal of Theoretical and Applied Computer Sciences,Volume 1 Number 1, 2006, pp. 9–34.

[13] Nelson, "Implementation of Image Processing Algorithms on FPGA hardware," MS thesis, May 2000.

[14] H. Taha, A. Sazish, A. Ahmad, M. Sharif and A. Amira, "Efficient FPGA Implementation of a WirelessCommunication System Using Bluetooth Connectivity," IEEE, 2010.

[15] R. Mehra and S. Devi, "Efficient hardware co-simulation of down converters for wireless communication systems," International journal of VLSI design & Communication Systems (VLSICS), Vol.1, No.2, June 2010.

[16] Z. Shanshan and W. Xiaohong, "Vehicle Image Edge Detection Algorithm Hardware Implementation on FPGA," International Conference on Computer Application and System Modeling ,ICCASM 2010.

[17] "Xilinx System Generator User's Guide," downloadable from;http:// www. Xilinx.com, 2010.

[18] "Xilinx System Generator User's Guide," www.Xilinx.com , www.Xilinxforum.

[19] K.A. Eldahshan, M.I. Youssef, E.H. Masameer and M.A. Mustafa. "Segmentation Framework on Digital Microscope Images for Acute Lymphoblastic Leukemia Diagnosis based on HSV Color Space," International Journal of Computer Applications 90(7):48-51, March 2014.

[20] Donida Labati, R., Piuri, V., Scotti, F., "ALL-IDB: the Acute Lymphoblastic Leukemia Image DataBase for image processing," 2011.

Sensitive Data Protection on Mobile Devices

Fan Wu
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

Chung-han Chen
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

Dwayne Clarke
Department of Computer Science
Tuskegee University
Tuskegee, Alabama, USA

***Abstract*—Nowadays, many mobile devices such as phones and tablets are used in the workplace. A large amount of data is being transferred from one person to another. Data transfer is used for several different fields. Many companies and institutions are focusing on research and development on the way to further protect sensitive data. However, sensitive data still get leaks on mobile devices. To analyze how sensitive data get leak, a simulation on transferring sensitive data is developed. In this paper, we present the analysis of mobile security problem dealing with sensitive data from getting out. The goals in our research are for users to have a greater understanding on how data is being transferred and prevention sensitive data from being stolen. Our work will benefit mobile device users and help to prevent sensitive data from being stolen. Our experiments show different ways to safely transfer information on mobile devices by testing three methods types, which are back-up, encryption, and lock plus wipe data.**

Keywords—Mobile Security; Sensitive Data; Data Protection

I. Introduction

Mobile Security is a very important field in the security world. Computer data transfer plays a very important role in daily life. The importance of the transfer data can range from business, schools, companies, and government documents. The process of transfer data is to focus on finding answers for life problem by transfer information from one person to another. In order to analysis how data being transfer from the mobile device and to prevent sensitive data from getting out there, we need to simulate several different ways how data can be transfer on the mobile device. Data transfer is achieving by safely copying or moving important data from one location to another. Some examples are, computer to computer, computer to mobile device, mobile device to mobile device, mobile device to the server, and computer to the server. Now it is much easier and faster to transfer data today than it was in the past few decades.

Nevertheless, it is even easier for hackers to get sensitive data from the users. As a result, many researches are needed to find safer ways to transfer data and information on the mobile device.

In this paper, we present analysis of mobile security problem dealing with sensitive data from getting out there. Data transfer is copying data from a storage device to memory also copying data from one computer to another [1]. As a result transfer data has increased it range for transfer data it just not only are computer, but are phones, tablets, and server.

Data transfer has many benefits, which include, offloading server work, robustness support environment, transferring only relevant data, backup data, and balancing resources in an application development environment. A redistribution of work load boosts response time for production systems that run on servers.

Increasing robustness to the decision support environment works in the case of a network failure that would temporarily eliminate access to the server's data. Transfers Only Relevant Data can transfer only the data that you need to use. Model of a Centralized Control Point automated jobs that can run during non-peak hours can distribute data and applications to multiple computers that need the data and the applications for the next day's work. Back-Up Client Data and applications can be copied from a client that has limited memory resources to a server that has more memory resources. This provides a backup in case of loss on the client. Balances Resources in Application Development Environment programmers can use Data Transfer Services to make efficient use of network resources [2].

This paper focuses on three solutions on how to prevent sensitive data from getting out on the mobile device. They are back-up, encryption, and remote lock plus wipe data. These solutions can be used for many different applications not only for personal use but for the business world as well also can be used on a number of mobile devices such as, phones and tablets. Although many approaches were use on the computer and had been applied with advantage to the solutions of some of these problems, we will explored this issues on an Android phone or tablets to see if one or both can be prevent sensitive data from getting taken from the user.

The rest of the paper is organized as follow: Section II introduces some previous related work; Section III describes the background on Sensitive Data on mobile device briefly; Section IV presents the experiments and research on prevent sensitive data from getting out there; and our experimental results are presented in Section V; Finally Section VI concludes this paper with our future directions.

II. Related Work

The smart mobile devices, such as smartphones and tablets, are becoming an essential tool in people's personal and business activities. A large amount of personal data and even more sensitive important company data are stored in these devices, which also exposes a severe risk to device users when their devices are lost or stolen. If no defense mechanisms were enforced a prior, the lost or stolen devices would leak user information: your passwords can be broken, your emails could be seen, e-commence data such as online purchasing or banking transaction might be viewed; The situation would be worse when a device has the access right to Enterprise networks, e.g., via VPN, in which company networks will be exposed to malware or could be hacked It is one of major focus of security concerns for Android mobile device [3].

Sensitive data have always been important but sensitive data on a mobile device just gaining attention. Also there is some work related to sensitive data on a mobile device. Here we just refer to some recent work closely related. University of Cincinnati and Southern Polytechnic focuses on the development of Mobile Security Lab ware which shows users on how to protect and prevent sensitive data before and after a device is lost or stolen. The different between their and our work is that what type of data can be prevented on the mobile device.

Another related work was from Dimensional Research. They mainly focused on is how many mobile devices store sensitive customer and business data. The statistics results show that users reported a significant level of very sensitive information was on their mobile devices, including customer data (47%), network login credentials (38%), and corporate information made available through business applications (32%) [4].

The Ponemon Institute research focuses on [5] conducting high quality, empirical studies on critical issues affecting the management and security of sensitive information about people and organizations. In addition, only 27% of users regularly update their passwords, again, leaving them vulnerable to security attackers [6]. Ernst & Young research was explaining that sensitive information or application configurations maybe accessible to users or unauthorized parties through various means [7]. American Health Information Management Association takes sensitive data on mobile security to medical point of view.

By saying, mobile devices are easily lost or stolen and thus pose increased risks to the confidentiality and security of patient health information. Loss or theft of a device could easily result in the need for patient breach notification and subsequent reporting to the Department of Health and Human Services and media as required under the American Recovery and Reinvestment Act [8].

III. System Architecture

The Android system that we used in our research and implementation is the API (Application Programming Interface), which connects with the devices to build functions program and create application to do many things. As the API level rises up the more add-ons. All Android system compatibles devices support 32 and 64 bit processing. This platform interacts with such mobile devices as phones and tablets.

The Sensitive Data Architecture is in Fig. 1 shows sensitive user data is only available through protected APIs [9]. The components of a sensitive data on the mobile device are personal information, input device, and metadata. Those are the types of sensitive User Data. API is the only way that another user can access sensitive data from the user.

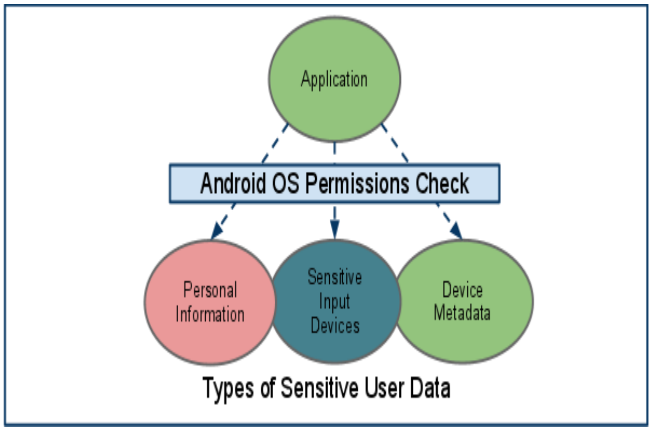


Fig. 1. Sensitive Data Architecture [9]

A. Personal Information

Personal Information is the information that identifies who you are. It will admission the user information and set it in a protected API. For example are contacts and calendar information on the device.

B. Sensitive Data Input Devices

Sensitive data inputs allow the applications or program to interact with the nearby environment, such as camera for taking pictures, microphone for speaking into or GPS for look for location. In order for third-party to gain access it needs the user permission for it.

C. Device Metadata

Device Metadata restricts access to data that is not natural, but also reveal some information about the user, user options, and the user method on the mobile device. For example are phones, logs, browser history, and text messages on the device.

IV. Protect Sensitive Data

Sensitive data has always been important not only for mobile security but computer relative issues. Some of the common problems with protect sensitive data on the mobile device are if your device get stolen, hacked, or damage. These are three methods that can be used to help the user protect sensitive data from getting out there which are back-up, encryption, and lock plus wipe data. Now for the programming, testing, and application parts to see if protect multiple types of sensitive data.

A. Back-Up

Back-Up is a copy of a file, program, or entire computer system in an event for the original get stolen, hacked, or damage. We are going to test if we can backup file such as, calendar, contacts, SMS, and even phone calls on the mobile device. To see how fast the process and where backup files will go. The way we are going to back-up files is by using a backup agent.

The algorithm 1 labeled “Back-Up Agent Code” shows class name for your backup agent, which is declared in your manifest with the android:backupAgent attribute in the <application> tag [10].

The algorithm shows the user how it works.

```
<manifest...>
...
 <applicationandroid:label="MyApplication"
          android:backupAgent="MyBackupAgent"
>
        <activity...>
             ...
             </activity>
      </application>
</manifest>
<applicationandroid:label="MyApplication"
            android:backupAgent="MyBackupAgent">
 <meta-data
android:name="com.google.android.backup.api_key"
        android:value="AEdPqrEAAAAIDaYEVgU6DJnyJdBm
U7KLH3kszDXLv_4DIsEIyQ"/>
</application>
// Get the oldState input stream
FileInputStream instream = new
FileInputStream(oldState.getFileDescriptor());
DataInputStream in = new
DataInputStream(instream);

try{
    // Get the last modified timestamp from the state file and
data file
    long stateModified = in.readLong();
    long fileModified = mDataFile.lastModified();

if(stateModified!=fileModified){
        // The file has been modified, so do a backup
       // Or the time on the device changed, so be safe and do a
backup
 }else{
    // Don't back up because the file hasn't changed
  return;
 }
}catch(IOExceptione){
    // Unable to read state file... be safe and do a
backup
}
public class MyFileBackupAgent extends
BackupAgentHelper{
  //The name of the file
          static final String TOP_SCORES = "scores";
         static final String PLAYER_STATS = "stats";
    // A key to uniquely identify the set of backup data
    static final String FILES_BACKUP_KEY =
"myfiles";

     // Allocate a helper and add it to the backup agent
   voidonCreate()                                          {
       FileBackupHelper helper = new
FileBackupHelper(this, TOP_SCORES,
PLAYER_STATS);
 addHelper(FILES_BACKUP_KEY,                    helper);
 }
}
```

Algorithm 1. Back-Up Agent Code [10]

B. Encryption

Encryption data is another way to protect your mobile device from leaks sensitive data. Encryption data transforms data into a secret code or message that unreadable form that uses algorithms. We tested to see what type of data can be encrypted and where the data is going to be storage. The process we are going to encryption data is by using encryption application which is call universal encryption app and show some coding for SMS encryption. The algorithm 1 labeled "SMS Encryption Code" shows how the encryption and decryption works with RSA encryption and decryption algorithm [11].

```
public static void generateKey() throws Exception
    {
     KeyPairGenerator gen =
KeyPairGenerator.getInstance(RSA);
     gen.initialize(512, new SecureRandom());
     KeyPair keyPair = gen.generateKeyPair();
     uk = keyPair.getPublic();
     rk = keyPair.getPrivate();
    }
   private static byte[] encrypt(String text, PublicKey
pubRSA) throws Exception
    {
     Cipher cipher = Cipher.getInstance(RSA);
     cipher.init(Cipher.ENCRYPT_MODE, pubRSA);
     return cipher.doFinal(text.getBytes());
    }
   public final static String encrypt(String text)
    {
     try {
      return byte2hex(encrypt(text, uk));
     }
     catch(Exception e)
     {
      e.printStackTrace();
     }
     return null;
    }

   public final static String decrypt(String data)
}
```

Algorithm 2. SMS Encryption Code [11]

C. Lock and Wipe

If all else failed the user have to option to lock or wipe all sensitive data for their mobile device. We tested to see if all sensitive data can be lock or if not at least wipe all sensitive data from the mobile. The application we use for our research is Lookout. The algorithm 3 labeled "Lock and Wipe Code" shows the DevicePolicyManager method wipeData() to reset the device to factory settings [12].

```
// Set device lock
```

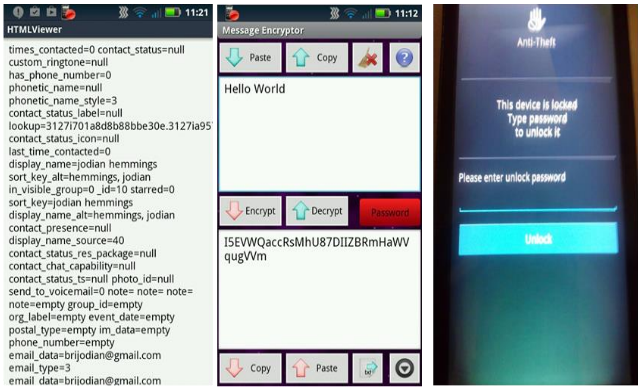
```
DevicePolicyManagermDPM;
ComponentNamemDeviceAdminSample;
long      timeMs=
1000L*Long.parseLong(mTimeout.getText().toString());
mDPM.setMaximumTimeToLock(mDeviceAdminSample,
timeMs);
DevicePolicyManager      mDPM;
mDPM.lockNow();

//Perform data wipe
DevicePolicyManager      mDPM;
mDPM.wipeData(0);
```

Algorithm 3. Lock and Wipe Code [12]

V. Experimental Results

We tested malware program to see whether malware can be removed by looking at coding. The reason we took a look at the coding is to understand how sensitive data can be protected. The methods we used in our research are back-up, encryption, and lock plus wipe data. Fig. 2 shows each method on protecting sensitive data for the mobile device. Many data were tested to see if they can be back-up, encryption, and lock plus wipe data. The results successfully show most of data can work using these three methods such as text message, phone calls, contacts, and etc. However, there are possibility better methods to deal with sensitive data.



```
times_contacted=0 contact_status=null
custom_ringtone=null
has_phone_number=0
phonetic_name=null
phonetic_name_style=3
contact_status_label=null
lookup=3127i701a8d8b88bbe30e.3127ia95
contact_status_icon=null
last_time_contacted=0
display_name=jodian hemmings
sort_key_alt=hemmings, jodian
in_visible_group=0 _id=10 starred=0
sort_key=jodian hemmings
display_name_alt=hemmings, jodian
contact_presence=null
display_name_source=40
contact_status_res_package=null
contact_chat_capability=null
contact_status_ts=null photo_id=null
send_to_voicemail=0 note= note= note=
note=empty group_id=empty
org_label=empty event_date=empty
postal_type=empty im_data=empty
phone_number=empty
email_data=brijodian@gmail.com
email_type=3
email_data=brijodian@gmail.com
```

Hello World

I5EVWQaccRsMhU87DIIZBRmHaWVqugVVm

This device is locked
Type password
to unlock it

Please enter unlock password

Fig. 2. Methods of protect sensitive data

VI. Conclusion and Future Work

A. Conclusion

In this paper we focus on and test sensitive data which can be protected from using a wide range options from back-up files encryption text, and lock plus wipe data. By using these methods and application was successful in protecting sensitive data from getting out in the open. It should not be too many problems to deal with sensitive data on the mobile device.

B. Future Work

There are some future problems in real world mainly in the business world. Since most sensitive data is protected from API. It is easily to avoid other users to get your information. However, it is difficult to tell how sensitive data will treat in the future and how it will change the mobile device. Future work will involve more and better methods on how to protect sensitive data on the mobile device by using more applications and different algorithms. We will focus on the applications such as Cosmos for Smartphones, and Super Backup for backup and protecting sensitive data. Also we will test if sensitive data can be protecting on mobile device using the cloud system.

Acknowledgment

This work has been supported in part by US. NSF grant # DUE-1241670 and US Department of Homeland Security Scientific Leadership Award grant # 2012-ST-062-000055.

References

[1] Inc, T.C., Data Transfer, 2014, http://www.pcmag.com/encyclopedia/term/40859/data-transfer

[2] SAS Institute, Data Transfer Services: Advantages, 2014, http://support.sas.com/documentation/cdl/en/connref/61908/HTML/default/viewer.htm#a000271140.htm

[3] University of Cincinnati., and Southern Polytechnic State University, 2013, https://sites.google.com/site/mobilesecuritylabware/1-threats-of-lost-or-stolen-mobile-devices/pre-lab-activity

[4] Dimensional Research., and Check Point Software Technologies LTD, 2014, http://www.checkpoint.com/downloads/products/check-point-mobile-security-survey-report.pdf

[5] Ponemon Institute, 2013, http://info.watchdox.com/rs/watchdox/images/WatchDoxWhite%20PaperFINAL2.pdf

[6] A. Lazou., and G. R. Weir, "Perceived Risk and Sensitive Data on Mobile Devices", UK. University of Strathclyde Publishing., pp. 183-196, 2011

[7] Ernst, & Young, Mobile devices security: Understanding vulnerabilities and managing risks, 2012, http://www.ey.com/Publication/vwLUAssets/Mobile_Device_Security/$FILE/Mobile-security-devices_AU1070.pdf

[8] G. Hughes, and C. A. Quinsey., Mobile Device Security, 2003, http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_049463.hcsp?dDocName=bok1_049463

[9] Developer Android, Android Security Overview, 2014, https://source.android.com/devices/tech/security/

[10] Developer Android, Data Backup, 2014, http://developer.android.com/guide/topics/data/backup.html

[11] University of Cincinnati., and Southern Polytechnic State University, 2013, https://sites.google.com/site/mobilesecuritylabware/3-data-location-privacy/lab-activity/cryptography/cryptography-mobile-labs/encryption-decryption/2-lab-activity/lab1

[12] Developer Android, Device Administration, 2014,http://developer.android.com/guide/topics/admin/device-admin.html

Architecture of a Mediation System for Mobile Payment

Boutahar Jaouad
EHTP,B.P 8108, Oasis,
Casablanca, Morocco

El Hillali Wadii
EHTP,B.P 8108, Oasis,
Casablanca, Morocco

El Ghazi El Houssaïni Souhaïl
EHTP,B.P 8108, Oasis,
Casablanca, Morocco

***Abstract*—Nowadays, the mobile phone has become an indispensable part of our daily. Exceeding the role of a communication apparatus, and benefitting from the evolution of technology, it could be used for several uses other than telephony, energy of photography, the geolocation, until the control of the health condition and the quality of the air, by measuring the cardiac pulsations, the temperature and the ambient content water. In this context, financial institutions wishing to take advantage of this wave of technological change and taking advantage of the telecom infrastructure robust and secure existing began to innovate to offer a new range of payment services based on mobile phone. Thus, in this article we present a proposal for an implementation of a mediation system of payment per mobile based on the technology of the Webservices.**

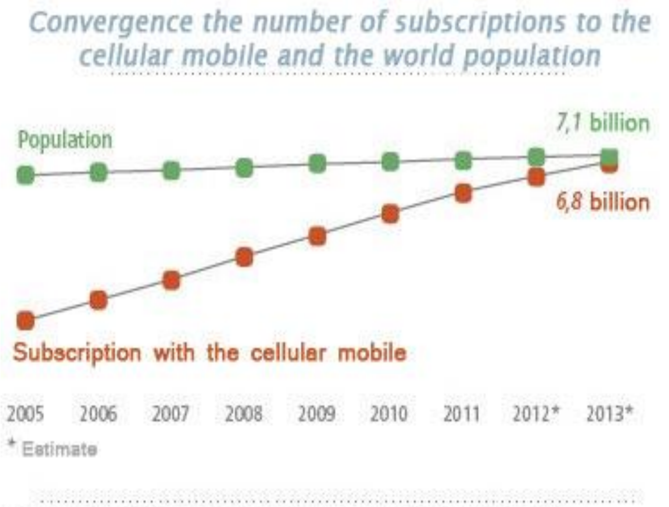
Keywords—M-paiement; M-Commerce; Android; Webservices; NFC; RFid

I. Introduction

Mobile technology knew a very strong evolution in the last decade, reserved at the beginning only with one privileged class, today the mobile phone is with the range of everyone[23,24,25]. International Telecommunication Union [1], estimates that in 2014 the mobile number of subscriptions will exceed the world population (Figure 1), in at the beginning of 2014 mobile number of subscriptions already reached 6.8 billion, that is to say a penetration rate of 93%. In the first quarter 2013, they are 425,8 million the number of mobile phones sold in the world, in rise of 0,7% compared with the sales of the first quarter 2012 [2].

The mobile phone exceeded the role of an apparatus of communication, the financial institutions wanted to benefit from the infrastructure telecom already present, to offer new banking services via the mobile. This started with the transfer of money per mobile, intended primarily for the countries in the process of development, allowing the people not having an bank account to carry out transfers of money per mobile, in addition a new concept saw the day, called the mobile banking or “Any where any time banking”, the bank in the vicinity making it possible to have the whole of the banking services offered by an agency via the mobile phone [3].

Also, with the growth of the sector telecom, of new technologies appeared, in particular the Internet on mobile with the 3G, the 4G, Wi-Fi Outdoor, and soon the WIMAX….



Source estimate: Database of the UIT on the indicators of ' telecommunication/TIC in the world

Fig. 1. Convergence of the number of subscriptions to the cellular mobile and of the world population

That supported the expansion of Smartphones. According to [4] the sales of Smartphones exceed those of the ordinary portables, thus exploding of 46,5% over one year, between April and June 2013, it was sold in the world more than 225 million Smartphones against 153 million one year earlier.

Vis-a-vis this revolution of mobile technology, a new form of purchase was born, the Mobile Commerce. The mobile phone became, thus, a system of payment offering a very high level technologies of safety compared to the traditional banking purchasing card [5], the manufacturers do not cease offering new means of safety such as the reading of the digital fingerprint or the recognition of the face of the owner, or technology NFC (NEAR Field Communication), allowing the payment without contact.

The banking main actors, in particular Visa, estimate that in the next years the mobile phone will replace the bank card, as means of payment, especially in the countries in the process of development. The bank card remains a means very vulnerable to the frauds, according to [6] the amount of the frauds with the bank card reached 1,55 billion euros in 19 European countries in 2013.

The service providers and the banks start to invest in the mobility solutions, by offering new services of payment per mobile, the use remains very limited, compared to the potential of this equipment. The service providers propose gravitational mobility solutions, but they are always based on the bank card like support of payment. Other shares the solutions suggested by the banks are only restricted with their customers, as these services are often limited to the payment of invoices (Electricity, Water, Telephone …).

After the comprehensive study of current systems of payment we propose to set up a central system of mediation based on the technology of Web services, allowing the customers to centralize the payment of the services of the suppliers to which it is adhered. For example the customer, will have the possibility to pay his invoices (Water, Electricity, Telephony.) in only one operation instead of going (the customer) on each site of the supplier. This operation will be carried out by the customer through a Mobile application (face-End), which will communicate with the Back-End part of the mediation platform.

A Mediation system makes it possible to have a uniform access for heterogeneous and distributed data sources. Figure 2., illustrates the general architecture of mediation systems, made up of a whole of data sources wrapped as a data source XML, and connected has a total interface (Integated Global View). Thus, the customer communicates with only one total data source instead of communicating with each base of give independently. The interfaces provide a uniform syntax of communication with the whole of the databases.

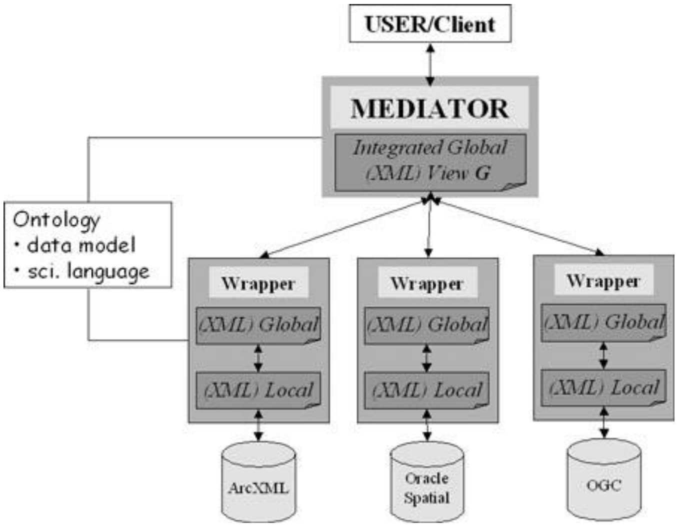


Fig. 2. Extended mediator architecture - Source: GEON: Toward a Cyber infrastructure for the Geosciences—A Prototype for Geologic Map Integration via Domain Ontologies

To give confidence to the client and the provider in our system, the operation of mediation is completely transparent to them, the platform leverages web services already offered by providers, the treatment is carried out at the platform at the time of client connection, it recovers in a single operation, all services payable by connecting to each supplier information system, after the confirmation of the customer, the platform is responsible for validating the payment to suppliers.

II. Related Works

Research on the architectures of mobile payment systems have inspired many researchers until today, it would be difficult to group the literature as a specific disciplines. Further evidence of this can be seen in the fact that the articles on mobile commerce and mobile payment are scattered across various journals in disciplines such as business, management, marketing, engineering, technology information (IT) and information systems (IS) [26].

The evolution of research on this field and reciprocal to that of mobile technology in the last decade with the advent of mobile searches have focused on the architecture of the payment systems and money transfer via mobile among researchers who are interested in this we find Jerry & Krishnaveni [17] (1), Ashutosh & Manik [27] (2).

(1) Jerry & Krishnaveni [17] propose a system of money transfer P2P-based mobile phone, set up a protocol for communication between the seller and the buyer, so they have developed a security strategy system implementation.

(2) Ashutosh & Manik [27] propose an architecture for mobile payments, to replace the credit card information by saving the EMV card chip on the SIM card.

On the other hand, and with the advent of the internet, and taking advantage of the implementation of the GPS system on the mobile phone, searches were interested in mobile commerce, in addition to the use of mobile phones as a means of replacing the credit card payment, a new layer is added to the purchase on the mobile.

Currently research is focused on contactless payment via NFC technology, this research found among those of Rahul & Shubham [28] propose to set up a platform for mobile payments based on NFC technology.

The common thread among most of the research is that they are always based on the existing banking network to make the payment or transfer of money, and they all focus on the payment interface part, the work that pluparts we met do not address the payment transaction from start to finish but they are only limited to the replacement of the card by the mobile phone.

III. Description of the Platform of Payment

The mobile operating systems were very well evolved; this made it possible to the merchants to propose richer and gravitational portable applications. Currently the majority of the merchants offer portable applications containing in addition to one window of the products, of the more relevant functionalities (points of fidelity, accounts - checks gifts, the geolocation and comparators of the prices…), this tendency is called the Mobile commerce, and it is regarded as an evolution E-commerce [7].

In addition, the customer thus finds himself vis-a-vis a whole of mobile applications installed on his smartphone, it is thus constrained to be identified with connection on each one of these applications, also for each operation of payment / purchase, it needs to obtain the information of his bank card, for example, for the payment of the invoices of electricity and of water, the customer is obliged to pay with that bank card for each service, knowing that comes from only one supplier. The operation of payment a remains complicated, and presents a risk of safety, one can say that until now, the applications of payment or purchase per mobile only of the E-commerce are not encapsulated in an portable application not exploiting the potential of the mobile phone, also, they are restricted to the people having a credit card.

To be more open, easier and more made safe, we set up a mediation system based on Web services allowing to centralize the proposed services by the suppliers and to ensure the payment of these services. The platform is connected automatically to the information system of the suppliers to recover the whole of the services to which the customer is registered, offering the possibility to the customer, in only one operation the payment of the whole of its services. The platform offers to its customers a virtual account containing his balance, enabling him to pay these services. That offers the advantage to the customers not having an bank account to also exploit the services of the platform. The platform is connected via channels of communication with the traditional systems of payment (interbank networks, systems of transfer of money) making it possible its customers to reload their virtual accounts. In addition the platform offers to the customers having an bank account to store encrypted information of the bank card in the data base of the platform, thus, and for safety reasons, the customer does not need more to seize them for each operation of payment, the platform as an intermediary, will undertake to restore this information in a protected way.

The main advantages of the platform rest on the following aspects:

1) *F*acility of use:

- Facility of adhesion (Customer/Supplier).
- Ease of payment.
- Dematerialization of the invoices paper in those electronic.

2) Opening:

- The platform does not suggest any restriction for adhesion.
- The platform makes it possible any supplier to propose the payment of its services to its customers.
- The platform is connected to all the interfaces of traditional payment.

3) Made safe:

- The platform respects the international security standards in particular the PA-DSS.

A. Use of the platform

The customer with the possibility of exploiting the services of the platform via two channels:

1) Via a Mobile application:

The customer with the possibility of exploiting the services of the platform on its mobile phone, in particular the creation of the account, the payment of the invoices of its services, the notification of the invoices.

2) Via an E-commerce website :

The platform can be useful like channel of payment on the E-commerce website of the merchant, that the customer can choose at the time of the payment, thus the customer can choose to pay is by virtual account or of its credit card via the platform.

B. Component of the platform

The platform consists of three essential layers (Figure 3):

1) The mobile customer: a multiOs portable application installed on the customer smartphone allowing exploiting the services of the platform.

2) The mediation system: a central server playing the part of offering federator of services proposed by the providers of the services exploitable by the customer and the suppliers. Having interfaces of communication with the traditional networks of payment. In addition to the operations of payment, the platform must be able simultaneously to deal with the communication with the various heterogeneous information systems from the suppliers and to be able to treat and to synthesize their various structures.

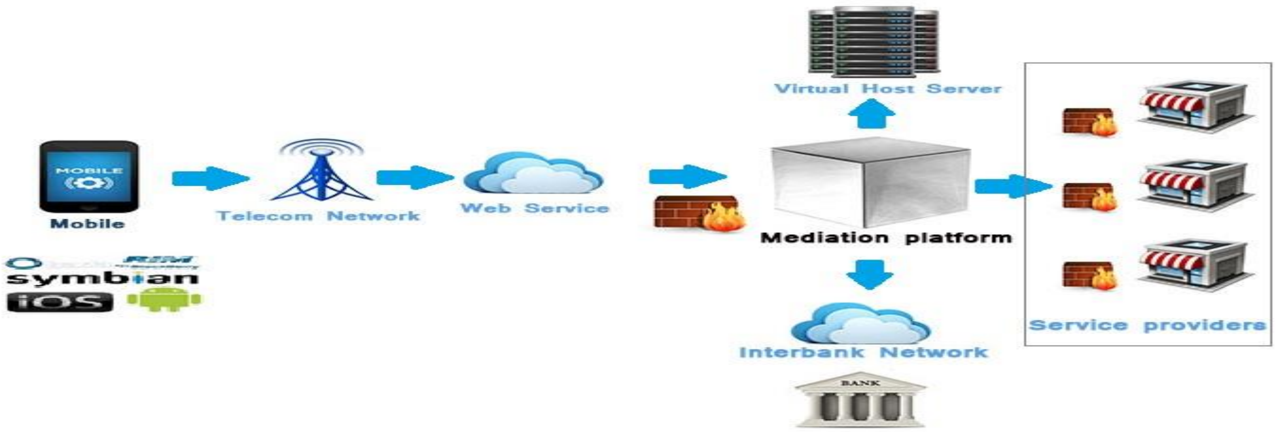


Fig. 3. Global description of the mediation platform

3) Partners: the principal partners of the platform are the suppliers of the traditional services (Electricity, Water, Telephones, Internet, Taxes ...), these partners are connected to the platform via adapters.

C. Description of the platform services

1) Creation of the account user:

To be able to exploit the platform services, the customer starts by creating an account by seizing his personal informations via the mobile application, the administrator of the platform validates the creation of the account, a virtual account is created automatically in the Host Server "Virtual Host Server" containing information of the balance of the platform customers, the account user is identified by a Single code UIC "**Unique Identifier of the Client**", it will be the support of principal payment of the user.

Once connected by the mobile application, the customer starts by choosing the list of the services for which he wishes to adhere, an authentication near the supplier is required to validate the adhesion of the customer to the service. The platform automatically retrieves all services payable on suppliers from the user login, the customer also receives a notification from the presence of an invoice for a service to be set.

2) Reload Virtual account:

During the creation of the account user, the platform offers to the customer to reload his virtual account by:

- Credit card: The customer with the possibility of reloading his virtual account by using its bank card, via interfaces of the platform with the interbank network.
- Transfer of money: the platform is connected to partners of transfer of money via Web services, making it possible to feed the virtual account of the customer in real-time at the time of a transfer operation of money.

3) Recording of the banking informations:

We designed our platform so that it is independent of the banking sector to allow people without bank accounts to access payment services via their mobile. The interface with the banking or money transfer structures is useful only to reload the customer's virtual account, if not all payment transactions will go through the virtual account platform. Storage of credit card information is only option offered by the platform for customers to use the platform as a means of payment instead of entering their data on the website of traders because of security.

In accordance with the PA-DSS, the information from the credit card (BIN, CSC, expiration date of the card) is encrypted before being stored in the database, for against it shall store the PIN . When a payment transaction, the customer can choose the payment channel or through virtual account or by bank account. If paying by credit card, the platform decrypts data and validates the payment.

IV. Platform Architecture

Among the great difficulties of taking into account in the installation of a payment system is the response time, especially for the systems connected to the interbank network. The leaders of the electronic money impose an optimal response time for the treatment of the transaction, VISA for example obliges a response time less than 10 seconds if not the transaction will be rejected by a answer code ''TIME OUT'' (BFFF0015) [8], to respect the banking standards, the architecture of our mediation system of payments will be built in two levels :

- "**Front office**" part: During the validation of the payment by the customer, this part takes care of the checking of the balance customer, confirmation of payment to suppliers, the payment and the generation and the recording of the journal of the transactions in the platform database, and the emission of a receipt to the customer.
- "**Back Office**" part: This part takes care of the consolidation operations between the platform and the suppliers; this concept is inspired from the banking environment called operations of clearing, allowing the interbank consolidation [9]. These operations are generally done in end-of-day in order to not occupy the performances of the production servers; it acts to generate the reports/ratios of accountancy while being based on transactions journals. These reports/ratios will be the support of suppliers payment starting from the clients' account.

A. Description of the Components of the platform

The mediation system platform respects a modular architecture, where its components are independent and communicate between them (Figure 4).

The interface "**Mobile Interface**" is charged to treat all the requests of the customer by mobile, the request for connection passes by this interface, by calling the module "**Authentication**" this module starts by authenticating the customer by basing on this information stored in the database "**Account Customer**", after the authentication of the customer, this module checks via the module "**Access Control**" the services to which the customer has the right to reach, this module recovers the rights of the customer while being based on information of the database "**Platform services**".

Once connected, the customer with the possibility of listing the invoices to pay via the module "**Service Management**". This module carries out a connection to the unit of the services to which the customer is registered to recover the invoices automatically to be paid. This with the advantage of avoiding with the customer explicit connection to the platform of each supplier to pay his invoices, but the platform centralizes the whole of these subscriptions, the communication with the suppliers passes via the interface "**Partner Gateway**" by using the Web services technology.

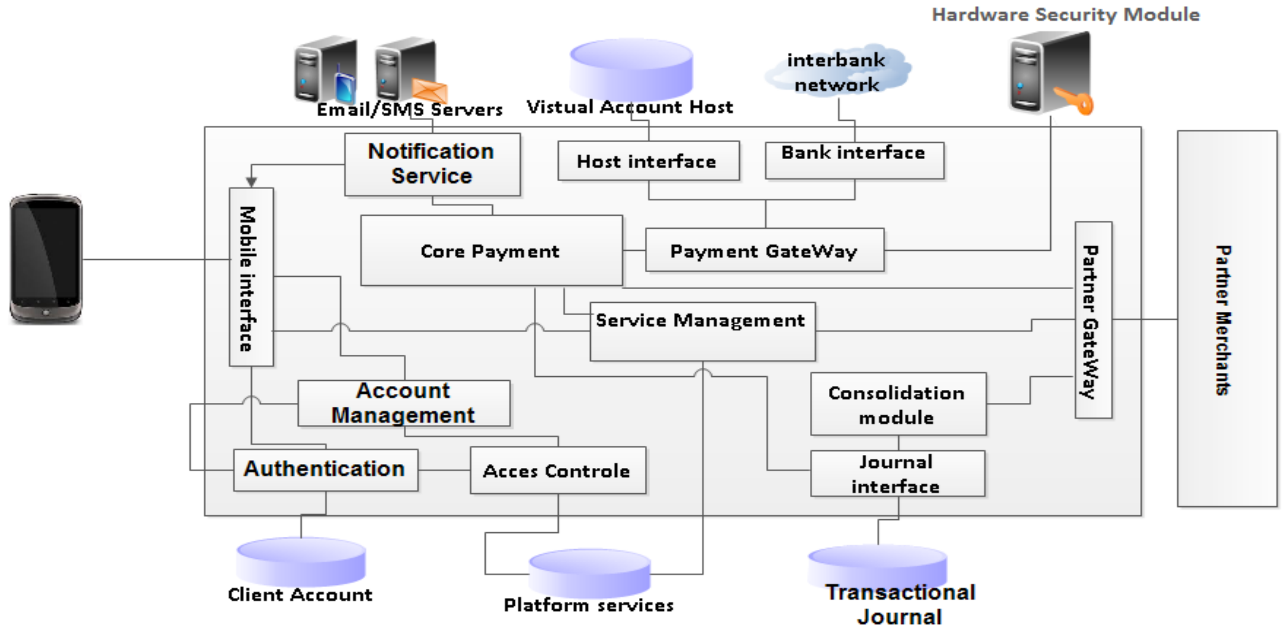


Fig. 4. The platform architecture.

The customer selects the list of the invoices to be paid and validates his payment, the module "**Service Management**" calls upon the module of payment "**Core Payment**" this unit deals with the operation of payment, it starts by calculating the amount of the transaction, and proposes to the customer the interface of payment wished via the module "**Payment Gateway**", if the customer chooses his credit card as payment medium, this last will pass by the interbank network via the interface "**Bank interfaces**", if not the amount of the transaction is debited from the customer' virtual account "**Virtual Account Host**" via the interface "**Host interfaces**".

The module "**Core payment**" validates the payment with the supplier and records the report of the payment transaction in the database "**Transactional Journal**" via the module "**Journal Interfaces**", once the report of the transaction recorded the module "**Core Payment**" informs the customer of the result of the transaction via the module of notifications "**Notification Service**" by SMS and transmits to the customer email address the electronic invoice.

The module "**Consolidation Interfaces**" takes care of the generation of the reports of the whole of the operations of day laborers payment generated via the platform while basing itself on information of the base "**Transactional Journal**", this information will be used like support of compensation thereafter.

V. The Platform Modeling

The platform modelling is carried out in two parts:

A. *The first part:*

The first part concern the mobile customer, it describes the whole of the functionalities suggested for the customer via the mobile (Figure 5) and which can be gathered in four essential modules:

1) The inscription at the platform: This module gathers the whole of the functionalities of inscription at the platform, starting by seizing these personal informations, the choice of the services and the methods of payment, this information will be transmitted to the administrator for the final validation of the creation of the account.

2) The account Management: This module covers the whole of the spots of administration of the account to knowing the consultation of the balance of the virtual account, the food of the virtual account by (transfer/Bank card), the personal modification of information, also it makes it possible to the customer to safeguard information of its bank card (the PIN code, the CSC, and it scratch date).

3) The suppliers/services Management: This module allows the management of the services suppliers partners of the platform, it makes it possible to the customers to adhere to the the various services suppliers.

4) The management of the services: This module takes care primarily of the payment of the services, it recovers the list of the services to which the customer is registered, it is connected to the suppliers servers to recover the services to be regulated, it also makes it possible automatically to notify the customer of the presence of the possible services to be regulated.

B. *The second part:*

The second part relates to the mediation platform (Figure 6), the modeling of this part is structured in three parts:

1) Implements customer services: This part covers the implementation of the Web services invoked by the customer via the mobile application, and which are the list of the services, the management of the account the inscription at the platform, the management of the services.

Implements customer services

2) Administration of the platform: This part covers the functionalities suggested with the administrator of the platform being an actor of the platform; these functionalities set out again in three parts which are, the management of the customers in particular the validation of the creation of the account, the management of the suppliers, and the follow-up of the transactions.

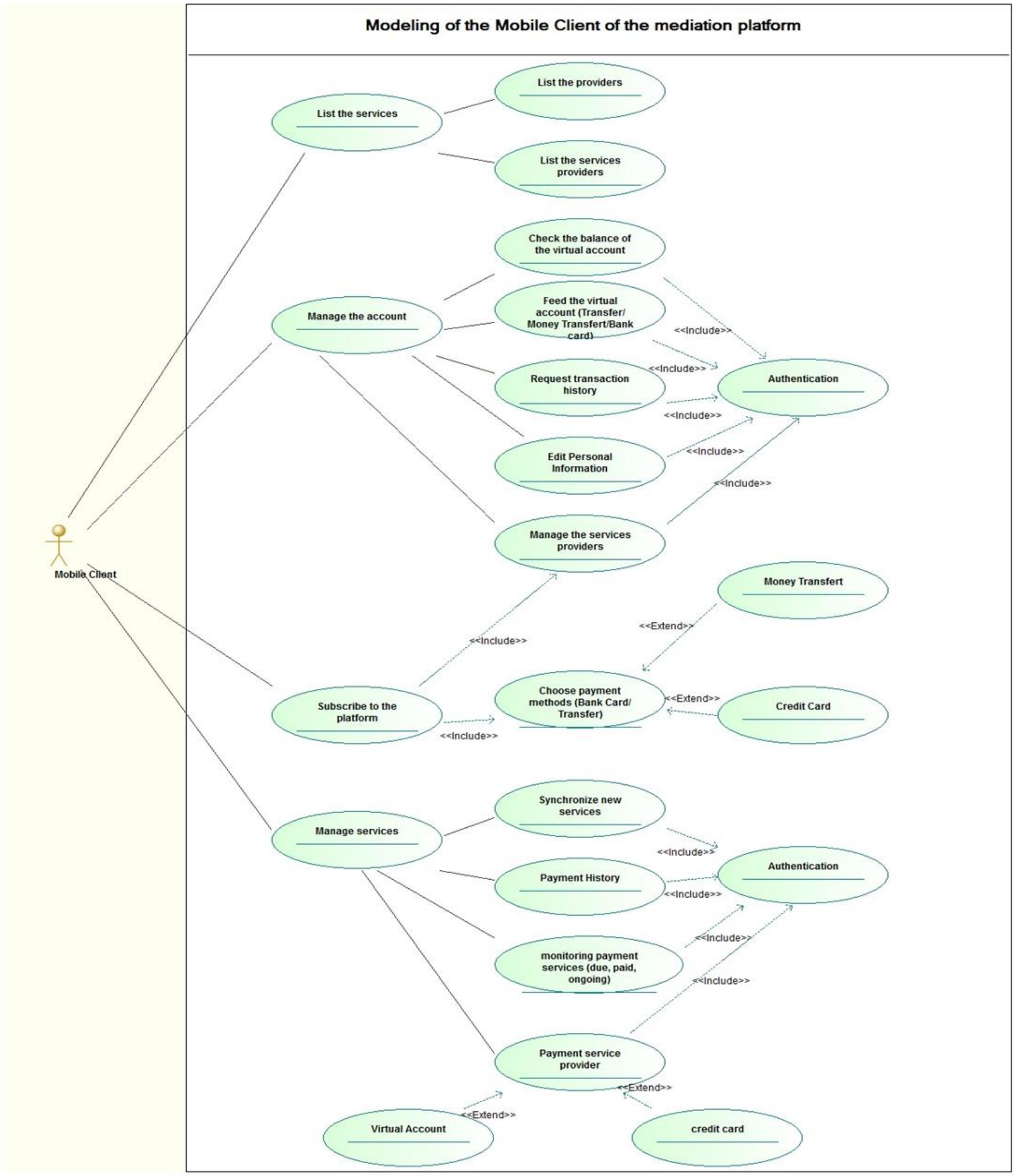


Fig. 5. Uses cases of Mobile application part.

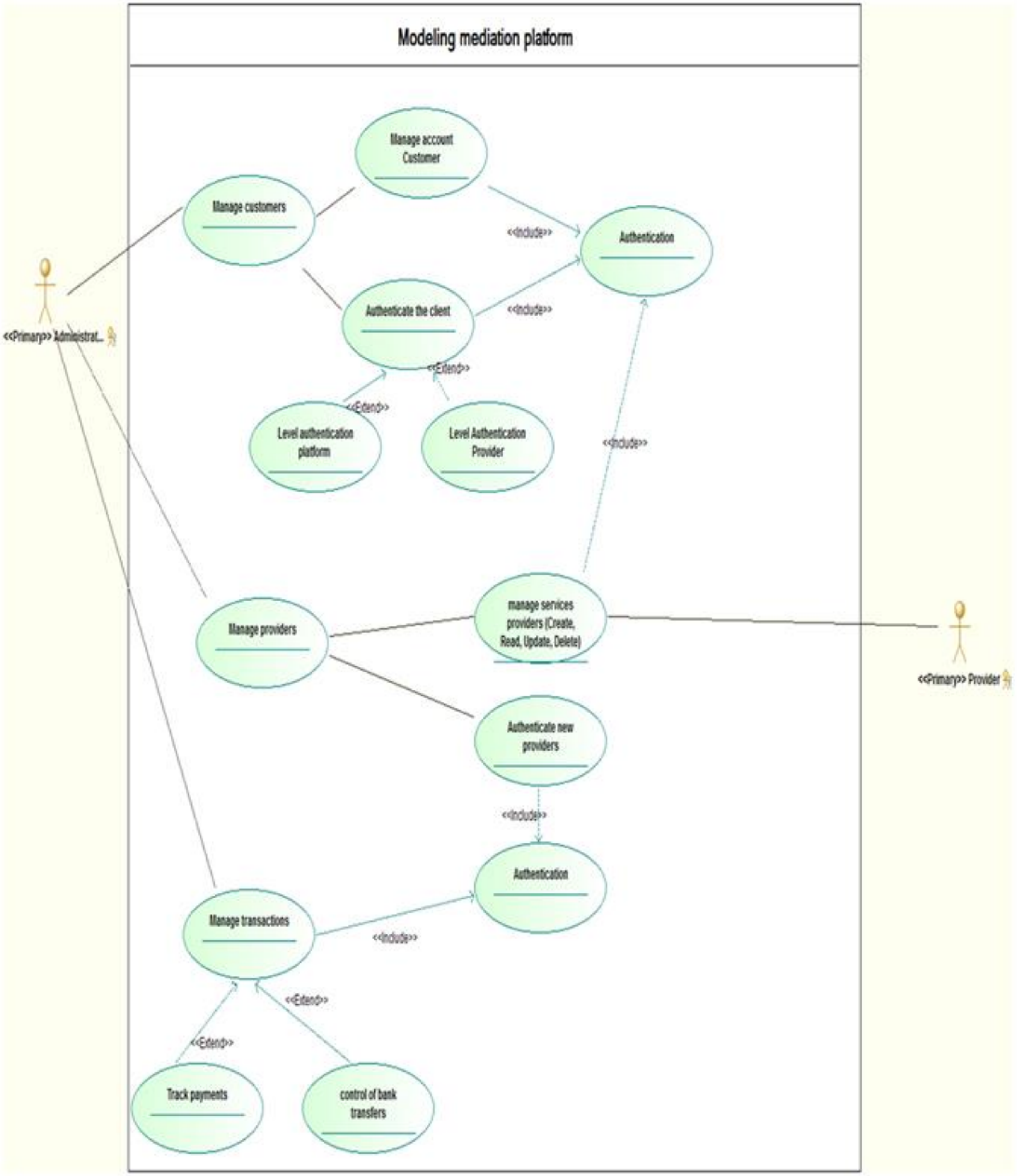


Fig. 6. Platform Uses cases.

VI. Technologies

A. *The Mobile custome*

To be able to cover a great mass of customers, the customer should be accessible via all mobile technologies.

We noted a very relevant point at the time of the choice of the mobile technology, which is mobile Frameworks of development, such as PhoneGap [10], and jQuery Mobile [11], and of others, generally based well on technologies HTML5, and JQuery, they implement the notions of the MDA (Model Driven Architecture) for the automatic generation of the source code, this starts by creating the model of the application, Framework is automatically given the responsability to generate the source code of the application for the whole of the operating systems mobile (Android, IOS, RIM, Windows Mobile…), this with the advantage of reducing considerably the time of development of mobile applications, also this reduces the complexity of development. One is obliged to have forcing of competences on all the Mobile computer programming languages.

In addition, we noted that these Frameworks are still flowering and cannot reach the whole of the native resources of the mobile. We intend to evolve/move our platform to support state-of-the-art technologies such as the NFC, and these Frameworks are likely to pose a blocking.

We chose to develop the Mobile customer under Android technology in a first place; we chose this technology considering it is most dominant compared to other mobile technologies.

The mobile operating system Android in first position with 79% of market share is followed of iOs with 14,2% of market share (Figure 7).

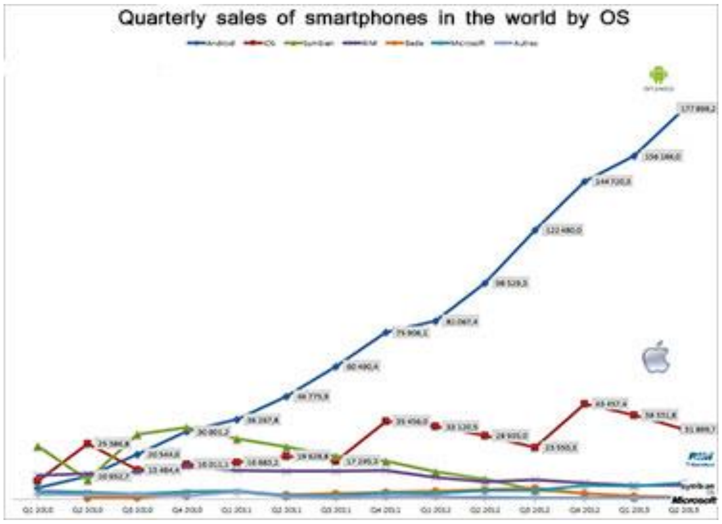


Fig. 7. Evolution of the market shares of the bones for smartphones (Android, iOS, Windows Mobile, RIM) # Gartner

1) The Web mobile services : The access to the Web services of the mediation system via the mobile is carried out using bookstore KSOAP2 [12], it provides a library of light and effective customer SOAP for the Android platform. It is Fork of the library kSOAP2 which is tested mainly on the Android platform, but should also function on other platforms using of the libraries Java [13].

B. *The mediation system*

For the development of the platform, we chose the J2EE technology in particular the JSP/Servlet, which is the standard of development of the applications of distributed companies, guaranteeing the robustness the evolution and safety [14], the richness of the bookstores and technologies and the weak blow of developments.

The application server used to implement the platform is the Tomcat server, which is a server of application supporting the JSP /Servlet [15]. Webservices are created by using Axis, the Apache solution, which is the Open Source container of Webservices more dominating [16].

1) The Web services SOAP : Although technology REST (REpresentational State Transfer) is more popular in the fields of the developments Web/mobile (Figure 8) considering its simplicity to implement, it was even adopted by the large firms of the Web world such as Amazon instead of SOAP, it is well-known that it is limited to the level safety.

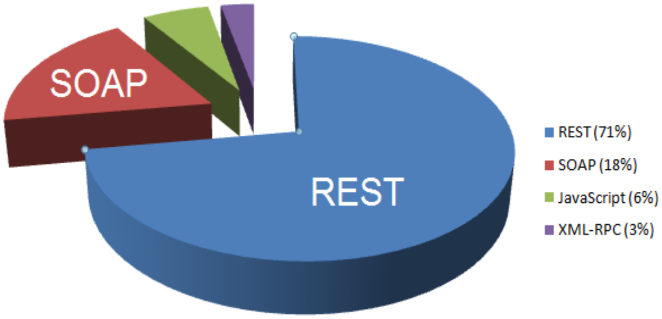


Fig. 8. Classification API web development. February 2012.

Technology REST is used mainly to make easy the access to the resources on the Web, while being based on protocol HTTP, via verbs which are defined by methods HTTP like DELETE, GET, POST and COULD, and it is from there that it holds are success, nevertheless although this technology was adopted by large firms of the Web such as Amazon [18], it still presents some limits. REST is a synchronous protocol and without state. When a customer subjects a request to the server, it does not have the means to know if its request were received automatically, but it is the server which must create a new task to inform the customer [19]. In our case, the platform is used especially for the operations of payment, therefore we need a robust and made safe technology, and this is why we chose an architecture SOAP.

2) Technology NFC : Near Field Communication, is a wireless technology for short-range use the mobile phone as a payment close [20].

In the short term we have used this technology to:

- Strong client authentication via an NFC card:

The platform offers customers the ability to use NFC as a method of physical security.

At the time of account creation, at the Security menu, the client begins by asking the setting of its new NFC card, the platform generates a key using HSM "Hardware Security Module" which will be sent to the customer's phone, the mobile application loads programming the Tag in the NFC card, the customer can choose the NFC card as the only means of authentication.

Therefore, the connection to the platform becomes easier because the client no longer need to enter their email address and password to log in, but just simply put his card on his NFC phone.

On the other hand at the time of payment services, the platform requires the client to validate his basket with NFC card.

In this way, we reduced the security risks for theft of telephone or mobile phone use by a third party.

- The transfer of money between customers:

The second use of NFC card money transfers between customers of the platform, or payment at the point of sale.

The money transfer is done via the mobile application, by entering the amount to be transferred, to validate the customer only has to ring closer its phone the recipient to complete the transaction. The mobile application retrieves the identifier of the recipient on this NFC card or Mobile phone (supporting NFC) and transmits to the platform a request for money transfer with the amount and beneficiary identifier.

VII. THE ASPECT SAFETY

To guarantee the security of our system, we propose to use PAD-DSS standard and the material potential of the smartphones in terms of safety:

A. PA-DSS standard

Currently, the financial institutions are conscious of the risks of safety, of the new standards were forced to reduce the risks of safety of the applications of payment in particular the standard PA-DSS, which is the program managed by the Council supervising before the program of Visa Inc. The goal of the PA-DSS is to help the suppliers of software and others to develop protected applications of payment which do not memorize data prohibited, such as the complete magnetic bands, the CSC number ("Card Security Code") or the data of PIN (“Personal Identification Number”), and to make sure that their applications of payment are in conformity with NCV DSS [21].

This standard has been implemented in our platform by:

1) Log: Avoid drawing confidential information in the logs to prevent any holdings of these data by an administrator.

2) The credit card information: stored on the platform are encrypted using RSA method, and they are hidden at the time of posting to the administrator.

3) OWASP implementation: At the platform level the possibility of the OWASP implementation library providing a source code open source to avoid the Ten Application Security Risks of the Most Critical Web. Ranked by various security agencies (DoD, PCI Security Standard)

B. Material

In addition, we will exploit the last tendencies of the mobile potential to make more safety and robust our solution, in particular, by using, the identification by digital fingerprint, the physical identification by NFC card, and the geolocation.

C. Prevent money laundering

To avoid bleaching and embezzlement, we propose to set up a data mining system to detect suspicious behavior based on data from the database "Transactional Log" by complying with rules such as:

- Do I know the customer?

- Is the transaction is consistent and compatible with the habits of the customer?

- This transaction is logical?

On the other hand the system checks the transaction amount, the position of the client and geolocation services purchased to control any possible fraud operation.

D. Two factor authentication:

The disadvantages of using NFC card for authentication, is that this key should be transported anywhere with the customer, if the customer risk for loss of the card not being able to access their account.

So for customer recognition, we have designed several authentication methods, taking into account available on the mobile phone technology.

The simplest method of authentication is the Email address and password, this mode presents security risks if a third party had access to this information, another layer of security that can be added to this mode is checking imei code of the phone or the phone number to see if it is indeed the owner of the account to make the payment.

NFC card: the platform provides a physical means which is the NFC card to authenticate the client, which in case of theft of the phone, the thief need this card to make these payments. setting the card passes through the customer's mobile phone during setup and the customer has the ability to generate a pin code on its NFC card in order to get his account for the loss of NFC card.

The fingerprint: We are currently testing a new API called Fingerprint_SDK, introduced by Samsung for its Galaxy mobile phone S5 enables developers to exploit the fingerprint chip in their projects. This API provides default parallel fingerprint the possibility for the user to enter a PIN if there are problems with the fingerprint authentication.

VIII. CONCLUSION AND PROSPECTS

This is a proposal for a mediation system based on the mobile phone, we noticed that the market of the smartphones gains ground in the next future years and we want profited from his potential to offer a solution has low costs, while guaranteeing the standards of performances and safety required in this field. Currently we are in the final phase of the development of the platform, the model under development and limited only to the payment of bills, knowing that the use of the platform is able to support other services. On the other hand we have limited the use of NFC technology for authentication and transfer of money, we are studying the prospects of using this technology in the platform, to extend it to the automation purchase tickets by NFC or RFID (Radio Frequency IDentification) [22]. We also plan to test security with robots to test the behavior of the platform against attacks and transactional response time. These results will be published in another article soon.

REFERENCES

[1] International Telecommunication Union,May 2013,WTID 2013.

[2] Gartner,"Worldwide Mobile Device Sales to End Users by Vendor",May 2012.

[3] Infogile Technologies, "Mobile Banking – The Future", August 2007 .

[4] Gartner "Market Share Analysis: Mobile Phones, Worldwide, 2Q13.".

[5] Ravi Tandon , Swarup Mandal and Debashis Saha, M-Commerce-Issues and Challenges.

[6] FICO by Euromonitor International http://fico.com/landing/fraudeurope2013/,2013.

[7] IACSIT, "Mobile Commerce and Related Mobile Security Issues »,2011.

[8] Agilent Technologies, Agilent VISA User's Guide".

[9] BANK FOR INTERNATIONAL SETTLEMENTS, "Payment, clearing and settlement systems in the CPSS countries Volume 2" ,November 2012.

[10] PhoneGap website, http://phonegap.com/.

[11] JqueryMobile website, http://jquerymobile.com/.

[12] KSOAP Poject website,http://kobjects.org/ksoap2/index.html.

[13] KSOAP for android,https://code.google.com/p/ksoap2-android.

[14] John Roth, SAS Institute Inc., Cary, NC "Configuring J2EE Application Servers for Use with the SAS BI Platform".

[15] Apache Tomcat Website, http://tomcat.apache.org/tomcat-5.5-doc/index.html.

[16] Beytullah Yildiz, Telecommunications, 2006. AICT-ICIW '06. International Conference on Internet and Web Applications and Services/Advanced International Conference on, "Experiences in Deploying Services within the Axis Container".

[17] Jerry Gao, Krishnaveni Edunuru, Jacky Cai, and Simon Shim, IEEE International Workshop on Mobile Commerce and Services (WMCS'05),"P2P-Paid: A Peer-to-Peer Wireless Payment System".

[18] Tim O'Reilly, O'Reilly," REST vs. SOAP at Amazon".

[19] National Security Agency USA," Guidelines for Implementation of REST" ,2011.

[20] Tom Igoe, Don Coleman & Brian Jepson , O'Reilly,"Beginning NFC: Near-Field Communication with Arduino, Android, and PhoneGap" ,2014.

[21] Pci Security Standards Council, "Payment Card Industry (PCI) Payment Application Data Security Standard (PA-DSS) " , 2014.

[22] Mabel Vazquez-Briseno, Interactive Multimedia Edited by Ioannis Deliyannis, "Using RFID/NFC and QR-Code in Mobile Phones to Link the Physical and the Digital World" , March 7, 2012.

[23] T. Kippenberger, Fasten your seatbelts, The Antidote 5 (1)(2000) 38–39.

[24] S. Kumar, J. Stokkeland, Evolution of GPS technology and its subsequent use in commercial markets, International Journal of Mobile Communications 1 (1/2) 2003) 180–193.

[25] H. Vogt, F.C. Gartner, H. Pagnia, Supporting fair exchange in mobile environments, Mobile Networks and Applications 8 (2) (2003) 127–136.

[26] E.W.T. Ngai, A. Gunasekaran , July 2005, "A review for mobile commerce research and applications", International Journal of Business Innovation and Research.

[27] Ashutosh Saxena, Manik Lal, Das Anurag Gupta, IEEE International Conference on Mobile Business (ICMB'05) "MMPS: A Versatile Mobile-to-Mobile Payment System".

[28] Rahul Gaikwad, Mr. Shubham Chaudhari, Ms. Dhanwanti Gaikwad, International Journal of Electrical and Electronics Engineering (IJEEE) 2011, "An Integrated Mobile Phone Payment System Based on 3G Network"

Using Social Signal of Hesitation in Multimedia Content Retrieval

Graphical Analysis of Selection Traces in the Matrix-factorization Space of Multimedia Items

Tomaž Vodlan
Agila d.o.o.
Ljubljana, Slovenia

Andrej Košir
Faculty of Electrical Engineering
Ljubljana, Slovenia

***Abstract*—This paper presents the graphical analysis of selection traces in matrix-factorization space of multimedia items. A trace consists of links (lines) between points that present a selected item during interaction between user and video-on-demand (VoD) system. User used gestures to select from among video on screen (VoD service), while additional user-produced social signal (SS) information was used to recommend more suitable new videos in the process of selection. We used a sample of 42 users, equally split into test (SS considered) and control and random (SS not considered) user groups. We assumed, for each user, there are areas of multimedia items in the matrix-factorization space that include preferred user items, called preferred areas. The results showed that user selection traces in the space of multimedia items (matrix-factorization space) better covered the user's preferred areas of items if the SS of hesitation was considered.**

Keywords—Human-computer Interaction; Social Signals; Hesitation; Matrix Factorization; Video-on-Demand; Graphical Analysis

I. Introduction

State-of-the-art research in human-computer interaction (HCI) ignores the user social behaviour, therefore the user interaction with the system is still not completely user-friendly experience. Social signal processing [1, 2, 3, 4] is a research domain that aims to understand social interactions through machine analysis of nonverbal behaviour [4]. Social signals (SSs) are initiated by the human body and present reactions to current social situations. They are expressed with nonverbal behavioural cues (e.g., gestures, postures, facial expressions, etc.).

One example of how SS can be used in HCI is a manual VoD system with a conversational recommender system (RS) where the user selected one video clip among several presented on the screen [5]. The system adjusted the list of video items to be recommended according to the extracted SS class {hesitation, no hesitation}. SS of hesitation was used because is commonly manifested when a user is faced with a variety of decision choices. The results of this study showed a significant difference in user satisfaction with the system between group for which the SS was considered and group for which the SS was not considered [5].

In this paper we present the results of graphical analysis of selection traces in matrix-factorization (MF) space of multimedia items. At each step user selected one video on screen (one point in MF space). A line links two consecutive selected videos (points). In that way we got selection traces for all interactions. Graphical analysis was based on two assumptions (i) the MF space of multimedia items is the best possible layout of multimedia items for all users and (ii) for each user, there are areas of multimedia items in MF space that include preferred user items, called preferred areas. We compared traces between group for which the SS was considered (test group, 14 users) and groups for which the SS was not considered (control group, 14 users; random group, 14 users). The results indicate that the use of the SS of hesitation in our VoD system provides better coverage of the user's preferred areas of multimedia items in MF space, resulting in better user satisfaction with the system.

The reminder of this paper is summarized as follows. Section II provides experimental design, experimental user scenario and additional explanations of the selected aspects of the experimental design. Section III describes the evaluation methods that were used, while the evaluation results are presented in Section IV. A discussion of the evaluation results are provided in Section V. Section VI concludes the study.

II. Experimental Design

We modelled an independent-measures experimental design and an associated experimental user scenario for the evaluation of SSs in HCI in an example where users use gestures to select from among videos on a screen (VoD service) (experimental design and user scenario are briefly described in [5]). Our experimental design allows the control of the effect of the SS expressed by the user during an interaction with the system and the control of other possible causes of differences in quality of experience (QoE) among tested users to reliably estimate the contribution of the use of the SS to the QoE. The experimental design allowed a fair comparison among test, control and random groups. A human operator provided a baseline for real-time action recognition and SS extraction. The main reason why we used a human operator was to avoid there being a new uncontrolled parameter in our design since the results obtained with current state-of-the-art automatic gesture-recognition algorithms still include errors. The human operator observed the user via a camera and reported his/her decisions through a human-operator interface.

The experimental user scenario was a manual VoD system with a conversational RS, where the user selected one video clip from among several presented on a screen (television) through a VoD user interface. The system adjusted the list of the video items to be recommended (RS) according to the extracted SS class {hesitation, no hesitation} and selected item. All scenario description below refer to the test user group. If the user is not hesitating, the system displays three similar items in addition to the selected one. If the user is hesitating, the system then displays four diverse items according to the items on the current screen. The new items are projected on-screen with sound feedback, which indicates how the system recognized the user's SS. The user repeats the selection process until he/she finds the item he/she wants to watch. When the user indicates with a gesture that the final decision has been made (i.e., the user selects the item he/she wants to watch), the system expands the selected item (video) to the whole screen and turns on the sound of the video. The user watches the selected item for about 20 seconds. To detect if the user was hesitating, we used hand movements, eye behaviour and the time between two selections. Before and after interaction with the system, the user fills in pre- and post-interaction questionnaires.

The scenario for the control user group and for the random user group is almost the same as for test group. The only difference is how the system presents new items to the user. In the control user group, the system provides three similar items related to the initially selected item (the decision of the system is based only on gestures for video selection). In the random group, the system randomly provides similar and diverse items related to the initially selected item. In this way, we ensure that any difference between the test and control groups of users and test and random groups of users is not only a consequence of the use of different selection functions.

Selection of the most significant behavioural cues that describe SS class {hesitation, no hesitation} was based on methodology presented in [5]. We obtained the best results by combining four features (three behavioural cues and one automatic feature (time)) and logistic regression as classification algorithm. These four features are: (a) the user watching video content, which is then selected for a longer viewing time, (b) the user making a quick gesture when selecting video content, (c) the user watching all video contents, but none for a longer time, and (d) the time between two selections. The proposed model was then used for the design of a human-operator interface through which the human operator reported his/her decisions about the extracted SS class and recognized gesture of selection.

A. *Selected Aspects in Experimental Design*

This paper focuses on graphical analysis of selection traces in the MF space of multimedia items, therefore we briefly describe the conversational RS, MF space of videos and video selection function in sub-sections below.

1) *Conversational Recommender System and Video Database*

A conversational RS with no previous knowledge about the user was used. Functions *getInitialItems()*, *getSimilarItems()*, and *getDiverseItems()* (see subsection below) were based on selected videos from the LDOS-CoMoDa research dataset [7] and MF-based recommender algorithms [8]. However, we did not use all videos from the LDOS-CoMoDa dataset. Our subset contained over 300 videos (movie trailers). All the videos had the same display resolution (632 x 274 pixels) and were in the same multimedia format. The minimum length of a video was 60 s. The distance between movies was computed in a two-dimensional space generated by the first two factors of the MF algorithm presented in our previous work [9] and briefly below.

2) *Matrix-factorization Space of Videos*

Input data for the RS were presented as a sparse matrix in two dimensions, where the first dimension represented users and second dimension items ($r_{u,i}$). Data of the matrix were item ratings; specifically, explicit user feedback taken from the LDOS-CoMoDa research dataset, which has more than 3600 ratings given by 150 users. The goal of the MF method is to explain ratings in the r_{ui} matrix by characterizing both items and users with factors inferred from the rating patterns. MF models map both users and items to a joint latent factor space of dimensionality *f*, such that user–item interactions are modelled as inner products in that space. Each item *i* is associated with a vector $q_i \in D^f$, and each user *u* is associated with a vector $p_u \in D^f$ (1) [8].

$$\hat{r}_{ui} = q_i^T p_u \tag{1}$$

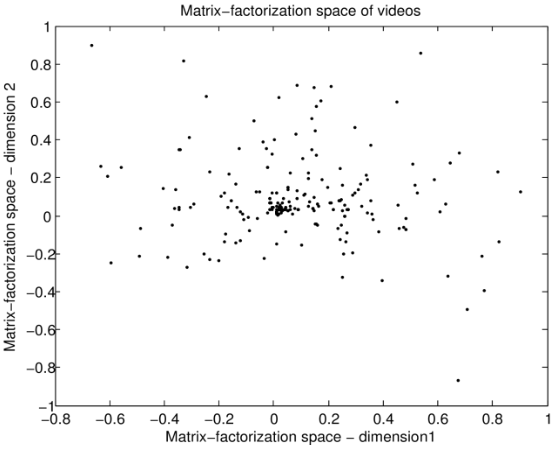


Fig. 1. MF space of videos. Each video is represented by a point in the two-dimensional MF space

The main challenge is the computation of the mapping of each item and user to factor vectors $q_i, p_u \in D^f$. In our case, the stochastic gradient descent approach [8, 10, 11] was used. We computed the factor space in two dimensions (*f = 2*) and each multimedia item was therefore presented as a point in two-dimensional MF space (Fig. 1).

3) *Video Selection Functions*

Employing our testing scenario (see Section II), videos were provided to the user according to the SS produced by the user. The VoD system simulates an event in the video rental store or at home. The user wishes to get a video, but is not sure which one. The support person provides the user with four videos (items) and the user expresses an opinion. If the user hesitates when selecting one item, four completely new items are provided. If the user does not hesitate when selecting one item, the selected item remains and three similar items are added. The selection procedure is repeated until a final selection is made. Therefore, we need three video selection functions provided by the conversational RS:

$$\left[hA, hB, \mathrm{hC}, \mathrm{hD}\right] = getInitialItems() \,, \tag{2}$$

$$\left[hS, hA, hB, hC\right] = getSimilarItems(hS, h1, h2, h3) \,, \tag{3}$$

$$\left[hA, hB, hC, hD\right] = getDIverseItems(h1, h2, h3, h4) \,. \tag{4}$$

Function *getInitialItems* ((2), Algorithm 1) provides four videos for the first screen, where the videos cover the whole MF space.

Function *getSimilarItems* ((3), Algorithm 2) provides four videos that are similar to *hS* (the selected video); one of them is *hS*. This narrows the search area.

Function *getDiverseItems* ((4), Algorithm 3) provides four videos that are not similar to *h1*, *h2*, *h3* and *h4*, which expands the search area. The function should diversely cover all of the factorized video space except the areas covered by *h1*, *h2*, *h3* and *h4*. The distance metric measuring similarity among movies is based on the MF space.

Algorithm 1: getInitialItems

```
   Input parameters:

   ComSub: matrix of all videos in our subset of
           LDOS-CoMoDa research dataset (videos ID
           and coordinates)
   n: number of items that are being looked for

   Output parameters:

   nIDs: vector of IDs of the selected items

Randomly select n videos from the ComSub subset
(using function rand in Matlab).
```

Algorithm 2: getSimilarItems

```
   Input parameters:

   vecC: vector of IDs of currently playing videos
   selID: ID of selected video
   vecE: vector of IDs of already played videos
   n: number of items that are being looked for
   ComSub: matrix of all videos in our subset of
           LDOS-CoMoDa research dataset (videos ID
           and coordinates)

   Output parameters:

   nIDs: vector of IDs of the selected items

Create a subset of IDs from which similar items
can be searched (SimSub). The subset does not
include IDs of already played videos (vecE) and
currently playing videos (vecC).
for all IDs in SimSub do
    Compute Euclidean distance between selected
    video ID (selID) and all video IDs in SimSub
    (use coordinates from ComSub).
end for
Put the distances in the order from smallest to
largest. Select first n videos (IDs) with the
smallest distance.
```

Algorithm 3: getDiverseItems

```
   Input parameters:

   vecC: vector of IDs of currently playing videos
   vecE: vector of IDs of already played videos
   n: number of items that are being looked for
   ComSub: matrix of all videos in our subset of
           LDOS-CoMoDa research dataset (videos ID
           and coordinates)

   Output parameters:

   nIds: vector of IDs of the selected items

Create a subset of IDs from which diverse items
can searched (SimSub). The subset does not
include IDs of already played videos (vecE) and
currently playing videos (vecC).
while number of selected items is less than n do
  for all items in SimSub do
    for all items in vecC do
      Compute Euclidean distance between items in
      vecC and current item in SimSub.
    end for
    Compute the square root of the sum of the
    squares of the distances.
  end for
  Find the maximum distance and the ID of the
  items that this distance belongs to. Add this
  ID in vector vecC and remove it from SimSub.
end while
```

III. Methodology

Graphical analysis was based on a comparison among selection traces in MF space (see Sec. II.A) obtained by users of all three groups. Each interaction can be presented in two-dimensional MF space as a trace. A trace consists of links (lines) between points that present a selected item during interaction (Fig. 2). A line links two consecutive selected items. If the line is coloured red, the selection function recommends items in the next step that are similar to the selected item in the current step (see Sec. II). If the line is coloured blue, the selection function recommends diverse items (see Sec. II). The green circle represents the starting point (first selected video), while the red circle represents the last selected video (final selection). According to these selection traces, we determined the effect of the coverage of the MF space on the user's QoE.

To explain selection traces in MF space we need to present methodology for the evaluation of the effect of an SS on the QoE. Methodology how we measured QoE is briefly described in [5]. In this paper we highlight only the most important parts.

The evaluation was based on pre- and post-interaction questionnaires. The pre-interaction questionnaire comprised 16 statements having a seven-point Likert scale [12] (from completely disagree to completely agree) and one question for which only five different replies were possible. The aspects considered were user knowledge about video contents, user trust propensity, persistence of user choice, user affection towards new technologies, and possible user pattern preferences. Psychometric characteristics such as reliability (Cronbach's Alpha [13, 14]) and validity (average variance extracted [15]) were measured for most aspects.

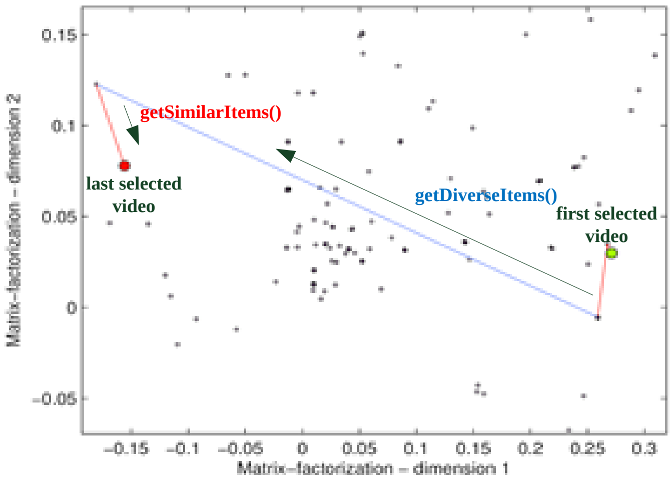


Fig. 2. Two-dimensional MF space of multimedia items, where each item denotes a point. Each line represents a link between two consecutive selections. If the line is coloured red, the system recommends similar items in the two selections; otherwise, it recommends diverse items (blue lines). The starting point is coloured green and the end point red

The post-interaction questionnaire consists of 25 statements and questions having a seven-point Likert scale [12], except in the case of demographics, for which data were collected in various ways. The questionnaire considered user satisfaction with the system, the system usability scale, past experiences with similar systems, the user selection time, user confidence in the accuracy of communication performance, user satisfaction with interpreted SSs, user satisfaction with recommended videos, user opinion about task complexity, and personal and demographic information. Psychometric characteristics were measured for most aspects.

To evaluate data from questionnaires, we used Fisher's exact test, the Mann-Whitney *U* test and an independent *t*-test for independent samples. An α-value of 0.05 was considered statistically significant.

Since a human operator was used for real-time action recognition and SS extraction, we estimated the possible effect of the human operator regarding his/her responsiveness and the consistency of his/her recognitions. Based on the results, we concluded that the use of a gesture-based user interface where a human operator performs gesture recognition does not have a negative effect on the interaction (his/her response time is fast enough). To check the consistency of human-operator recognitions in real time, we introduced two additional human operators for gesture and SS class recognition. Both results indicate that human-operator decisions made in real time do not critically affect the results of our experiment. Brief explanation of these results is given in [5] and [6].

IV. Results

Graphical analysis is based on the following assumption. The MF space of multimedia items (Sec. II) is built on more than 3600 ratings, which we can reasonably assume is the best possible layout of multimedia items for all users. We thus introduce the notion of the MF spatial area as an area of multimedia items with similar characteristics (short distances among items within the same area). Therefore, in our layout of items, there are several areas that combine items with similarities.

A user's past experiences with multimedia items are reflected in the way that the user prefers some items over others. Therefore, our second assumption is that, for each user, there are areas of multimedia items in our MF space that include preferred user items, called preferred areas. We graphically estimated the coverage of the preferred areas for the users in all three groups. All following analyses are based on the procedures described in Section III.

We used a sample of 42 users (N=42); there were 14 users for each of the control, random and test user groups. Since the evaluation in this paper is related to our previous research results, we firstly present the results of hypothesis testing that were published in [5] and [6].

A. Hypothesis Testing

To test hypothesis "*The use of the SS of hesitation in the RS improves the QoE when the user interacts with a VoD system*" we used statements from the post-interaction questionnaire that represented user satisfaction with the system. The first tested statement was "*The system is useful*." (St1) and the second statement was "*Overall, I am satisfied with the system*." (St2). The Mann-Whitney *U* test was employed to measure the *p*-value. Results are shown separately for the two pairs of groups (control and test groups and random and test groups) (Table I).

Before we measured QoE we detected and eliminated other possible causes for the difference between groups. We compared users according to (i) their basic demographics, (ii) their answers to the pre-interaction questionnaire, and (iii) the video content provided. We indicated two possible causes for the difference in QoE between user groups. Significant difference exists in age between user groups and in average rates of all videos that were recommended to the user. We concluded that difference in both cases does not give any advantage to the users in test group (SS considered) in measuring the effect of the SS.

Table I shows that in comparison between control and test group there is a significant difference in both cases (St1, St2), while in comparison between random and test groups there is not a significant difference. We can thus accept the null hypothesis only for comparison between control and test user groups.

B. Coverage for the Control Group of Users

Users in the control group are limited to one area of MF space that is not always suited for them. The MF space is poorly covered in terms of the items that the user sees. Therefore, the user does not always get an item that he/she wants.

Fig. 3 shows two typical traces made by users in the control group. The traces between selected items are short because every selection results in the recommendation of similar items (red line). Items thus cover only a small area of MF space. The users see only the items from one area of MF space, which may not correspond the preferred areas, possibly resulting in lower QoE (Table I).

TABLE I. THE RESULTS OF A USER SATISFACTION WITH THE SYSTEM (QOE MEASURE) AS COMPARED BETWEEN CONTROL AND TEST GROUPS AND BETWEEN RANDOM AND TEST GROUPS. THE NULL HYPOTHESIS WAS TESTED USING STATEMENTS ST1 AND ST2. A MANN-WHITNEY *U* TEST WAS USED. THE RESULTS ARE PRESENTED WITH MEAN VALUES FOR ALL THREE GROUPS (MEAN C – CONTROL GROUP, MEAN T – TEST GROUP, MEAN R – RANDOM GROUP) AND *P*-VALUE (P_{C-T} – CONTROL, TEST; P_{R-T} – RANDOM, TEST). ROWS WHERE A SIGNIFICANT DIFFERENCE WAS FOUND BETWEEN GROUPS ARE SHADED RED

	Variable	Mean C	Mean T	Mean R	p_{C-T}	p_{R-T}
User satisfaction	St1	4.86	5.64	5.29	0.022	0.051
	St2	4.64	5.64	5.00	0.045	0.069

C. Coverage for the Random Group of Users

Users in the random group are not limited to one area of MF space. The MF space is better covered in terms of the items seen by the user. Because the recommendation of similar (red line) or diverse (blue line) items is generated randomly, the items may cover areas that do not suit the user, and the user therefore does not always get an item that he/she wants. Fig. 4 shows two typical traces for users in the random group. The traces are interlaced because the selection function is selected randomly. Consequently, users see more areas in the MF space that could suit them but they cannot manage these recommendations. A mismatch between areas seen and the user's preferred areas can be reflected in poor QoE (Table I).

D. Coverage for the Test Group of Users

Users in the test group are not limited to one area of MF space. The MF space is better covered in terms of the items seen by the user. The items cover areas that are suited to the user because the system allows the user to manage the item recommendation through his/her SS. In this way, the user has a better chance to find an item that he/she wants to watch. Fig. 5 shows two typical traces made by users in the test group. Items better cover different areas of the MF space. The SS manages the recommendations and thus guides the user trace. If the user hesitates, a diverse-selection function is used. In contrast, if the user does not hesitate, a similar-selection function is used. The users see more preferred areas and select the most appropriate item from one of the suitable areas. This can be reflected by better QoE (Table I).

E. Analysis of Test-group Scenarios

As assumed, there are several preferred areas in MF space for each user. The function of diverse items (D) is used when the user hesitates, while the function of similar items (S) is used when the user does not. Function D allows the passing from one area of MF space to another, while S allows 'walking' only in one area of MF space. Below we present the most common scenarios for the test group of users.

1) A few S then D and then again a few S

The user finds one (preferred) area he/she is interested in and he/she wants to explore. The user does not hesitate and therefore receives similar items. Even if this is one of the preferred areas, the user does not find a suitable item after few steps. The user hesitates and gets four diverse items, which represent four diverse areas in MF space. The user selects another area that he/she finds is suitable and explores it until finding the item he/she wants to watch. The cycle can be repeated several times in one interaction. After each D, it is possible for one or more S to follow. In the case of only one S, the user probably thinks that the current area is of interest, but after getting more items from that area, the user changes his/her mind. In the case of more S, the user explores the selected area. The described scenario can be seen in Fig. 6.

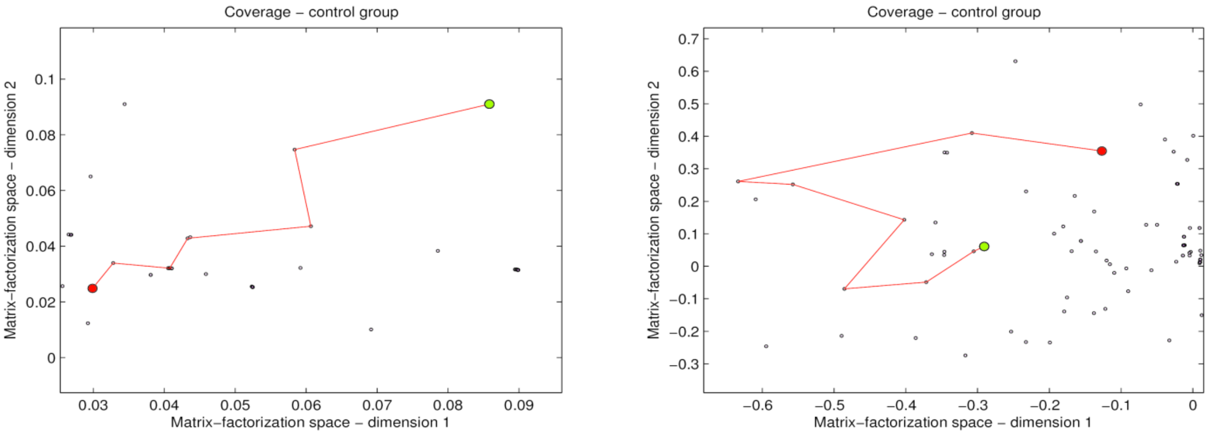


Fig. 3. Typical traces among selected items during the interaction of users in the control group with the VoD system. Lines between selected items are coloured red to indicate a similar selection. The user is limited only to one area in the MF space, which may not be his/her preferred area and may be reflected by poor QoE

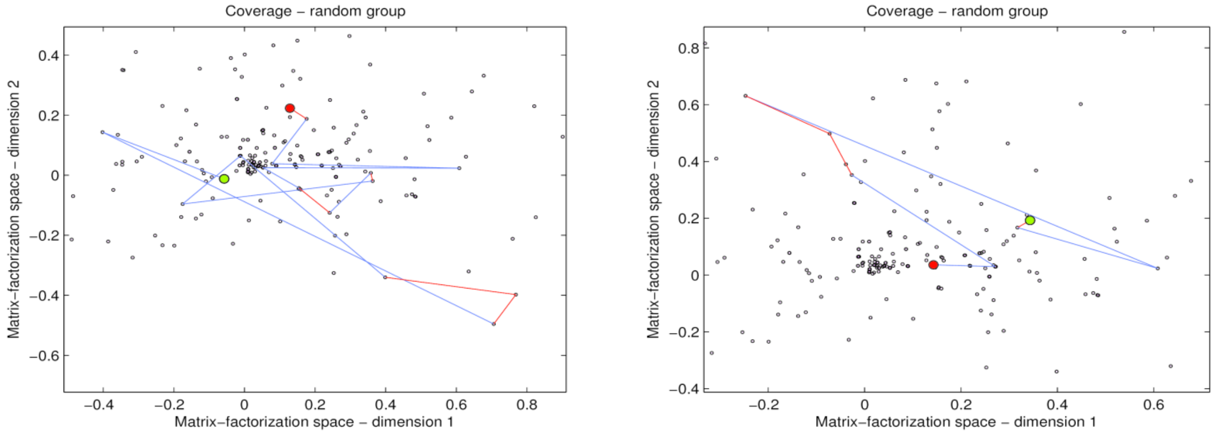


Fig. 4. Typical traces of selected items in the interaction of users in the random group with the VoD system. Lines between selected items are coloured red for a similar selection and blue for a diverse selection. The user sees more areas of the MF space but he/she cannot manage the recommendations and therefore cannot always get an item from a preferred area

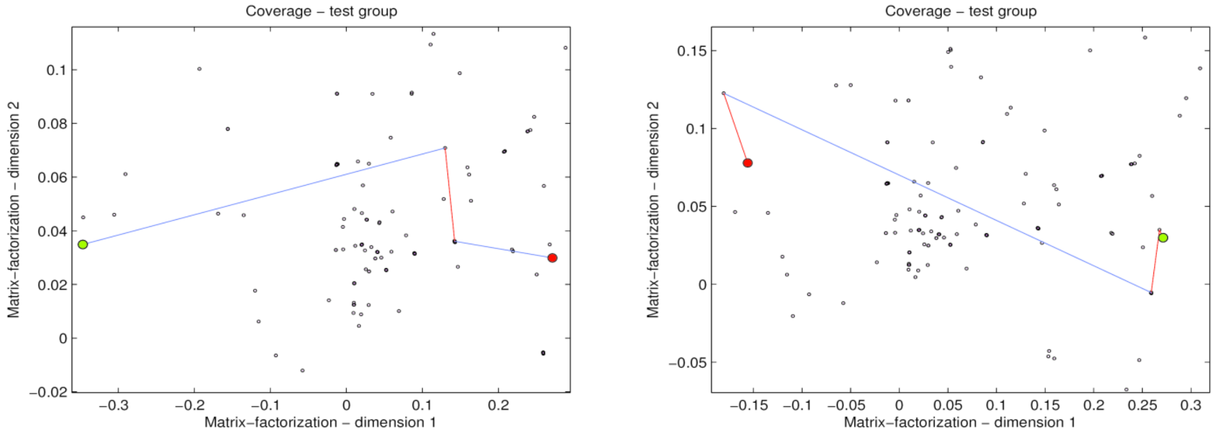


Fig. 5. Typical traces of selected items in the interaction of users in the test group with the VoD system. Lines between selected items are coloured red for a similar selection and blue for a diverse selection. The user sees more preferred areas in the MF space because he/she can manage the recommendations. Therefore, the user can always get an item from one of his/her preferred areas

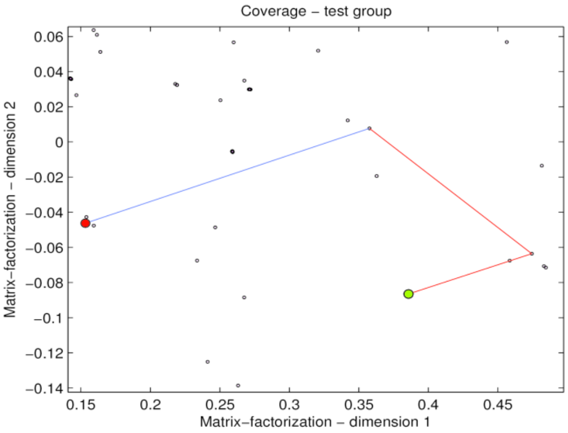


Fig. 6. The user finds an item from one of his/her preferred areas on the first screen and therefore explores this area. Since the user is not hesitating, the system provides similar items (red line). After a few steps, the user still does not find an appropriate item and thus hesitates, and the system provides diverse items (blue line). The user is interested in one of the four new items (in one of the preferred areas) and therefore does not hesitate to select it. After a few steps within this area, the user finds an appropriate item

2) One or several D at the beginning of the interaction

The user is not satisfied with the first screen (i.e., there is no single item from the preferred areas) and therefore hesitates. The user then gets items from four new areas. The scenario repeats until the user finds an appropriate (preferred) area. The user then explores this area until he/she finds an appropriate item. The described scenario can be seen in Fig. 7.

3) Only S in interaction

The user finds a (preferred) area that he is interested in on the first screen. The user does not hesitate and gets only similar items. After a few steps in this area, the user finds an appropriate item. This scenario is common for users who love to watch movies but have never used a similar system. The described scenario can be seen in Fig. 8.

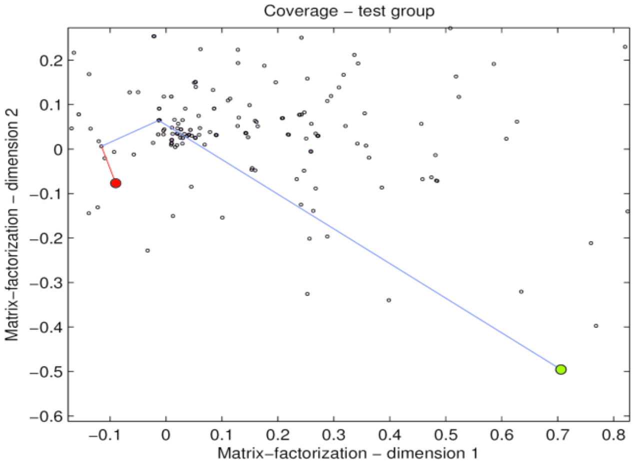


Fig. 7. The user does not find an item from a preferred area and therefore hesitates (blue line). After finding an appropriate area, the user explores (without hesitation (red line)) it and selects an appropriate item

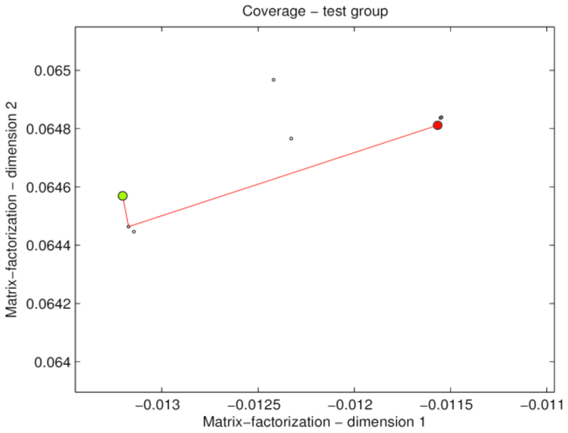


Fig. 8. The user finds an item he/she is interested in on the first screen. Since the user does not hesitate, the system provides similar items (red line). In a few steps, the user explores the selected area until he/she finds an appropriate item

V. Discussion

Comparison of the coverage of preferred areas in MF space of multimedia items among groups gave expected results. Since the users in the test group can manage the system recommendations (similar or diverse items) with their expressed SSs, their preferred areas of multimedia items in MF space are better covered, which is reflected in better QoE. The users in the random group cannot manage the system recommendations because the recommendations are generated randomly (random selection between similar and diverse selection functions). Despite this, the system can cover a user's preferred areas in MF space since the diverse function allows transition among areas in MF space. Users in this group have worse satisfaction with the VoD system than users in the test group but better satisfaction than users in the control group. Users in the control group can select an area in MF space only on the first screen and then can only explore within this area. There is thus a high probability that the user does not see any item from his/her preferred areas. These findings indicate that the use of the SS of hesitation in our VoD system provides better coverage of the user's preferred areas of multimedia items in MF space, resulting in better user satisfaction with the system.

VI. Conclusions

We presented, to the best of our knowledge, the first attempt to use the user's SS expressed during the interaction as part of feedback information. We modelled an experimental design and an associated experimental user scenario where users make gestures to select among videos on screen (i.e., VoD). Additional user-produced SS information was used to recommend new videos that were more suitable in the process of selection. Our previous work [5, 6] includes comparison between a group for which the SS is considered (test group) and a group for which SS is not considered (control and random group). Comparison based on pre- and post-interaction questionnaires. Our findings were (i) there was a significant difference between the test group (a user group for which the SS was considered) and the control group (a user group for which the SS was not considered) in user satisfaction with the system and (ii) there was a non-significant difference between the test group and random user group (another user group for which the SS was not considered) in user satisfaction with the system.

In this paper we present the results of graphical analysis of users' selection traces made in MF space of multimedia items to estimate the effect of the MF space coverage on the user's QoE. We concluded that the use of the SS in our VoD system provided better coverage of a user's preferred areas of multimedia items in MF space, which is reflected by better satisfaction with the system.

Our future work should focus (i) on increasing the size of the sample of the tested users and (ii) on the realization and testing of repeated-measures experimental design. Each user should test the scenarios of all three groups and decide only which of the scenarios offers him the best user experience.

ACKNOWLEDGMENT

Operation part financed by the European Union, European Social Fund.

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain", Image Vision Computing, vol. 27, no. 12, pp. 1743―1759, 2009.

[2] A. Vinciarelli, M. Pantic, C. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing", IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 69―87, 2012.

[3] A. Vinciarelli, H. Slamin, and M. Pantic, "Social Signal Processing: Understanding social interactions through nonverbal behavior analysis", Proceedings of the Computer Vision and Pattern Recognition Workshops, pp. 42―49, IEEE 2009.

[4] A. Pentland, "Social signal processing", IEEE Signal Processing Magazine, vol. 24, no. 4, pp. 108―111, 2007.

[5] T. Vodlan, M. Tkalčič, and A. Košir, "The impact of hesitation, a social signal, on a user's quality of experience in multimedia content retrieval", Multimedia Tools and Applications, DOI: 10.1007/s11042-014-1933-2, March 2014.

[6] T. Vodlan, and A. Košir, "Does the use of the social signal of hesitation in the recommender system improves the quality of experience when the user interacts with a video-on-demand system" Proceedings of the 8th International Conference on Interfaces and Human Computer interaction, July 15-17, Lisbon, Portugal, 2014.

[7] A. Košir, A. Odić, M. Kunaver, M. Tkalčič, and J.F.Tasič, "Database for contextual personalization" Elektrotehniški vestnik, vol. 78, no. 5, pp. 270―274, 2011.

[8] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model", Proceedings of the 14th ACM SIGKDD, pp. 426―434, ACM 2008.

[9] A. Odić, M. Tkalčič, J.F. Tasič, and A. Košir, "Predicting and detecting the relevant contextual information in a movie-recommender system", Interacting with Computers, vol. 25, no. 1, pp. 74―90, 2013.

[10] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering", Proceedings KDD Cup and Workshop, pp. 39―42, ACM Press 2007.

[11] G. Takács, "Major components of the gravity recommendation system", SIGKDD Explorations 9, pp. 80―84, 2007.

[12] J.G. Dawes, "Do data characteristics change according to the number of scale points used? An experiment using 5 point, 7 point and 10 point scales", International Journal of Market Research, vol. 50, no. 1, pp. 61―78, 2008.

[13] L.J. Cronbach, "Coefficient alpha and the internal structure of tests", Psychometrika, vol. 16, no. 3, pp. 297―334, 1951.

[14] J.C. Nunnally, "Psychometric theory", 1st ed.. New York: McGraw-Hill, 1967.

[15] C. Fornell, and D.F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error", Journal of Marketing Research, vol. 18, no. 1, pp. 39―50, 1981.

Study of Chaos in the Traffic of Computer Networks

Evgeny Nikulchev
Moscow Technological Institute
38A, Leninckiy pr., Moscow, Russia, 119334

Evgeniy Pluzhnik
Moscow Technological Institute
38A, Leninckiy pr., Moscow, Russia, 119334

***Abstract*—Development of telecommunications technology currently determines the growth of research with an aim to find new solutions and innovative approaches to the mathematical description of the processes. One of the directions in the description of traffic in computer networks is focused on studying the properties of chaotic traffic. We offer a complex method for the dynamic chaos determination. It is suggested to introduce additional indicators based on the absence of trivial conservation laws and weak symmetry breaking. The conclusion is made that dynamic chaos in the example of computer network traffic.**

Keywords—chaos; traffic of computer networks; nonlinear dynamics

I. Introduction

The article is focused on the computation of invariant characteristics of dynamic chaos based on the flow of corporate computer networks.

A significant amount of work on modeling of traffic in computer networks based on queuing theory. This, of course, involves the application of Poisson flow hypothesis, but this hypothesis is often not confirmed by the practice. The hypothesis of the Poisson streams can be used in networks with large redundancy across the width of the channel, in other cases there are other types of distribution and the process requires a fundamentally different approach to modeling.

Today's networks are characterized by the distribution of computing resources and a variety of end-users (from gadgets to appliances that have access to the Internet), with simulation aimed at communication channels control systems creation being a particularly urgent task.

The study specifies distribution which serves as a basis for analyzing data about downloading online channel from the monitoring work of university corporate network, measured in the course of the year. Statistics obtained by removing information from the router interfaces on the number of transmitted data and loading port, protocol snmp, using packet Paessler Router Traffic Grapher, which generates a table with data and graphics load (see Fig. 1).

Empirical histogram frequency channel load is shown in Fig. 2 On the basis of the use of the criteria of fit test and Kolmogorov - Smirnov observed probability distribution is not consistent with a Poisson distribution. Empirical histogram has a "heavy tail", indicating the presence of the peak moments of the network load, in which there is a strong increase in latency and packet loss.

Information on downloading channels was also obtained by monitoring the external communication channels of one of the companies and providers of on-site optimization. The resulting histogram also possesses ponderous tails, indicating the presence of the peak moments of the network load, in which there is a strong increase in delays and loss of information.

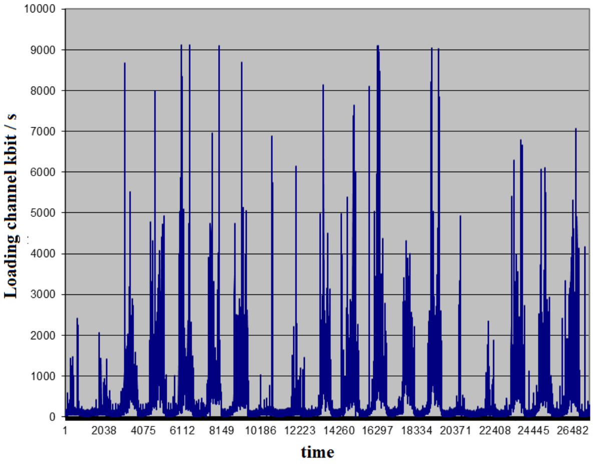


Fig. 1. Fragment of an annual progression loading of the channel network (20 days)

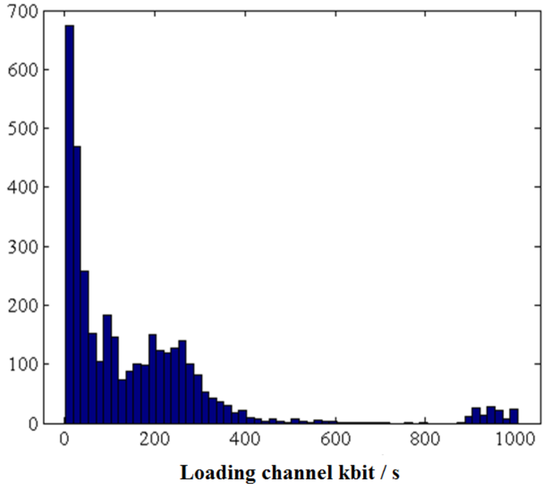


Fig. 2. Empirical histogram loading of the channel network (6 months)

Due to the fact that the distribution function has a heavy tail, and is not consistent with the Poisson distribution, queuing theory for the considered network cannot provide an adequate mathematical description.

As it was noted in [1] for the TCP / IP protocol distribution with ponderous tails makes a major contribution to the self-similar nature of the traffic and, consequently, the chaotic nature of the dynamics.

A number of works were focused on the study of chaotic traffic . In [1] the aim is to evaluate the values of the largest Lyapunov exponent on the basis of the traffic generated on the test bench; In [2, 3] Internet traffic is an example for the calculation of various characteristics; In [4] the dynamic properties of the chaos is used to solve telecommunication problems of data exchange, but the study of chaotic properties remained outside publications.

II. Calculation of the Characteristics of Dynamic Chaos

It is assumed that the time series generated by the discrete

$$\mathbf{x}_{k+1} = f(\,\mathbf{x}_k\ \mathbf{x}_0\,), \qquad (1)$$

or a continuous system

$$\frac{d\mathbf{x}(t)}{dt} = F(\mathbf{x}(t), \mathbf{x}(0)). \qquad (2)$$

Here, $\mathbf{x} = (x_1(t), ..., x_n(t))$; n — the dimension of the phase space; *t* — time; *k* — discrete time (number); *F*, *f* — vector function. Phase trajectory of a continuous system is an n-dimensional curve, which is a solution of the system of coordinates of the state space for given initial conditions $\mathbf{x}_0$. For discrete systems able to connect lines in accordance with the sequence of samples k= 1, 2, ...

An important concept of dynamical systems is the attractor. For systems in equilibrium, the attractor is a point (with the time change state x does not change), for oscillatory systems - closed paths (cycles). For chaotic systems, there is an attractor, which is called the odd, in this case the trajectories are drawn, but not to the point, a curve, a torus, and in a subset of the phase space. Attractor is an invariant feature of the system, ie. is preserved under a conversion action

Unambiguous characteristics of chaotic signal are a spectrum of Lyapunov exponents. Positive maximum of Lyapunov exponent is a measure of chaotic dynamics, zero maximum Lyapunov exponent denotes a limit cycle or quasi-periodic orbit and negative maximum Lyapunov exponent is a fixed point [2]. System of dimension n has n Lyapunov exponents: : λ_1, λ_2,. . . , λ_n, ranked in descending order. According to the definition introduced by Lyapunov:

$$\lambda_i(x_0) = \lim_{t\to\infty} \frac{1}{t} \ln \frac{|\,\delta_1(t)\,|}{|\,\delta_i(0)\,|}.$$

here $\{\delta_i(t)\}$ — the fundamental solution of the system, linearized in the neighborhood of $\mathbf{x}_0$.

Dynamical systems, for which the n-dimensional phase volume decreases are called dissipative. If the phase space is conserved, such systems are called conservative. In conservative systems there is always at least one conservation law. The presence of the law of conservation often implies the existence of the corresponding zero Lyapunov exponent. For dissipative dynamical systems sum of Lyapunov exponents is always negative. In dissipative systems, Lyapunov exponents are invariant with respect to all initial conditions.

In terms of Lyapunov, it is possible to provide much information on the observed mode of the dimension of the attractor, if any, and on the entropy of a dynamical system. Dynamic chaos meets the instability of each individual trajectory, ie presence of at least one positive Lyapunov exponent. The attraction of the attractor requires that the phase volumes of large dimensions shrank, then reflected in the Lyapunov spectrum. Knowledge allows us to estimate Lyapunov exponents and the fractal dimension of the attractor [1].

Nevertheless, the number of independent frequencies cannot always find out as zero indicators may be associated with the presence of conserved quantities. The presence of dissipative systems of conservation laws, in general, is not typical, but there are relevant examples.

There is a considerable amount of numerical methods for calculating Lyapunov exponents from time series [2]. It is important that the condition that the number generated by the system under study (1) or (2), a senior figure could be calculated. However, it is impossible to estimate the entire spectrum. For distributed systems, even knowing the system of equations, the evaluation of the Lyapunov exponent is a significant computational complexity.

For the test series, shown in Fig. 1, 2, we calculate the largest Lyapunov exponent. For the calculations we used a system TISEAN. The results of calculations by different methods showed a positive value of the highest exponent.

However, the positivity of the largest Lyapunov exponent cannot be a necessary condition for the existence of chaos. For example, even in a system of Lorentz with positive leading indicator is known in a number of conditions, have a limit cycle.

An additional criterion to use the property of absence of trivial conservation laws was suggested that is — symmetries broadcast, tension and compression. A lowest symmetry violation is used to identify the chaos [5, 6]. Note that the compression phase volume does not mean conversion ratio.

To check the transformation fragment trajectories genetic algorithm and program for MATLAB were developed, their description is given in [3]. At the same time there is a check of the following assumption [6], which proves that the system allows conversion in low-symmetry breaking, i.e. there is some small value, slightly deviating from the symmetric display. Visually, it is geometrically evident at almost similar hinges on the attractor. It is obvious that in such a test, with different initial conditions for systems with regular dynamics is was discovered that they identical symmetry, for more complex but not chaotic — translation (shift of the phase portrait) for systems that tend to stable equilibrium position — compression, etc., and for the chaos — almost repeated portions of phase trajectories.

Reconstructed, according to traffic load, the attractor is shown in Fig. 3.

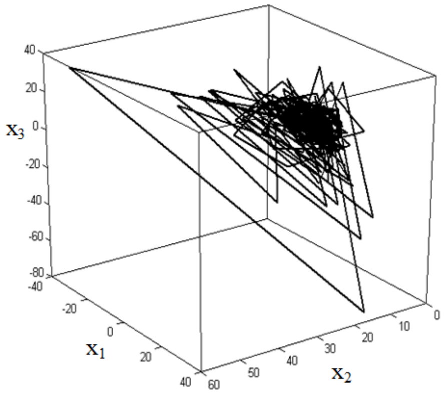


Fig. 3. Attractor, built on the basis of network traffic

In general, studies of chaotic signal can be formulated as follows.

1) Construction of the histogram. If there is heavy-tailed, it is necessary to check the chaos.

2) Calculation of the necessary conditions - Lyapunov exponent, Hurst exponent.

3) Construction of the attractor and the identification of symmetry breaking.

If all three tests are accomplished, there is a chaos in the system, and this property should be considered when dealing with such networks.

Confirmation of chaos can be the basis for building dynamic models. For example, in the form of an ensemble of pendulums [1], affinity controlled systems [8] or in the form of rows.

Identification of the parameters of the system using the method of [3, 8], gives the following result:

$$d\mathbf{x}/dt = \mathbf{A}\mathbf{x}(t) + \mathbf{\Psi}_0(t),$$

$$\mathrm{A} = \begin{bmatrix} 0.9413 & -0.1805 & 0.1164 & -0.0295 \\ -0.0545 & 0.8226 & 0.1622 & 0.1056 \\ 0.0014 & -0.0105 & -0.4455 & 0.8474 \\ -0.0062 & 0.0341 & -0.8860 & -0.5404 \end{bmatrix};$$

$$\Psi_0 = \begin{bmatrix} 0.0399 \\ 0.0463 \\ -0.4848 \\ -0.1851 \end{bmatrix} (\exp(t^{0.0001})\sin(t^{0.4})).$$

III. Conclusion

The paper deals with the chaotic phenomena in computer data networks. Based on the chaotic properties can be constructed mathematical models of the dynamic behavior of traffic. Models can be used to provide guaranteed quality of service (QoS), the analysis of bottlenecks in the structure of the corporate network, data sharing in cloud environments [9, 10].

At the same time, and the chaos of the indicators themselves, the structure of the attractor can have the value. Changing the values of the highest Lyapunov exponent, topology change attractor is an indicator of changes in network activity. For example, computer attacks [6, 10], the failure (denial of service) enterprise data exchange, or a reason for the change of policy administration - an extension of communication channels or by completing a list of banned network resources. For example, this is the case of recently observed popularity of social networking, video sharing resources.

References

[1] A. V. Karpukhin, I. N. Kudryavtsev, A. V. Borisov, D. I. Gritsiv and H. Cho "Computer Simulation of Chaotic Phenomena in High-Speed Communication Networks," Journal of Korean Institute of Information Technology, 2013, vol. 11, no. 3, pp. 113–122.

[2] Z. Liu "Chaotic Time Series Analysis," Mathematical Problems in Engineering, 2010, vol. 2010.. (doi:10.1155/2010/720190)

[3] E. V. Nikulchev and O. V. Kozlov "Identification of Structural Model for Chaotic Systems," Journal of Modern Physics, 2013, vol. 4, no. 10, pp. 1381–1392. (doi: 10.4236/jmp.2013.410166)

[4] W. Xiong, H. Hu, N. Xiong, L. T. Yang, W. C. Peng, X. Wang, Y. Qu "Anomaly secure detection methods by analyzing dynamic characteristics of the network traffic in cloud communications," Information Sciences, 2014, vol. 258, pp. 403–415. (doi: 10.1016/j.ins.2013.04.009).

[5] R. O. Grigoriev "Identification and Control of Symmetric Systems," Physica D, 2000, vol. 140, no.3–4, pp. 171–192. (doi: 10.1016/S0167-2789(00)00014-2).

[6] P. Chossat and M. Golubitsky "Symmetry-increasing bifurcation of chaotic attractors," Physica D: Nonlinear Phenomena, 1988, vol. 32, no. 3, pp. 423–436.

[7] E. V. Nikulchev "Geometric method of reconstructing systems from experimental data," Technical Physics Letters, 2007, vol. 33, no. 3, pp. 267–269. (doi: 10.1134/S1063785007030248)

[8] E. V. Nikulchev "Geometric Method of Reconstructing Evolution Equations from Experimental Data," in Evolution Equations, A. L. Claes, Eds. New York : Nova Science Publishers, 2011, pp. 373–379.

[9] E. V. Pluzhnik and E. V. Nikulchev "Use of dynamical systems modeling to hybrid cloud database," Int'l J. of Communications, Network and System Sciences, 2013, vol. 6, no. 12, pp. 505–512. (doi: 10.4236/ijcns.2013.612054)

[10] E. Pluzhnik, E. Nikulchev and S. Payain "Laboratory Test Bench for Research Network and Cloud Computing" Int'l J. of Communications, Network and System Sciences, 2014, vol. 7, no.7, pp. 243–247. (doi: 10.4236/ijcns.2014.77026)

Case Study: the Use of Agile on Mortgage Application: Evidence from Thailand

Kreecha Puphaiboon
Faculty of Computer and Information Technology
Kasem Bundit University
Bangkok, Thailand

***Abstract*—This paper presents a case study of a mortgage loan origination project using SCRUM Agile model and Business Process Management and Business Rule Management System (BPMS and BRMS). From the Waterfall model (Stage 1), a web-based self-developed had been developed using opensource frameworks: Spring and Sarasvati. But, several problems were detected and the project failed due to insufficient project management, rapid requirement changes and developer coding skills. The project was continued (Stage 2) selecting a BPMS and BRMS tool. Later, Stage 3 SCRUM was executed with proper project management and the new tool, which suited better for rapid business needs, and minimum coding. An efficient team communication and the frequent delivery of code releases increasingly contributed to the sponsor and user's satisfaction. However, due to political influenced timeline, inexperienced project management and requirement changes, the budget exceeds and SCRUM is not favored. Nonetheless, Open-end questionnaire and interview results with core team members both business users and developers as well as software usability measurement inventory (SUMI) conducted with 14 users, it shows that SCRUM and the new tool rescue the project. Empirically, this paper demonstrates a method to evaluate the use of Agile augmented with usability measurement to Agile development community.**

Keywords—SCRUM; BPM; BRE; Mortgage Loan; LOS; Usability

I. Introduction

Mortgage loan origination process is complex, although common process flows are: Application, Processing, Underwriting, Closing and Post Closing [1]. However, a system that can handle a large number of Loan numbers is hard to find. In addition for IT, the pressure to revise flows, policy changes and credit risk calculations in a few days and automatically make correct lending decisions has been a great challenge in retail banking [2]. There's no room for customer and financial information and loan process errors as banks need to have certain confidence in every lending decision.

Agile's principles encourage the formation of collaborative and self-organization teams [3]. The Agile Manifesto is as follows: 1) Individuals and interactions over processes and tools. 2) Working software over comprehensive documentation. 3) Customer collaboration over contract negotiation. 4) Responding to change over following a plan. However, a continual debate surrounds the effectiveness of agile software development practices. Some organizations adopt agile practices to become more competitive, improve software development processes, and reduce costs. Other organizations are skeptical about whether agile development is beneficial. Additionally, large organizations face an additional challenge in integrating agile practices with existing standards and business processes [4].

Whilst it is generally accepted that SCRUM development improves the cost reduction and it helps to accelerate the software product to the market. Importantly, it improves customer satisfaction [5] [6]. However, no field studies research has been reported when Agile and SCRUM is first being used in a large organization in Thailand. Research [5] mentioned factors to run Scrum in aligned with PMP BOK principles (see Table I). Hence, the researcher wants to study and report the usage of Agile/SCRUM to satisfy business needs and assess the impacts over the IT development an example for software engineering community.

TABLE I. step 1. 4, often the projects have been assigned with the timeline [6]. Political forces at work within a project or company can often drive estimation inaccuracy [8]. This is usually in the form of managerial pressure to stay within or meet the estimate timeline [8]. The estimation process can be impacted negatively by these pressures resulting in project timeline or cost constraints [8] [9].

The rate of change in business and bank is accelerating [1] [2] . A number of techniques for addressing that change have emerged independently to provide for automated solutions in this environment. Business Process Management (BPM) and Business Rule Engine (BRE) that are large as well as distributed are becoming more prevalent [10] [11]. Both technologies tend to offer the promise of easy to change. As change is common in large projects; the case where the entirety of a project's complexity is understood in the early stages is quite rare. Large, distributed projects that involve user requirements present a unique challenge that neither agile methods nor waterfall approaches alone can effectively address. Hence, combining an effective software development tool with agile process may be very beneficial.

Koch [12] has proposed three criteria for evaluating the effectiveness of the agile method adopted: 1) project performance with schedule performance and budget performance; 2) management acceptance; 3) customer relationship and 4) team satisfaction. However, usability was not included. Thus, it is important to evaluate all these five criteria for agile adoption for which they were deserved.

I am thankful to Kasem Bundit who has sponsored this research

TABLE I. THE KEY PROCESSES OF RUNNING SCRUM

The key processes of running scrum		
1. Determinate phase	***2. Planning phase***	***3. Start-up phase***
1.1. Develop the real requirements of customers;	2.1. Define all the work of the project;	3.1. Recruit project manager;
1.2. Write a one page project description;	2.2. Establish the schedule of initial project;	3.2. Build the scope change management process;
1.3. Recode the requirement of customers;	2.3. Assess the time required to complete the project;	3.3. Recruit the project team members;
1.4. Gain the senior managers' permission to run the project;	2.4. Analyze and adjust the project schedule;	3.4. Manage the team communication;
1.5. Discuss how to meet the requirements with the customers.	2.5. Assess the resource required to complete the project;	3.5. Write the descriptive document of project;
	2.6. Write the risk management plan;	3.6. Determine the schedule;
	2.7. Assess the whole cost of the project;	3.7. Build the team operating rules;
	2.8. Record the project plan;	3.8. Write the work package.
	2.9. Sort the work in chronological order;	
	2.10. Get the senior management's permission to start the project.	
4. Supervision and control phase	***5. Decided to start the iteration phase***	***6. Closeout phase***
4.1. Build the running and reporting system;	5.1. Decision-making process for customer management;	6.1. Get the confirmation of the customer;
4.2. Report the schedule;	5.2. Customers must be fully involved in this process;	6.2. Prepare for the deliverables and installations.
4.3. Supervise the running;	5.3. The atmosphere must be complete open and honest;	6.3. Write the closeout report;
4.4. Deal with the request of scope change;	5.4. Determination must base on the expected commercial value;	6.4. Start the audit of the running.
4.5. Supervise the risks;	5.5. Solution must be formed according to the project's goal.	
4.6. Identify and solve the problems.		

Agile and usability aim to build quality software. As noted in research [13], agile and usability the two methods have much to offer when they share iterations because the iterations used in agile facilitate usability testing and allow developers to incorporate results of these tests in subsequent iterations. However, research [13] commented that improving the usability of a product does not come without costs. In order to integrate agile and usability and at the same time minimize these costs and risks, we need the use of usability artifacts and practices in a condensed form. SUMI for Software Usability Measurement Inventory is commonly used for usability evaluations [14]. SUMI consists of over 50 questions; it is method of measuring software quality from the end user's point of view. There is a need of usability measurement integrated with agile methodology to determine whether the software supported mortgage loan needs or any domains.

This article presents the case of mortgage loan origination project called LOS. The purpose of LOS was to replace an existing mortgage application as there were problems such as: legacy application (over 15 years old), lack of Power Builder developers to support, difficulty in changing business flows, incapable of scalability and performance issue. There were over 8,000 users using the mortgage system, roughly around 500 concurrent users across the nation. There are needs for new application with features: easy-to-change, web-based and scalable, user to make routine changes. Also, the emphasis is on IT to have a fast delivery of the software application to compete with market trends and attract customers, the faster the delivery of software are the chances that the bank will gain profits. Note that the writer is the technical manager of this project. The bank is ranked in the top five banks in Thailand. The revenue of mortgage loan was over 1 billion us dollars in year 2012. So, it is a highly critical system for the bank.

The paper is structured as follows: Section 2 presents the development stages of delivering a prototype, including detected problems. Section 3 explains ways in which BPM/BRE tools were selected. Section 4 describes the full development SCRUM model and some properties of the methodology are summarized of the system and its iterations. Section 5 shows team comments on Agile/SCRUM base on questionnaire and feedbacks via Sticky Notes and user experience assessment to the software via SUMI usability questionnaire. Lesson learned and conclusions are presented in Section 6.

II. STAGE ONE OF MORTGAGE APPLICATION – PROTOTYPE SELF-DEVELOPMENT

Using Waterfall model, the development of web-based mortgage application the project was initiated (2011–2012) timeline and main tasks were planned by senior management with four months for each step of requirements, coding and UAT testing. Two main representatives' of business users were given. Requirement gathering in June – September 2011, there were 18 main flows from start to finish entire loan process e.g. data entry, manager, credit approval officer, credit approval manager, legal contract and legal managers. The direction was to used open-source framework and in-house development the J2EE, Spring Framework (2009) and Sarasvati proposed by IT developers. Spring is used to handle simultaneous runs and as an interface to database DB2 (see in Appendix for Fig. 4). The development effort was plan roughly for 14,400 man hours for the entire project for one project manager, two business users, 10 Java developers (average experienced 3.5 years) and 3 testers. The budget was 180,000 USD.

Application development was carried out in October 2011 - January 2012, but different problems threatened the project continuity. Considering the SIT was unable to finish within the first two month (1 month delayed) due to numerous bugs. The team had difficulties to follow defined plans and the senior executives were unsatisfied with the progress of the software development after 6 months. The following main deficiencies were detected:

- Deficient project management and communication; project manager, domain experts and end users were present only six-hours per week. They also sit in different buildings with IT developers. Telephone and e-mail were used. These means did not result efficient when problem arisen and need immediate action from BAs.
- Requirements were broad without enough details and cannot integrate entire flow. For example, details of different collateral types between two units are not in sync (processing and underwriting). Therefore, most of the coding tasks required impact analysis, modifications and re-implementations, causing continuous delays to the development process.
- The implications are poor change control, developers' software design skill and software framework to adopt dynamic changes.
- The developer stated that complicated flow, business rule calculations and changes were the root cause of failures.

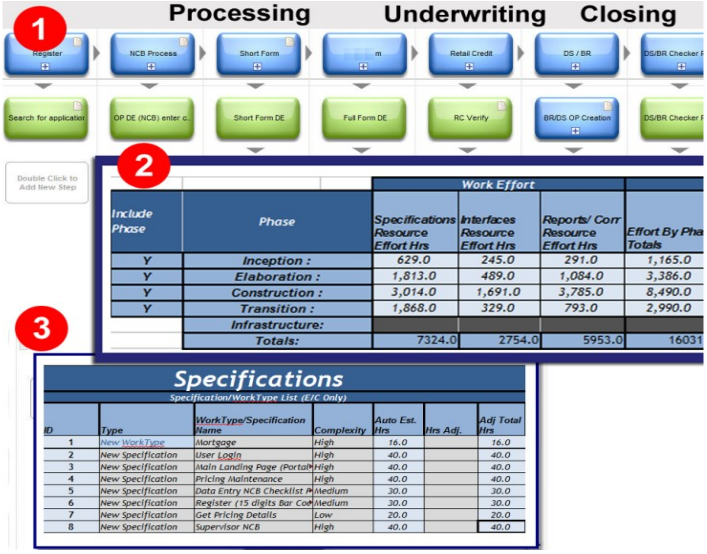
The objectives of self-develop web application project were clearly not met. Significantly the budget was exceeded by 60,606 USD. After a meeting between the team and executives (CFO, COO and CIO), all decided to discontinue the self-development. CIO advised the project to acquire BPM and BRE tools which support developer to code as less as possible and users can manipulate the flow and rules within the system. This is to downgrade possible failure risks. Importantly, the team requested access to the expert user on a daily and full-time basis.

III. Stage Two of Mortgage Application – Selecting and Assessing the Tool

On February 2012 the BPM and BRE Vendor Selection was executed. The top product listed in Gartner and Forrester ranking reports were invited. Importantly and practically, two weeks of POC project to build almost half of an entire mortgage process arisen from Stage 1 including SIT and UAT were conducted. Moreover, stress test with 500 concurrent users was conducted and passed 3 seconds respond time criteria. Developers and users from the bank also involved in the development as well as evaluation of the software. This is to prove that the selected BPM/BRE framework provided software flexibility for fast development without much coding and ease of change when users want to change various calculation schemes. For end users, they appreciated the fact that they can input decision rules and the software provides friendly and intuitive screens for users. Another additional benefit for developers is that the software supports Agile development: 1) users can draw flow and design screen with developers in real time without coding this helps to improve business gaps with users; 2) flows and rules can be drawn/changed without coding; and 3) requirements and documents are saved in the system so developers can identify changes. Another benefit for the bank is that financially, 5 Year Total Cost of Ownership (TCO) is less than other products. The product also offers Cloud Amazon EC2 service, thus it leverages maintenance agility and investment cost. 5 year TCO of the project is around 10 million US dollars. But for mortgage application is funded with 1.5 million USD where outsource developers budget is around 330,000 USD. Time-and-Material contracts and Labor Hour (LH) contracts are used. Three weeks of training was also provided to local staff including business analysts.

IV. Stage Three of Mortgage Application – Application of SCRUM

On June 2012, the Stage 2 of the software development finally started. As it was learnt that requirement documentation was not clear, hence the re-documentation/requirement were captured and put directly into the system. Two weeks period with multiple sessions captured the details of the use cases and flow which give the project traceability and determine application requirements. During the requirement gathering, a close seating and direct communication environment within a big room with a projector, design sketches and white-board, notes and mocked up screen were conducted. These meetings focus on efficiency, getting 2 subject matter experts (SME) from 2 departments into the room to focus on the implementation details of mocking up UI, validation of inputs and outputs and aware of each other impacts. They agreed up front on how the application processes will work, avoiding costly rework later. The outcomes determined the number of iterations, sprints and effort required for the project. Due to business confidentiality, Fig. 1 shows some of business process flow of Application Registration, Processing and Underwriting (without Closing and Post Closing). More details will be discussed in Section 4.1 – 4.5. The sizing effort by developers was produced with 11,480 man hours for coding and unit testing (see Fig. 1 circled number 2) with total of 213 use cases (see some specification below in Fig. 1 circled number 3).



		Work Effort			
Include Phase	Phase	Specifications Resource Effort Hrs	Interfaces Resource Effort Hrs	Reports/ Corr Resource Effort Hrs	Effort By Pha Totals
Y	Inception :	629.0	245.0	291.0	1,165.0
Y	Elaboration :	1,813.0	489.0	1,084.0	3,386.0
Y	Construction :	3,014.0	1,691.0	3,785.0	8,490.0
Y	Transition :	1,868.0	329.0	793.0	2,990.0
	Infrastructure:				
	Totals:	7324.0	2754.0	5953.0	16031

Specifications

Specification/WorkType List (E/C Only)

ID	Type	WorkType/Specification Name	Complexity	Auto Est. Hrs	Hrs Adj.	Adj Total Hrs
1	New WorkType	Mortgage	High	16.0		16.0
2	New Specification	User Login	High	40.0		40.0
3	New Specification	Main Landing Page (Portal	High	40.0		40.0
4	New Specification	Pricing Maintenance	High	40.0		40.0
5	New Specification	Data Entry NCB Checklist P	Medium	30.0		30.0
6	New Specification	Register (15 digits Bar Cod	Medium	30.0		30.0
7	New Specification	Get Pricing Details	Low	20.0		20.0
8	New Specification	Supervisor NCB	High	40.0		40.0

Fig. 1. Flows of Mortgage Application

However, TABLE II shows the output sizing sheet and effort for the project which was influenced by the senior executives to 7,524 man-hours (reduced 34%).

TABLE II. ITERATIONS PLAN AND SIZING EFFORTS

Steps	**Political Forced Effort and Initial Effort in Man-Hour**		
	Name	***Political***	***Plan***
1	Application Registration	1,332	440
2	Application Processing	720	380
3	Underwriting and Closing	480	3,296
4	User Management and Change of Condition	2,256	502
5	Risk Analysis and Interface	2,736	6,862
Total		**7,524**	**11,480**

In this project, the system development team was integrated by four groups with the following roles:

- Developers: agile defines specific categories team lead designer and programmers. In this case, they were cross-functional and allocated dynamically depending on particular needs of the running iterations. Four developer teams were agreed between three different vendors and the bank. 23 developers were engaged and 16 were outsources from India, Singapore and Hong Kong averaged 4.0 years experienced with the product. Higher number of outsource the reason being that because bank staff had little experience with the product.

- Product owner is IT business analyst lead manager: she works with SME and users.

- Scrum master is technical manager who does the code review, release management, network, and security. He serves as a resource to help the teams make appropriate system and component level design decisions during implementation. He defines and split use cases and features for the program backlog, and allocating respective items to the individual team for implementation.

- Team involves 3 SIT testers, an architecture leader from the vendor who establishes software architecture design support of upcoming user and business needs and helped developers when required. Subject matter experts: three domain experts each for processing, underwriting and closing who can evacuate doubts or give a rapid opinion as required by developers. Different users carried out this role during the development depending on the issue under review. A dedicated team for BRE was also provided to deal with all underwritings and A-Score models.

A daily "scrum" or standup meeting was held with all the stakeholders. Every day, developers answered three questions: 1) what have you done since yesterday; 2) what are you planning to do by tomorrow; and 3) Do you have any problems preventing you from accomplishing your goal. To satisfy quality assurance of development, unit test with capture screens was applied by developers and reviewed by business lead. Frequent delivery and test with every two weeks, a release was delivered to SIT environment with bug fixes and new features in product backlog. For fast testing, QTP was used as an automated testing tool. Once a month, the product was released to UAT and a meeting with steering committees was conducted to inform the status of bug fixes and project status. This is to ensure that the team and all stakeholders have reviewed the product and meet the expectation.

A. Iteration 1 (Application Registration)

Following the plan, a short first iteration (15 days) was designed (02/07/12 - 23/07/12) but it was taken 34 days to complete - 19 days later than planned. This iteration involves the first three columns in **Error! Reference source not found.**1 where users register a mortgage application with the barcode and confirm application creation. The data entry person will enter customer information from a hard copy document and use the citizen identification to interface with National Credit Bureau and obtain the credit score result which will be used later in underwriting calculation. User can check and search if the borrower or co-borrowers information exist in the core bank systems, if so retrieve all the information. Once completed, they can send to the supervisor whose role is to review and revise wrong/missing data given by their staff. Later, if required they can submit the work to another unit whose roles are to comment and record missing data or document for further analysis of sale sufficiency. They also need to follow up with sales or customers about the missing information.

The first integration of the GUI presentation layer with the bank systems was achieved as a working software delivery. However, there were two factors which affected this iteration timeline. Firstly, Bank of Thailand regulation states that the bank cannot keep customer's sensitive information e.g. name, surname and id outside of Thailand. As a consequence, a special HTML/Ajax control was developed which maps all sensitive information kept in Cloud and the bank. The control is developed and can be reusable on all screens. The bank utilized the 'Mapper' server using Java and Spring Framework developed in Stage 1 to keep the real customer information (see Fig. 2) and interface with the bank internal systems via MQ. Secondly, during the reflection workshop carried out at the end of the present iteration, different code conventions were specified and refactored to facilitate maintenance and readability of the code. As a result the iteration was finished 17/08/2012, 19 working days delay.

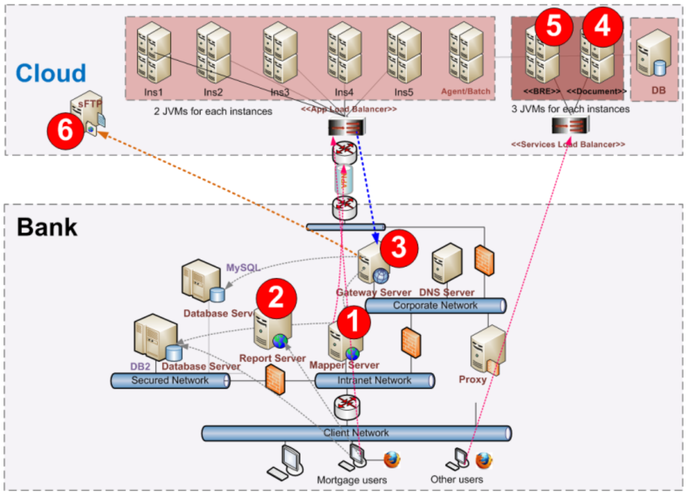


Fig. 2. Architecture of the System

B. Iterations 2 (Application Processing)

The second iteration started on 25/07/2012 to 03/09/2012 with an incremental on top of the previous iteration, developers enhanced features: 1) employee loan and employee search capability (edit, delete, and add) which can be reused in all screens. 2) Borrower search with edit, delete, add and online information retrieval of related loan and customer details resided in old mortgage application as data migration is not implemented and 3) validation for previous screen input fields were done.

To enhance business value, the iteration developed routing capability which is to deliver task to designed users and unit. In addition, all managers can track and store for key performance indicators (KPIs). So the manager can analyze process performance as well as create service level agreements (SLAs). However, a challenge of this iteration is that processing a mortgage loan, there were over 180 input fields in one page such as pricing plan detail, fee detail, insurance service, payback plan, guarantor, secure collateral types i.e. land and building, deed, condominium, debenture, bond and etc. These field values were highly inter-related and significant for the loan outcome. Therefore, developers required business knowledge, times for develop and unit test. On this iteration, prior to the delay 5 additional users came into help after office hour to test and identify missing requirements on the screen and validation rules and prevent further delay. Nevertheless, due to business complexity of all fields, the iteration faced 18 days delay and completed in 03/09/2012.

C. Iterations 3 (Underwriting and Closing)

This iteration was carried out from 23/08/12 to 13/11/12 (58 days). Two main processes were developed where underwriter officer approved the loan will be routed to loan closing unit. In this project underwriter officers' main tasks include: 1) calling the borrower about his/her income, wealth, credit history; 2) verifying borrower information with third parties such as social security department, other banks and employers via phone and online government website; and 3) approving or overriding the underwriting result (risk analysis, see Section 4.4). Loan closing process which is triggered automatically when underwriter approved. The loan application will be delivered to appraisal unit, once having an appraisal approved. The system initiated the process of finalizing documentations and printing between the bank and customer i.e. insurance, contract and cheque.

There were challenges in this iteration resulting in 48 days delay. With underwriting development (20 days delay), as data model was from the previous processing unit. However, the underwriters can add comments and adjust all fields resulting in 225 input fields appeared on screen (see Fig. 5). In addition, auto population of historical data (e.g. statements, debt, and insurance) via 7 interface bank systems was requested to reduce their time to key-in in one single page. Nonetheless, midway development underwriter users were informed on performance issues with HTML streaming and interface time (time > 3 seconds), so data grouping was proposed. UI was redesigned collaboratively with users where all input fields were groups into 3 tabs: customer (name and address, guarantor), finance (income, debt) and loan (mortgage term, rate, installment amount) information. Once, the user wants to view or edit then they can click each tab individually to improve loading time of less than 3 seconds. For interfacing issue, a manual click was proposed and accepted with seven interface buttons provided for users to enquire when needed.

With the closing state, there were resource and technical problems with the development of legal contracts. Not knowing before, the BPM tool the bank bought is not sufficient in generating documents. There were over 140 legal contracts in mortgage loan. So, legal contract users were trained to use iReport and designed the contracts which fulfill legal compliances such as font size, paragraph, margin and etc. Additional developers also were employed. Technically, loan data was kept in Cloud and real time data was required. So a dedicated web-service server was used to transfer loan data from Cloud to the on-premise report server (see **Error! Reference source not found.** circled numbers 2 and 4). It was also found out that legal contracts consumed JVM memory, as each loan requires 30 legal contracts at least. As a result, a dedicated server was provided with 2 JVMs inside. The iteration was once finished but during the iteration review, business executives requested a flow change (delete of appraisal manager), they were not cognizant of the impact these changes would have on the project budget or timeline, leading to significant tension across the project teams.

D. Iterations 4 (User Management and Change of Condition)

The fourth iteration started on 05/09/2012 and finished on 19/12/2012 19/12/12 taken 76 days. By then, the team understood many requirements of housing loan and in a continuous improvement, favoring a high team motivation. Another module was developed to handle user profile management, which defines allowed access rights, authority levels, skills, approval limits, department, views and outcomes by profile of different user groups. This is used to control startup features for the user's screen and session, the types of authorized approval limit and screen that s/he can operate.

The late part of business process is the 'change of condition'. The purpose of change of condition is to enable any information adjustment of the latest approved loan such as details of customer, finance and loan. The business process starts by allowing users to search by borrower or loan application number. System will return the latest approved loan application for the borrower or the loan. But specific complication arose as a 'change in requirement' when validation required recursive search whether all the related parties (borrower, borrower-spouse, authorize signature person, all the parties in collateral) in the latest application having any in-progress loan application within the bank. For example, a borrower may be a co-borrower or a guarantor in multiple loans. So, a change of borrower's name should not impact other loan. After retrieving all results in the background of the latest loan, all data will be auto-populate, users can register the change of condition loan process with minimum to key-in and validation is ensured. After, the loan operation reiterates the same steps as 4.1 – 4.3. Due to complexity of this recursive validation requirement, the project tried multiple ways to achieve the objective and stress tests to prevent future system failure. Hence, the iteration was delayed 76.

E. Iterations 5 (Risk Analysis and Interface)

The fifth iteration started on 16/07/2012 and finished on 18/03/2013 taken 176 day. However, every two weeks some rules and interfaces were delivered in accordance with the previous iteration sprints. Significant usages of BRE are used to empower business users to quickly create and manage underwriting rules with minimal involvement from IT staff. To facilitate the mortgage underwriting process, reduce costs, and promote consistency for all loans, ''credit scoring'' models have been developed that numerically weigh or ''score'' some or all of the factors considered in the underwriting process and provide an indication of the relative risk posed by each application (see TABLE III). At the time, the bank's mortgage underwriting rules had over 2,600 rules. So, four dedicated developers and three business users involved in the underwriting and analytic score development. BRE and its complex calculations were also involved such as loan to value, loan term, loan history, delinquent and etc. During the development, we also found with many accounts and debt history can result in a long processing. So, a dedicated BRE server was employed (see Fig. 2 circled number 5).

In the loan process, there were needs for real time online integration with 13 external systems and databases such as credit bureau, deposit, debt, fraud and so on. The application controlled the response sent in turn to a response received from bank systems via web service. Information was mapped directly to loan application properties or parsed and transformed. The application servers served as the endpoint for an external connection—as a means to provide data to other systems (see Fig. 2). Additionally, at night time, for non real time data mortgage system sent/received file integration via SFTP with external systems e.g. enterprise data warehouse (EDW), insurance, human resource, anti-money laundry and etc. Time was also a constraint as users finished work at midnight only 6 hours permit to complete file transfer. As there were many loan applications only new and updated loans information were extracted and sent to the external systems. A dedicated agent node and SFTP was utilized (see Fig. 2 circled number 6).

TABLE III. PARTIAL EXAMPLE OF APPLICATION SCORE MODEL

Mortgage and characteristic	**Credit score**							
	Mean	Median	Low		Medium		High	
			% of characteristic	% of score range	% of characteristic	% of score range	% of characteristic	% of score range
Loan-to-value ratio (%)								
< 81	817	842	3.8	86.8	8.2	85.5	88.1	86.8
81 to 90	801	821	3.7	12.6	8.8	13.6	87.5	12.6
> 90	770	782	3.4	0.6	12.8	1	83.9	0.6
Loan size								
2 M	814	840	4.4	44.2	8.3	37.7	87.3	37.5
2 M –5 M	812	836	3.9	39.1	8.7	39.9	87.5	37.8
> 5 M	819	840	2.6	16.8	7.6	22.4	89.8	24.7
Location characteristic								
ZIP code								
< 80	788	811	5.5	13.4	12.6	13.9	81.9	8.5
80 to 120	811	836	4.2	52.9	8.6	50.1	87.2	47.5
> 120	824	847	2.9	33.6	6.9	36	90.1	44.1
….	….	….	….	….	….	….	….	….

The SFTP file integration with other systems was executed last, as these was between system to system and the assumption that business application needed to be processed correctly first. However, during the test, it was continuously detected by related systems that our interface data caused their systems to collapse regularly. For examples with the insurance system, firstly, date of birth need to be in Christian year format "yyyy-mm-dd". So, the software needed to convert the Thai data before transferring.

Secondly, insured company document identification number needed to be 13 digits, so there was a need for screen formatter and validation to ensure users have entered data correctly. Thirdly, KYC Level the system needed to set a default level value at 100 (low risk). Fourthly, in the case of selecting a life insurance of a particular company, we needed to set a sum insured to be greater than 0. These file integrations significantly impacted the project timeline, as meetings and agreements with IT and business owners were required to agree upon the inputs and signed-off. These systems have some regulations that they have to comply and cannot change. As a result of these new findings, various parts of the software were revised e.g. data format and types, business rules and screens. SIT, UAT and regression tests were done resulting in 124 days delay in total.

V. EVALUATIONS OF SCRUM AND MORTGAGE APPLICATION

The research aimed to answer five areas: 1) project performance with schedule and cost; 2) management acceptance; 3) customer relationship; 4) team satisfaction and 5) usability acceptance.

For project performance, TABLE IV shows actual effort and duration, the last column highlighted differences from the plan. The actual efforts were over two thousands man hours (33%) more than estimated and 200 days delayed resulting in 300,000 USD over budget. Therefore, top senior management decided that in the next phase a turn-key project managed by a vendor will be utilized in order to manage the cost. Agile will no longer be favored.

TABLE IV. ACTUAL EFFORT AND DURATION

Process	Plan (hr)	Duration (day)	Act. Work (hr)	Act. Duration (day)	Diff (hr)
Application Registration	1,332	16	420	35	912
Application Processing	720	12	774	29	-54
Underwriting and Closing	480	11	4,164	59	-3,684
User Management and Change of Condition	2,256	32	790	76	1,477
Risk Analysis and Interface	2,736	52	3,885	176	-1,149
Total	7,524	123	10,033	375	-2509

A week after the production date (3/06/2013), assessment to find customer relationship and team satisfaction with Agile, an interview with project sponsor, was conducted. She was happy with Agile as its approach promoted teamwork, facilitated the deliveries of periodic working software. However, she was disappointed with the control over costs. From working team (PM, Dev, BA, SME, testers) points of views towards the project (with writing comments on sticky note colored in green and red to represent 'good' and 'no good'), the result shows that developers liked the Agile approach, new tool and office environment. They felt the team was congenial (outsources and in-house developers) and were comfortable on working and understanding with each other despite language challenges. Developers felt SME and business leads are open to discussions, able to explain and clarify queries on existing business processes. However, the team shared similar negative opinions: 1) high resource turnover, resulting in substantial time and effort spent on orientation and initiation of new resources; 2) There are many situations where changes to requirements were made without analysis and approvals from stakeholders "*Better utilization of time and effort could have been achieved if there was a more comprehensive process in assessing suitable resources*". "*Change management process needs further refining and governance*"; 3) Project status updates in weekly meetings with senior management of outsource companies were often postponed.

In terms of interface usability evaluation, fourteen participants, from all units, 7 males and 7 females, voluntarily participated. Their ages ranged from 22 to 31 years with a mean age of 25 (std. = 0.48). LOS was given an overall usability of 58% which is considered above average. The other factors met the standard requirement of usability scales (see **Error! Reference source not found.**). For each scale, the median value is shown circled in the middle of the line; the 95% confidence levels are shown by the opening and closed points. These limits mean that we can be 95% certain that true scale median for the software can be found. LOS made the circles over 50% line, except for the Efficiency scale showing that users felt the software and navigation were complicated.

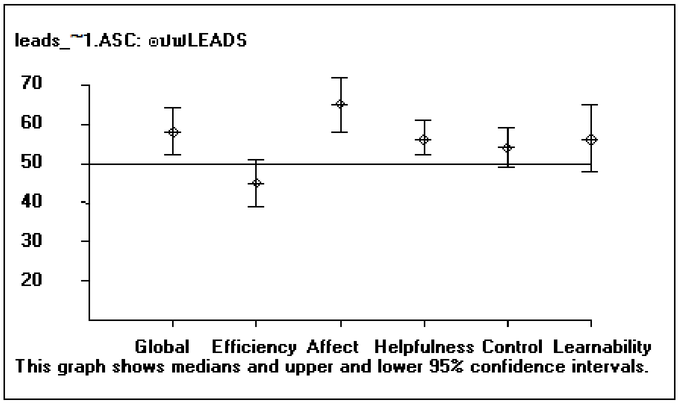


Fig. 3. SUMI Result

The main conclusion for each of the Sub-scales is summarized in TABLE V.

TABLE V. SUMMARY OF SUMI RESULTS

Subscale	Main Results
Efficiency	LOS was complicated to navigate. It required too many interactions (text inputs, buttons and conditions) to achieve an intended task. The software is robust and sufficient to work in a network environment. Even though, the software did take the issues of sensitive data into account, many of the save buttons were too many. The users need to save of sensitive data for any modification because of Cloud.
Affect	The users were satisfied with working with this software and did not feel tense while using it. Still, the presentation needs to be improved.
Learnability	The interface is informative; most functions of menu and buttons represented what it did quite clearly except for the sensitive data.
Control	The software was fast and robust. The user could move from one part to another fairly easily. However, there were too many clicks and keystroke and the user felt they were not in control.
Helpfulness	The help file was informative but some texts were difficult for the staff. The error and software messages were adequate. Each screen had its own help presentation.

From the SUMI questionnaire, LOS was usable (> 50). Overall end users had no problem in operating the software. This indicated that interface's factors such as clearly seen buttons and layout of UI elements received positive feedback from the end users.

VI. Discussion and Conclusion

This paper aims to contribute with an understanding of agile development failure in a large scale project, by identifying learning lessons which may contribute to other financial systems and other complex domains. Importantly, this study shows that Agile and the project was not failed, but due to political pressure on reducing the effort by senior management (Stage 3) which aligned with previous studies [7] [8]. Currently, Agile is not adopted by the bank as a result of effort, timeline and cost overruns. Waterfall model is currently employed. However, TABLE VI shows that if using the initial estimated (last column); the total project effort was in a safe zone (over estimated by 1447 man-hours or 12%). Indeed, agile should have been adopted, if efforts were not determined by senior management.

TABLE VI. Comparison of Actual and Initial Planning

Process	Political	Actual	Team plan
1. Application Registration	1,332	420	440
2 Application Processing	720	774	380
3 Underwriting and Closing	480	4,164	3,296
4 User Management and Change of Condition	2,256	790	502
5 Risk Analysis and Interface	2,736	3,885	6,862
Total	7,524	10,033	11,480

Even though, one of the main principles of agile methods is to "welcome changing requirements" [4] [7] [12] however this research showed that changing requirements especially with technically complicated challenge (Iteration 4) can contribute to extended timeline and cost [9]. The project appeared to have timeline and budget is strictly determined by senior management, then it implies that working over-time is mandatory. In this project, developers worked over-time on a regular basis to meet the political forced timeline. Their efforts were mostly un-clocked and un-billed to help the project and team. Therefore, the actual hours of over-time were not recorded. Consequently, six outsources and two local staff resigned adding problems to the project.

Rigid timeline increased pressure, when timeline is restricted, poor planning and analysis of related interface systems (Iteration 5) were not focused causing re-works significantly and miss of timeline for the project. Future empirical research is needed to investigate under which related interface systems and their messages should be reconciled in terms of their required fields to avoid reworks and delay of project. For example, if a middle named is a required field in a customer information system, then the field is a required field in mortgage system.

As reported in the PMKBOK [4] lack of fulltime SME staff was an important side effect of the detected problems (Stage 1), leading to project delays and failure. Besides, the previously mentioned problems potentially linked to political force and project management, there were additional difficulties during software development. At the beginning of Stage 3 of our project, the local team was novice. Employing new technology in any project implies certain inherent risks. Although, a training period to use BPM/BRE tool was carried out in Stage 2, it resulted insufficient; since there were areas of the tool to serve complicated mortgage requirements. Special care was employed by *side-by-side programming* practice between local and outsource developers. This turns beneficial to outsource as well for handling complex nature of mortgage domain where multiple disciplines interact [1] and specific Thai mortgage rules. It is therefore recommended that local and foreign staff sitting together.

Efficient communication is one of the key issues of Agile [3] and frequent delivery of tested working software in an iterative way brought high visibility of project progress [10]. The email update of bug fixes and project status update for every two weeks too all stakeholders give direct feedback to development amelioration. This context helped to share the 'big picture' of the project state and to build a strong camaraderie and team spirit, which definitely were the key drivers to sustain focus and commitment during all stages.

Although, the SUMI score was adequate, a range of interface problems was uncovered. Firstly, the major problem arising from the SUMI analysis related to the 'Efficiency' scale (46%). It was found that the users were required to key-in substantial data in one single page, check data synchronization between tabs and sections in order to complete the procedure. For example, firstly, in financial data section, users were not allowed to step next if they had not completed typing in eight tabs of the account. Secondly, the application instructed the user to save sensitive information in many places (see Fig. 6). In the next version of LOS in terms of Efficiency scale, therefore the software will minimize the number of save button. Also, collapsible section will be used (see Fig. 7) without displaying all fields[1]. This may reduce the user perception that they have to key-in all.

One of the limitations of this research study was the constitution of the sample i.e. mortgage application specific. Nonetheless, mortgage was only part of the activities of this research in agile development where an institute attempted to adopt Agile. BPM/BRE tool used is proprietary where line of code cannot be counted to assess software size. Additionally, from the authors' knowledge, the BPM/BRE cloud-based loan origination system is the first experience in the world. So in terms of Koch Agile evaluation [10] there is no benchmarking available to compare in terms of speed of delivery, software size, performance, robustness or adaptation of Agile in banking industry. Therefore, the results might not generalize to other agile development, particularly those in different culture or free of political influences.

While the need for agile approach has been widely recognized, making an agile approach work in a long established waterfall bank is challenging. A number of recommendations can be drawn; firstly even a political pressure on the effort estimated by the developers, still in the end the actual effort reflects the early estimate. Secondly, inexperience developers may not be able to design a core engine of BPM/BRE for the bank. Tool may be needed.

[1] At the time of this writing, it has been implemented as a result of SUMI and interviews where users requested for screen redesigned changes

Thirdly, constantly changing requirements increases difficulties and workload during development where timeline cannot be changed producing a significant dissatisfaction from developers and the sponsor. The research revealed several paradox-like phenomena that need further research and investigation. The agile method was found feasible in a large project; within the team Scrum was highly effective in rescuing this mortgage web-based project. The use of Scrum resulted highly positive at working team since it improved the communication between team members and as a consequence increases the team flexibility and productivity and maintaining focus on those tasks more relevant to the project.

ACKNOWLEDGMENT

The author is very grateful to Anders Markvardsen for continued reviews of this paper.

REFERENCES

[1] K.Temkin, D. Levy and D. Levine, "A Case Study of the Mortgage Application Process”," in *Mortgage Lending Discrimination: A Review of the Evidence*, Washington DC,, Urban Institute Press, 1999, pp. 137-160.

[2] R. Avery, R. Bostic and P. Calem, "Credit Risk, Credit Scoring, and the Performance of Home Mortgages," Federal Reserve Bulletin July 1996, Washington, 1996.

[3] L. Williams, "What Agile Teams Think of Agile Principles," *Communications of the ACM,* vol. 55, no. 4, pp. 71-76, 2012.

[4] B. Jordan, J. Giboney, M. Keith, J. Mark, D. Wilson, R. Schuetzler, P. Lowry and A. Vance, "Overview and Guidance on Agile Development in Large Organizations," *Communications of the Assosiation for Information Systems,* pp. 25-44, July 2011

[5] H. Mohammad, T. Alwada and J. Ababneh, "Agile Software Methodologies: Strength and Weakness," *International Journal of Engineering Science and Technology,* vol. 5, no. 3, pp. 455-459, 2013

[6] D. Batra, W. Xia, D. VanderMeer and K. Dutta, "Balancing Agile and Structured Development Approaches to Successfully Manage Large Distributed Software Projects: A Case Study from the Cruise Line Industry," *Communications of the Association for Information Systems,* vol. 27, no. 21, pp. 378-394, 2010.

[7] J. Wan, Y.Zhu and M. Zeng, "Case Study on Critical Success Factors of Running Scrum," *Journal of Software Engineering and Applications,,* vol. 6, pp. 59-64, 2013.

[8] M. Santos, P. Bermejo, M. Oliveira, A. Tonelli and E. Seidel, "Improving The Managment of Cost and Scope in Software Projects using Agile Practices," *International Journal of Computer Science & Information Technology,* vol. 5, no. 1, pp. 47-64, 2013

[9] S. Keaveney and K. Conboy, "Cost estimation in Agile Development Projects," in *ECIS*, 2006.

[10] S. Kruba, S. Baynes and R. Hyer, "BPM, Agile, and Virtualization Combine to Create Effective Solutions," *International Journal of Advanced Computer Science and Applications,* 2012.

[11] K. G. Royce, "Integration of a Business Rules Engine to Manage Frequently Changing Workflow: A Case Study of Insurance Underwriting Workflow," in *Americas Conference on Information Systems*, Colorado, 2007.

[12] A. S. Koch, Agile Software Development Evaluating the Methods for Your Organization, Boston: Artech House, 2005.

[13] T. Silva, M. Silveira, F. Maurer and T. Hellmann, "User Experience Design and Agile Development: From Theory to Practice," *Journal of Software Engineering and Applications,* vol. 5, pp. 743-751, 2012.

[14] A. Narasimhadevara, T. Radhakrishnan, B. Leung and R. Jayakumar, "On Designing a Usable Interactive System to Support Transplant Nursing," *Journal of Biomedical Informatics,* vol. 42, p. 137–151, 2008.

[15] G. G. Angel, PMP Certification A Beginner’s Guide, New York: McGraw-Hill, 2010.

APPENDIX

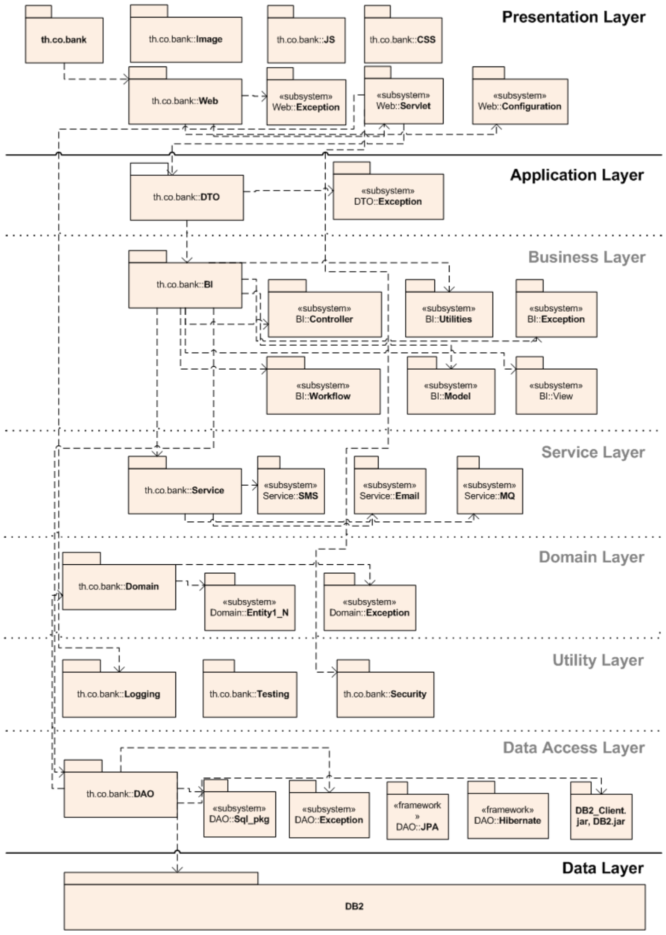
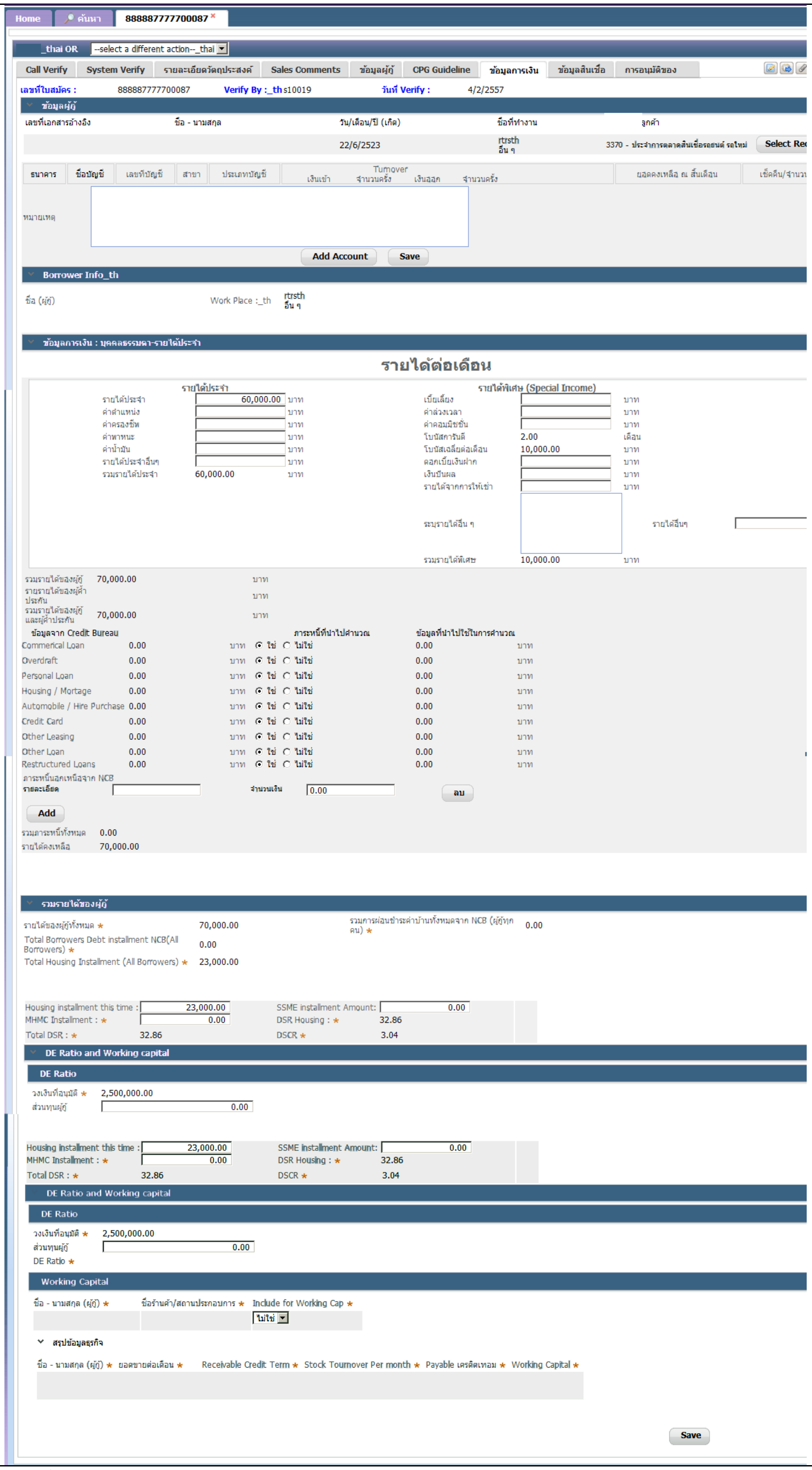


Fig. 4. Three Tier Architecture of Mortgage Application



เลขที่ใบสมัคร : 888887777700087 Verify By :_th s10019 วันที่ Verify : 4/2/2557

ข้อมูลผู้กู้

เลขที่เอกสารอ้างอิง	ชื่อ - นามสกุล	วัน/เดือน/ปี (เกิด)	ชื่อที่ทำงาน	ลูกค้า
		22/6/2523	rtrsth อื่น ๆ	3370 - ประจำการตลาดสินเชื่อรถยนต์ รถใหม่

ธนาคาร	ชื่อบัญชี	เลขที่บัญชี	สาขา	ประเภทบัญชี	Turnover เงินเข้า	 จำนวนครั้ง	 เงินออก	 จำนวนครั้ง	ยอดคงเหลือ ณ สิ้นเดือน	เช็คคืน/จำนวน

หมายเหตุ

Borrower Info_th

ชื่อ (ผู้กู้) Work Place :_th rtrsth อื่น ๆ

ข้อมูลการเงิน : บุคคลธรรมดา-รายได้ประจำ

รายได้ต่อเดือน

รายได้ประจำ			รายได้พิเศษ (Special Income)		
รายได้ประจำ	60,000.00	บาท	เบี้ยเลี้ยง		บาท
ค่าตำแหน่ง		บาท	ค่าล่วงเวลา		บาท
ค่าครองชีพ		บาท	ค่าคอมมิชชั่น		บาท
ค่าพาหนะ		บาท	โบนัสการันตี	2.00	เดือน
ค่าน้ำมัน		บาท	โบนัสเฉลี่ยต่อเดือน	10,000.00	บาท
รายได้ประจำอื่นๆ		บาท	ดอกเบี้ยเงินฝาก		บาท
รวมรายได้ประจำ	60,000.00	บาท	เงินปันผล		บาท
			รายได้จากการให้เช่า		บาท
			ระบุรายได้อื่น ๆ		รายได้อื่นๆ
			รวมรายได้พิเศษ	10,000.00	บาท

รวมรายได้ของผู้กู้ 70,000.00 บาท
รายรายได้ของผู้ค้ำประกัน บาท
รวมรายได้ของผู้กู้และผู้ค้ำประกัน 70,000.00 บาท

ข้อมูลจาก Credit Bureau			ภาระหนี้ที่นำไปคำนวณ	ข้อมูลที่นำไปใช้ในการคำนวณ	
Commerical Loan	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Overdraft	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Personal Loan	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Housing / Mortage	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Automobile / Hire Purchase	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Credit Card	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Other Leasing	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Other Loan	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท
Restructured Loans	0.00	บาท	● ใช่ ○ ไม่ใช่	0.00	บาท

ภาระหนี้นอกเหนือจาก NCB
รายละเอียด จำนวนเงิน 0.00

รวมภาระหนี้ทั้งหมด 0.00
รายได้คงเหลือ 70,000.00

รวมรายได้ของผู้กู้

รายได้ของผู้กู้ทั้งหมด * 70,000.00
Total Borrowers Debt installment NCB(All Borrowers) * 0.00
Total Housing Installment (All Borrowers) * 23,000.00
รวมการผ่อนชำระค่าบ้านทั้งหมดจาก NCB (ผู้กู้ทุกคน) * 0.00

Housing installment this time : 23,000.00
MHMC Installment : * 0.00
Total DSR : * 32.86
SSME installment Amount: 0.00
DSR Housing : * 32.86
DSCR * 3.04

DE Ratio and Working capital

DE Ratio

วงเงินที่อนุมัติ * 2,500,000.00
ส่วนทุนผู้กู้ 0.00

Housing installment this time : 23,000.00
MHMC Installment : * 0.00
Total DSR : * 32.86
SSME installment Amount: 0.00
DSR Housing : * 32.86
DSCR * 3.04

DE Ratio and Working capital

DE Ratio

วงเงินที่อนุมัติ * 2,500,000.00
ส่วนทุนผู้กู้ 0.00
DE Ratio *

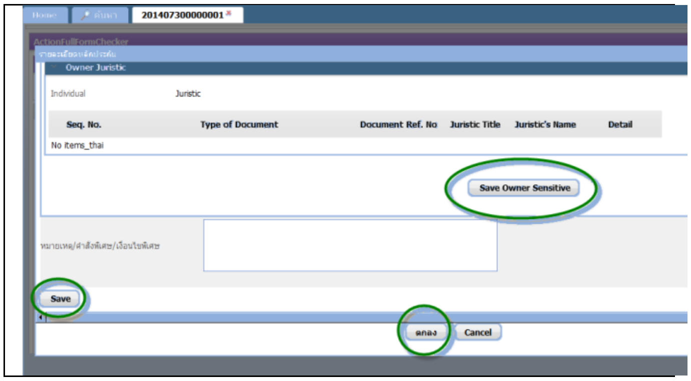
Working Capital

ชื่อ - นามสกุล (ผู้กู้) *	ชื่อร้านค้า/สถานประกอบการ *	Include for Working Cap *
		ไม่ใช่

สรุปข้อมูลธุรกิจ

ชื่อ - นามสกุล (ผู้กู้) *	ยอดขายต่อเดือน *	Receivable Credit Term *	Stock Tournover Per month *	Payable เครดิตเทอม *	Working Capital *

Fig. 5. Screen Example

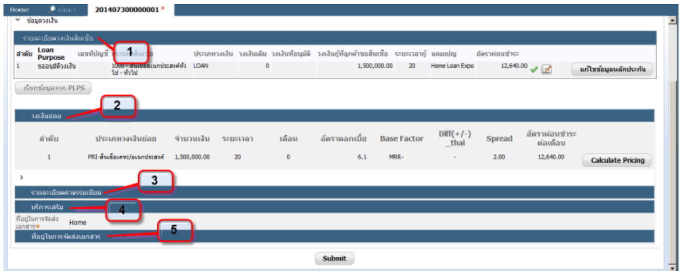


Individual Juristic

Seq. No.	Type of Document	Document Ref. No	Juristic Title	Juristic's Name	Detail
No items_thai					

หมายเหตุ/คำสั่งพิเศษ/เงื่อนไขพิเศษ

Fig. 6. Save Buttons



ลำดับ	Loan Purpose	เลขที่บัญชี	ประเภทสินเชื่อ	ประเภทวงเงิน	วงเงินเดิม	วงเงินที่อนุมัติ	วงเงินกู้ที่ลูกค้าขอสินเชื่อ	ระยะเวลากู้	แคมเปญ	อัตราผ่อนชำระ
1	ขออนุมัติวงเงิน		1000 - สินเชื่ออเนกประสงค์ทั่วไป - ทั่วไป	LOAN	0		1,500,000.00	20	Home Loan Expo	12,640.00

ลำดับ	ประเภทวงเงินย่อย	จำนวนเงิน	ระยะเวลา	เดือน	อัตราดอกเบี้ย	Base Factor	Diff(+/-)_thai	Spread	อัตราผ่อนชำระต่อเดือน
1	PRI-สินเชื่อเคหะ/อเนกประสงค์	1,500,000.00	20	0	6.1	MRR-	-	2.00	12,640.00

ที่อยู่ในการจัดส่งเอกสาร* Home

Fig. 7. Collapsible Sections

XML Schema-Based Minification for Communication of Security Information and Event Management (SIEM) Systems in Cloud Environments

Bishoy Moussa
Information Technology Department
Faculty of Computers and
Information, Helwan University
Cairo, Egypt

Mahmoud Mostafa
Information Systems Department
Faculty of Computers and
Information, Helwan University
Cairo, Egypt

Mahmoud El-Khouly
Information Technology Department
Faculty of Computers and
Information, Helwan University
Cairo, Egypt

***Abstract*—XML-based communication governs most of today's systems communication, due to its capability of representing complex structural and hierarchical data. However, XML document structure is considered a huge and bulky data that can be reduced to minimize bandwidth usage, transmission time, and maximize performance. This contributes to a more efficient and utilized resource usage. In cloud environments, this affects the amount of money the consumer pays. Several techniques are used to achieve this goal. This paper discusses these techniques and proposes a new XML Schema-based Minification technique. The proposed technique works on XML Structure reduction using minification. The proposed technique provides a separation between the meaningful names and the underlying minified names, which enhances software/code readability. This technique is applied to Intrusion Detection Message Exchange Format (IDMEF) messages, as part of Security Information and Event Management (SIEM) system communication hosted on Microsoft Azure Cloud. Test results show message size reduction ranging from 8.15% to 50.34% in the raw message, without using time-consuming compression techniques. Adding GZip compression to the proposed technique produces 66.1% shorter message size compared to original XML messages.**

Keywords—XML; JSON; Minification; XML Schema; Cloud; Log; Communication; Compression; XMill; GZip; Code Generation; Code Readability

I. Introduction

XML-based communication governs most of today's systems communication, due to its capability of representing complex structural and hierarchical data. However, XML document structure is considered a huge and bulky data that can be reduced to minimize bandwidth usage, transmission time, and maximize performance. This contributes to a more efficient and utilized resource usage. In cloud environments, this affects the amount of money the consumer pays. Several techniques are used to achieve this goal. This paper discusses these techniques and proposes a new XML Schema-based Minification technique. The proposed technique works on XML Structure reduction using minification. The technique separates the original structure names from the minified names, to better achieve code readability while reducing data sent in the wire. This technique is applied to Intrusion Detection Message Exchange Format (IDMEF) messages, as part of Security Information and Event Management (SIEM) system communication hosted on Microsoft Azure Cloud.

This paper starts with an overview of the key concepts, required throughout the paper in section II. Section III presents related work. Then, section IV introduces the proposed system architecture and the minification process. After that, two experiments and test results are presented in section V. Conclusively, the proposed solution is discussed, and ideas for future work are suggested in section VI.

II. Key Concepts

A. XML-based communication

1) XML, DTD, and XSD

Extensible Markup Language (XML) is a data representation technique used to represent structural and hierarchical data. An XML document is composed of a set of nested nodes with only one starting node. Each node may have a number of attributes. XML document is defined by a Document Type Definition (DTD), or alternatively, an XML Schema Definition (XSD). DTDs and XSDs define the structure of the corresponding XML document, the number and type of children nodes included within any node, and some validations and constraints regarding each attribute values or possible combination of children nodes [1].

XML message structure is very lengthy and redundant. Figure 1 shows a sample Intrusion Detection Message Exchange Format (IDMEF) Heartbeat message in XML. The Bold nodes and attributes represent redundant and descriptive structure elements that are sent with each message.

2) XML Schema Definition (XSD) Components

The building block in XML schema is **Element**, because it is directly mapped to an XML node. Element has a *name* attribute (representing the XML node name) and a *type* attribute (representing the XML node data type). XML schema types can be primitive types, found in XML Schema namespace, (e.g. integer, string, etc.) or new types, defined in other user-defined schemas. Schema types are categorized into two different categories; **simple types** (types composed of a single element), and **complex types** (types composed of multiple elements). Schema **Attribute** node defines an attribute of an XML node. Similar to **Element**, **Attribute** node

```
<?xml version="1.0" encoding="UTF-8"?>
  <idmef:IDMEF-Message version="1.0"
xmlns:idmef="http://iana.org/idmef">
    <idmef:Heartbeat messageid="abc123456789">
      <idmef:Analyzer analyzerid="hq-dmz-analyzer01">
        <idmef:Node category="dns">
          <idmef:location>Headquarters DMZ
Network</idmef:location>
          <idmef:name>analyzer01.example.com</idmef:name>
        </idmef:Node>
      </idmef:Analyzer>
      <idmef:CreateTime
ntpstamp="0xbc722ebe.0x00000000">2000-03-09T14:07:58Z
      </idmef:CreateTime>
      <idmef:AdditionalData type="real" meaning="%memused">
        <idmef:real>62.5</idmef:real></idmef:AdditionalData>
      <idmef:AdditionalData type="real" meaning="%diskused">
        <idmef:real>87.1</idmef:real></idmef:AdditionalData>
    </idmef:Heartbeat>
  </idmef:IDMEF-Message>
```

Fig. 1. Sample Intrusion Detection Message Exchange Format (IDMEF) XML – Heartbeat message (with redundant data in bold) [2].

has *name* and *type* attributes. Schema **Enumeration** node defines a single possible value for the specified type. All of the above components form the XSD, which is defined by a *Target Namespace*. It is easier to think of the *Target Namespace* as a name governing the current schema such that there should not be two similar sibling schema items of the same name. XML Schemas can reference other schemas via two types of tags / nodes; **Import**, and **Include**. Both of them has *schemaLocation* attribute, describing the location of the Schema to reference. The difference between schema **Import** and schema **Include** is schema **Import** allows importing other schemas of different target namespace, while schema **Include** allows importing schemas of the same target namespace. Therefore, schema **Import** must specify the imported schema target namespace via *namespace* attribute [1].

B. Serialization and Deserialization

Serialization is the process of converting complex data objects into a serial format, before sending it via transmission medium. Deserialization is the process of converting the received serial format to its original complex data objects, in order to make it ready for direct member access via code. Serial format may include Binary Stream (Byte Array), XML, or JavaScript Object Notation (JSON) [3]. In order to realize the serialization and deserialization processes, a mapping between the data object and the serial format is essential. Members that can be serialized and deserialized are marked. Serializers and Deserializers are implemented to convert data objects to and from the serial format, respectively. Examples of Serializers and Deserializers are Memory serializers (serial format is Byte Array), XML serializers (serial format is XML message), and JSON serializers (serial format is JSON message). Serialization of complex objects is done recursively for each object member, until primitive data type is found (e.g. integer, float, double, character, etc.).

C. Cloud Computing and Service Models

Cloud computing is based on providing consumers with different services in an elastic and measurable way. So that, consumers only pay for their usage of different computing resources. They still get the benefits of elastic resources, which can expand or shrink based on requests load. Cloud computing offers different service models. It includes Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), Security as a Service (SecaaS), and other Emerging Services [4].

IaaS model provides consumers with different types of resources (e.g. storage, network, and processing power). Consumers are required to build their own platform (operating system installation and configuration, and development runtime environment (RTE) installation), and application software. PaaS model is built on top of IaaS model. It provides consumers with different types of resources, and platform. Consumers are required to build their own application software. SaaS model is built on top of PaaS model. It provides consumers with different types of resources, platform, and specific software. Consumers are required to create accounts and use the offered software. Pricing is measured per account or resources usage. SecaaS model provides consumers with security-related solutions for any environment [5]; e.g. Logging Solutions (which are used to centralize logging), and Security Information and Event Management (SIEM) systems (which are complete solutions for providing security and events information storage, normalization, correlation and analysis, incident reporting, and incident interaction) [6]. Emerging service models are new services. They include Financial Software as a Service (FSaaS) model, Health Informatics as a Service (HIaaS) model, and Education as a Service (EaaS) model.

As in Figure 2, SIEM systems collect security and events information from different sources via sensors. Most information is represented in the form of formats/protocols; e.g. Syslog, IDMEF, Common Event Expression (CEE), and Simple Network Management Protocol (SNMP). Most of these protocols are based on XML [7].

Syslog is used to send log information. It is based on simple plain text; no structured format is used. It is difficult to represent structured, complex data using Syslog [8]. CEE is XML-based format, used to represent log and audit data. It also allows an organization to demonstrate compliance with audit requirements [9]. SNMP is a protocol for managing devices on IP networks. It is used for status monitoring, and configuration of network devices [10]. IDMEF is used to report an Intrusion Detection System (IDS) alert, or a device status as a heartbeat. IDMEF is based on XML. It supports structured and complex data. It also supports XML/XSD extensions, to cover any needed extra information that is not supported by the current specification of IDMEF [2]. Because of the previously mentioned benefits of IDMEF, IDMEF is selected for the study.

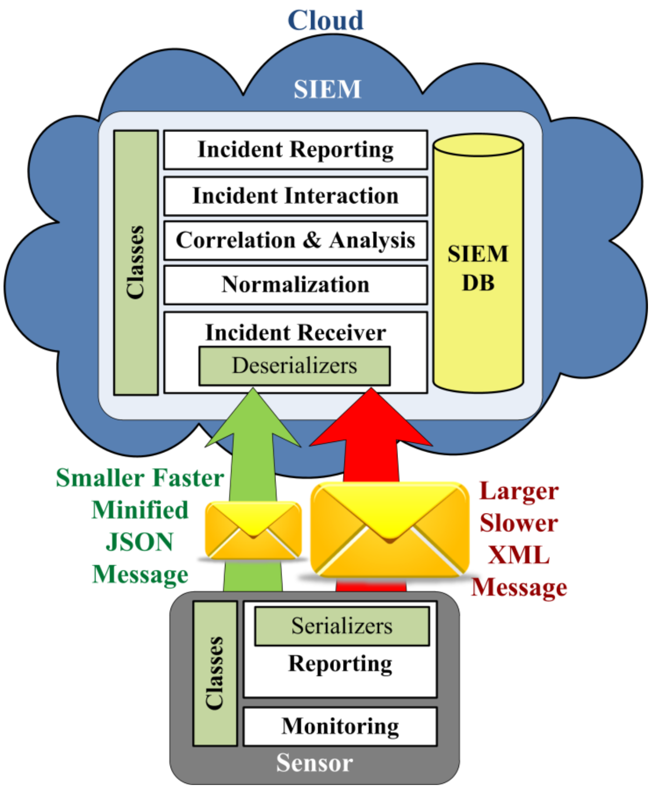


Fig. 2. Security Information & Event Management (SIEM) System Components and Communication (proposed components are in light green).

III. Related Work

Related work covers different topics. Attempts to reduce the unnecessary white spaces in XML are discussed. A lighter format (JSON) is used in different web systems communications. Then, the concept behind reduction in JSON is introduced. After that, the advantages and disadvantages of parsing different message types are discussed. Finally, time-consuming compression techniques are presented.

A. XML Minification

XML messages are built based on hierarchical structure. It is common to represent them with tabs or spaces to add indentation to enhance readability. Unfortunately, these whitespace characters increase message size, regardless of the huge amount of data maintained to store structure (e.g. opening and closing tags with descriptive names).

XML Minification techniques aim to reduce message size; however, most techniques are focused on whitespace characters, and comments removal. Advanced minifiers can collapse tags that does not have content; e.g. "<idmef:real></idmef:real>" is changed to "<idmef:real/>". Examples of XML Minifiers include THE XML MINIFIER (http://www.nathanael.dk/tools_thexmlminifier.php) and WEB <MARKUP> MIN - XML Minifier (http://webmarkupmin.apphb.com/ minifiers/xml-minifier).

B. XML vs. JSON

JavaScript Object Notation (JSON) is another data exchange format. It is lighter than XML, and easier to generate and parse by machines. It is commonly used in web systems communications. It is recommended for data communication due to its performance and message size [11] [12].

Figure 3-a shows the sample IDMEF message (of Figure 1) after conversion to JSON format. The XML message size, in Figure 1, is 686 bytes. While the JSON message size, in Figure 3-a, is 403 bytes.

JSON message is composed of a single parent object. Objects are enclosed by curly braces "{}". Objects are composed of members. Each member has a member name and value, separated by a colon ":". Member names are strings, enclosed by double quotes """". Member Values can be of simple data type, like integers (e.g. 2), strings (e.g. "dns"), or date-time (e.g. "2000-03-09T14:07:58Z"). Member Values can also be of complex data type (e.g. instance of another complex data type). Different members within an object are comma-separated ",". Array items are enclosed by square brackets "[]", with a comma separating each two consecutive items.

```
{
 "IDMEF-Message": {
   "Heartbeat": {
     "messageid": "abc123456789",
     "Analyzer": {
       "analyzerid": "hq-dmz-
analyzer01",
       "Node": {
         "category": "dns",
         "location": "Headquarters
DMZ Network",
         "name":
"analyzer01.example.com"
       }
     },
     "CreateTime": {
       "ntpstamp":
"0xbc722ebe.0x00000000",
       "value": "2000-03-
09T14:07:58Z"
     },
    "AdditionalData": [
      {
        "type": "real",
        "meaning": "%memused",
        "real": "62.5"
      },
      {
        "type": "real",
        "meaning": "%diskused",
        "real": "87.1"
      }
     ]
   }
  }
}
```

(a)

```
{
 "a": {
   "a": {
     "a": "abc123456789",
     "b": {
       "a": "hq-dmz-analyzer01",
       "b": {
         "a": "dns",
         "b": "Headquarters DMZ
Network",
         "c":
"analyzer01.example.com"
       }
     },
     "b": {
       "a":
"0xbc722ebe.0x00000000",
       "b": "2000-03-09T14:07:58Z"
     },
     "c": [
       {
         "a": "real",
         "b": "%memused",
         "c": "62.5"
       },
       {
         "a": "real",
         "b": "%diskused",
         "c": "87.1"
       }
     ]
   }
  }
}
```

(b)

Fig. 3. Representation of IDMEF message in Figure 1: (a) JSON representation; (b) Proposed Minification with JSON representation.

C. JavaScript Minification

JavaScript Minification is most common in websites development and websites optimization for mobile access. It is preferred as a finalization step after development completion and before website deployment. Minification offers the following benefits: (1) File size reduction, which will minimize transmission time and network latency. (2) Faster handling and processing. (3) Minified files are better candidates to compression techniques, resulting in higher compression ratios [13]. Trivial minification includes comments, and whitespace characters removal (tabs, spaces, new lines, carriage returns, etc.). Some advanced minifiers do a more complex step, which is renaming variables, as shown in Figure 4. Examples for JavaScript Minifiers include JSCompress (http://jscompress .com), YUI Compressor (http://refresh-sf.com/yui/), and javascript-minifier (http://javascript-minifier.com).

D. Code Generation

Parsing and generating XML document manually is error prone. Some parsers work based on strings; e.g. element extraction is based on its name string, and setting element value is passed as a string, no matter what element data type is (http://search.cpan.org/~erwan/XML-IDMEF-0.11/IDMEF. pm)

Code generation is used to generate object oriented classes that map the corresponding XML messages based on their schemas. Messages are based on objects serialization, whereas objects are based on messages deserialization. The benefits of using code generation are: (1) Faster development time; intelligent Integrated Development Environments (IDEs) offer code auto-completion (in Microsoft Visual Studio, it is called IntelliSense), that helps developers to find the wanted member (in this case, XML element) with minimum effort. (2) Correct reference of an XML element, since elements are object's members and no strings are used. Strings are vulnerable to spelling mistakes. (3) Correct typed values assignment restricts setting each element to its value according to its element data type, rather than setting elements values as strings.

To send and receive XML messages using classes, code generation tools exist. These tools are based on the corresponding XSD. Tools for Microsoft .NET Framework include Microsoft's XSD tool (http://msdn.microsoft.com/en-us/library/x6c1kb0s%28v=vs.110%29.aspx), the open source XSD2Code (http://xsd2code.codeplex.com), etc. Tools for C++ include XSD: XML Data Binding for C++ (http://www.codesynthesis.com/products/xsd/). Tools for Java include JAXB and XmlBeans (http://www.jetbrains.com/idea/ webhelp/generating-java-code-from-xml-schema.html).

E. Compression Techniques

XMill is a specialized XML compression technique. It compresses XML data by separating it into three components: The element and attribute names, the text values, and the tree structure of the XML document. The text values are grouped by parent element name. The three components are then compressed using standard text compression techniques [14].

```
function product(num1,num2)
{ return num1*num2; }
```

(a)

```
function product(n,r){return n*r}
```

(b)

Fig. 4. JavaScript Minification: (a) original sample function. (b) the same function after minification.

Dong Zhou implemented a Structure Extraction and Encoding technique. An XML Structure is extracted; an MD5 hashing function is used to get a unique structure ID, then receiver stores Structures with their IDs in a cache. Data are sent with no structure information, associated with Structure ID only [15]. The advantage of this technique is that it works generally on any XML. The disadvantages are: (1) Similar structures are treated as new structures with new Structure ID and stored as different instances in the cache; e.g. number of items in a list, optional node or attribute, etc. (2) it is based on a cache to be available. (3) The process is considered an overhead, especially if a cache-miss occurs. A good comparison between different XML compression techniques is introduced in references [16] [17] [18].

GZip is a general-purpose compression technique. It is widely used in HTTP communication due to its good performance and high compression ratio [19] [20]. It uses the DEFLATE algorithm [21].

Compression techniques are considered a conversion process, which means it has an overhead processing time before sending the message, and after receiving the message. Direct communication techniques with message size reduction are preferred.

IV. PROPOSED SOLUTION

Proposed solution applies the JavaScript Minification techniques to XSDs, which are used to generate code that does the serialization and deserialization of objects in the minified XML format. Furthermore, it can be used with JSON serializer/deserializer, in order to make use of JSON advantages. In this case, the output will be minified JSON messages. The solution applies names minification to the underlying data format. It does not affect the generated members' names. This reduces message size but maintains software code readability. Proposed solution is implemented using Microsoft .NET Framework in C# language.

A. Solution Architecture

In order to achieve this goal, the solution architecture (Figure 5) shows two main tools: XSDMinify and Code Generators. (1) XSDMinify: it works on an XML Schema file and applies structure names minification. It produces two files; the first is the minified XML Schema, and the second is a dictionary file mapping each minified element name to its original element path in the original XSD (Figure 6). This process is performed only once per original XSD or any of its referenced schemas change. (2) Code Generators: XSD2Code is an open-source code generator from an XSD. It generates serializable C# classes from an XSD. Some changes are made to support the minified XSD and dictionary files as input. This tool generates serializable data fields with the minified names and getter/setter properties/methods to get/set the data fields. The properties/methods names are based on the original, meaningful and descriptive names from the dictionary file (See Figure 7). For other programming languages, the

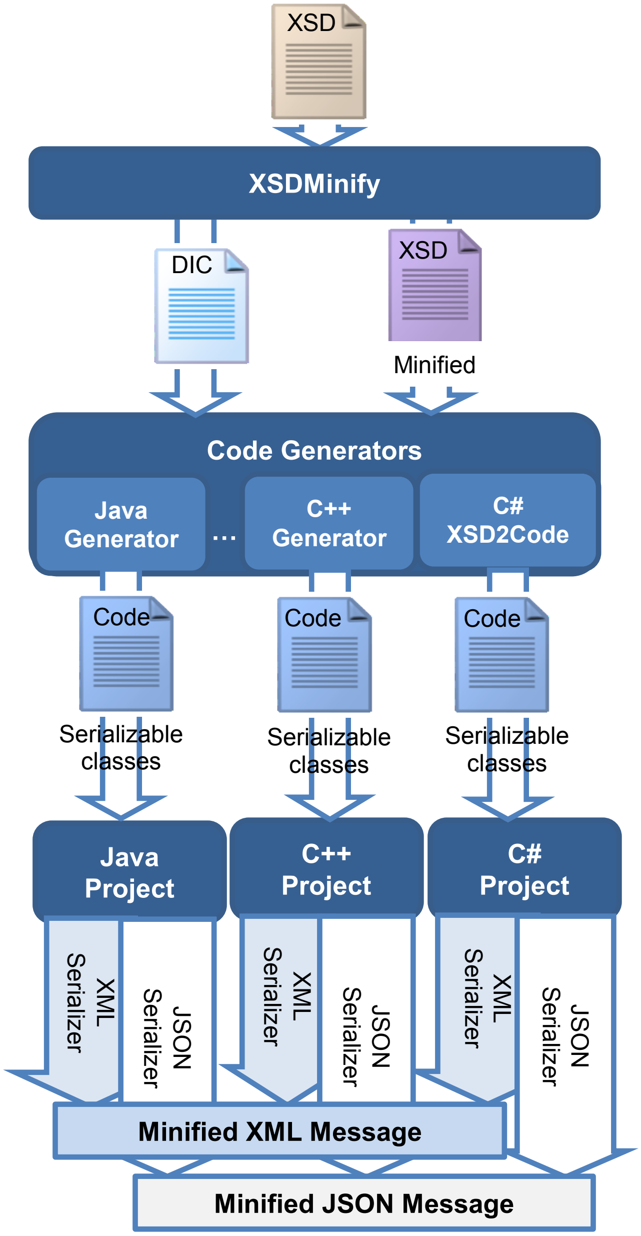


Fig. 5. Proposed solution architecture.

corresponding code generation tool needs to be customized to generate object oriented classes using the same technique.

As in Figure 2, the generated code is then included in the sender and receiver development projects. In this case, sender project represents a sensor, and receiver project represents SIEM system module. Typed messages are composed at the sender, serialized with any serializer (preferably JSON serializer), and transmitted to the receiver. The receiver receives the message, deserializes it using the appropriate deserializer. Now the message is ready for use as an object, at receiver's side.

```
a,xsd:schema/xsd:element[name=IDMEF-Message]
b,xsd:schema/xsd:element[name=Alert]
a,xsd:schema/xsd:complexType[name=IDMEF-
Message]/xsd:attribute[name=version]
```

Fig. 6. Sample of the dictionary (Generated from XSDMinify).

B. *XSDMinify*

XSDMinify is the tool that reduces the XML documents structure by applying schema structure names renaming. The original XSD file is the only input the tool requires. XSDMinify automatically detects schema Imports or schema Includes, fetches these referenced schemas, and applies minification to the referenced schemas first.

XSDMinify has two main passes for processing and minifying any XSD file. The first pass checks for schema Import or schema Include tags, then pushes the referenced schema in a stack. Therefore, the children/referenced schemas are at the top of the stack; while their parent/referencing schemas are at the bottom of the stack. The second pass represents the main minification process. In this pass, schemas are popped from the stack for minification. Schema's Target Namespace is detected, and a new Target Namespace is specified for the minified schema. Then, processing Import and Include tags is done through updating referenced Target Namespaces and schemas' new locations. This is followed by a search for any mention of the referenced schema, and an update with the minified names. After that, minification process of the current schema starts. Search for any node with "name" attribute or enumeration node with "value" attribute is carried out. An order generator generates new shorter names (e.g. a, b, c, etc. or 0, 1, 2, etc.) for each of the found name or value, respectively. Node path is also considered during order generation to allow reuse of short names. Such that, two nodes with different paths can have the same short name. A dictionary is built to store the short name mapping with the node path, and saved to a file with DIC extension (See Figure 5). Figure 6 shows a sample of the dictionary file; where short names are associated with its corresponding node path (starting from the root "schema" node to the leaf node). As processing original XSD continues, original names are replaced by the short names. Any references to the original names are updated as well. The changes are saved as the minified XSD (See Figure 5).

C. *Code generators (XSD2Code)*

XSD2Code is changed to handle code generation differently. Code Generation is based on the minified XSD and the dictionary (DIC) files, which are generated from the XSDMinify tool. As in Figure 7, Properties are generated with serializable short name fields, while property names with the

```
[DataMember(EmitDefaultValue
= false)]
private Analyzer a;
public Analyzer Analyzer {
        get { return this.a; }
        set { this.a = value; } }
```

(a)

```
public enum usercategory : uint {
  [XmlEnumAttribute("0")]
unknown = 0u,
  [XmlEnumAttribute("1")]
application = 1u,
  [XmlEnumAttribute("2")]
osdevice = 2u,}
```

(b)

Fig. 7. Sample of the generated code using modified XSD2Code: (a) Code Generation of a Property (XSD Element); (b) Code Generation of an Enum (XSD Enumeration).

original meaningful names are accessible through code. Similarly, Enumerations are based on integer series. These integer values are used in serialization while Enumeration members are the original meaningful name. This way, a typed and meaningful access to the object's properties is achieved, resulting in maintaining software code readability. However, shorter structure elements names are used for transmission.

V. Experiment and Results

A. *Experiment*

Generated Code using proposed tool (XSD2Code), and original technique (using Microsoft's XSD tool) is included in two projects: (1) First project is a desktop application, simulating software alert source/sensor. It generates alert messages and sends them to the receiver end (the second project). (2) Second project is a web application project, simulating SIEM system, which receives alerts via a web service, processes alerts, and calculates results statistics.

Two experiments are established to compare the proposed technique's message size reduction and performance. The first experiment compares the proposed technique against traditional XML messages. The second experiment "Compression" compares the proposed technique against XMill compressed messages and GZip compressed messages.

B. *Test Data*

Several types of IDMEF messages are used:

1) Empty Message: Almost empty message with necessary parts sent only (AnalyzerTime, CreateTime, DetectTime, and messageid fields only set.)

2) Full Message: IDMEF Message with all fields filled with appropriate data.

3) Sample IDMEF Message: IDMEF messages as represented in Examples section of the IDMEF protocol at IETF [2]. Samples are Tear Drop, ping of death, Port Scanning – 1 (Connection to a Disallowed Service), Port Scanning – 2 (Simple Port Scanning) , loadmodule – 1, loadmodule – 2, phf, File Modification, System Policy Violation, Correlated Alerts, Analyzer Assessments, and Heartbeat messages.

Analysis of message structure is performed, including Raw XML Message size, Total Nodes Count for the whole message, Total Attributes Count for the whole message, and XML Complexity / Levels (representing the number of levels for nesting nodes). Table I and Figure 8 show the results of this analysis. For larger numbers, it is expected to have longer message processing time, and larger reduced message size as well.

C. *Test Environment*

The sender project is hosted on a Desktop PC (Intel Pentium 4, with 3.4 GHz Processor and 3 GBs of RAM, with Network Connection of 512 Kbps Download Speed and 128 Kbps Upload Speed).

Receiver (the web services) project is hosted on Microsoft Azure Cloud Small instances. Small instance is a virtual machine with a single core 2.10 GHz processor, 1.75 GBs of memory. Instances run Microsoft Windows Server 2008 R2 Enterprise Edition – 64 bit. Cloud instances use AutoScale feature for elasticity, with one to four instances. The services are hosted as Cloud Services, somewhere in West Europe.

TABLE I. Message Structure Analysis.

Message Type	XML Message Size (Bytes)	Total Nodes Count	Total Attributes Count	XML Complexity / Levels
Empty Alert	558	5	7	3
Complete Alert	5219	107	70	6
Tear Drop	1461	23	20	6
Ping Of Death	1387	25	22	6
Port Scanning 1	1623	30	26	6
Port Scanning 2	1304	22	19	6
Load Module 1	1076	19	17	6
Load Module 2	1581	35	22	6
phf	1450	27	19	6
File Modification	2352	51	31	7
System Policy Violation	1618	30	23	6
Correlated Alerts	1674	31	21	6
Analyzer Assessments	1772	37	20	6
Heartbeat	736	11	9	5

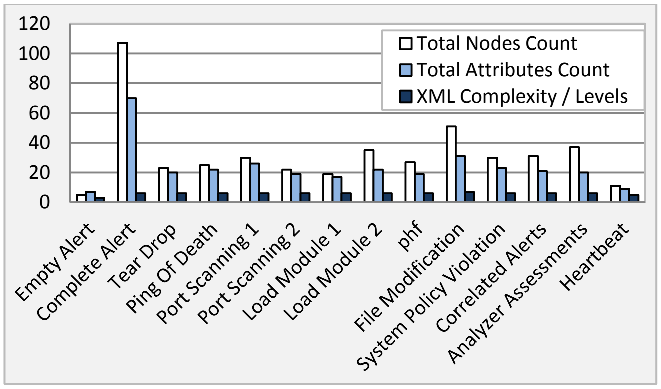


Fig. 8. Message Structure Analysis.

D. *Test Results*

For Experiment 1, the sender sends a burst of 500 messages. This results in total of 1000 messages for each type of the 14-message types. For experiment 2, a burst of 100 messages is sent for each message type, for each compression technique. Averages are recorded. Experiment 1 Test results for message size and transmission time (including serialization and deserialization time) (Table II) are recorded for normal XML message (Figure 1) against the proposed minified JSON message (Figure 3-b). Results show message size reduction ranging from 8.15% to 50.34%. Performance is enhanced by 35 to 342 milliseconds. The cloud instances' overall CPU usage did not exceed 5.55% of the CPU speed.

Experiment 2 results include Execution Time and Message Size analysis for compression techniques. Operations are abbreviated. Table III illustrates the abbreviations used and the function of each abbreviated process. Table IV shows Average Execution Time results. Table V shows Message Size results. Figure 9 shows combined average Execution Time for different techniques, including encoding and decoding times.

TABLE II. EXPERIMENT 1 RESULTS (SHOWING MESSAGE SIZE RESULTS IN BYTES, AND TRANSMISSION TIME RESULTS IN MILLISECONDS).

Message Type	Message Size Results Size (Bytes)			Transmission Time Results					
				Minimum (ms)		Maximum (ms)		Mean (ms)	
	XML	*Minified JSON*	*Reduction %*	*XML*	*Minified JSON*	*XML*	*Minified JSON*	*XML*	*Minified JSON*
Empty Alert	558	450	**19.35**	1027	1010	8803	2471	1455	1420
Complete Alert	5219	2592	**50.34**	1503	1169	3782	2732	1926	1585
Tear Drop	1461	1096	**24.98**	1120	1056	2234	2166	1523	1463
Ping Of Death	1387	1274	**8.15**	1119	1070	3759	3825	1525	1480
Port Scanning 1	1623	1061	**34.62**	1135	1046	5487	2203	1553	1460
Port Scanning 2	1304	957	**26.61**	1110	1044	4779	2157	1520	1454
Load Module 1	1076	894	**16.91**	1074	1046	2200	2238	1492	1449
Load Module 2	1581	1092	**30.92**	1142	1061	2276	2195	1545	1463
phf	1450	996	**31.31**	1125	1042	2260	2178	1525	1449
File Modification	2352	1450	**38.35**	1218	1082	2343	2187	1621	1489
System Policy Violation	1618	1066	**34.11**	1139	1053	2270	2310	1548	1460
Correlated Alerts	1674	1185	**29.21**	1142	1061	2282	2216	1550	1470
Analyzer Assessments	1772	1195	**32.56**	1160	1061	2259	2252	1558	1468
Heartbeat	736	404	**45.10**	1047	1005	2385	3647	1453	1418

TABLE III. EXPERIMENT 2 “COMPRESSION” ABBREVIATIONS.

Process Abbreviation	Description
XML	Serialization of traditional XML messages. No compression used.
De XML	Deserialization of traditional XML messages (inverse of the “XML” process).
XMill	Compression of XML messages using the specialized XMill compressor.
De XMill	Decompression of XML messages using the specialized XMill compressor (inverse of the “XMill” process).
GZip XML	Compression of XML messages using GZip compressor (a cyclic redundancy check value for detecting data corruption is included).
De GZip XML	Decompression of XML messages using GZip compressor (inverse of the “GZipXML” process).
Min JSON	(Proposed Technique) Serialization into Minified JSON messages. No compression used.
De Min JSON	(Proposed Technique) Deserialization from Minified JSON messages (inverse of the “MinJSON” process).
GZip Min JSON	(Proposed Technique) Serialization into Minified JSON messages, plus using GZip compression (a cyclic redundancy check value for detecting data corruption is included).
De GZip Min JSON	(Proposed Technique) Decompression of the compressed Minified JSON messages (inverse of the “GZipMinJSON” process).

TABLE IV. EXPERIMENT 2 AVERAGE EXECUTION TIME IN MILLISECONDS.

Message Type	De XML	XML	GZip XML	GZip Min JSON	De GZip XML	Min JSON	De Min JSON	De GZip Min JSON	De XMill	XMill
Empty Alert	0.01	0.04	0.15	0.01	0.04	0.08	0	1.12	10.6	17.3
Complete Alert	0.27	0.08	1.27	1.02	1.17	4.81	2.42	3.51	16.4	25.4
Tear Drop	0.09	0.07	0.21	0.07	0.24	0.59	0.61	1.02	11.9	18.9
Ping Of Death	0.05	0.13	0.17	0.13	0.19	0.16	1.09	1.25	11.7	18.8
Port Scanning 1	0.04	0.49	0.05	0.17	0.14	0.07	0.97	1.06	11.6	19.3
Port Scanning 2	0.17	0.02	0.13	0.18	0.5	0.02	0.37	1.56	11.8	18.7
Load Module 1	0.03	0.07	0.23	0.16	0.19	0.05	0.28	1.08	12	19.5
Load Module 2	0.23	0.01	0.28	0.2	0.22	0.02	1.3	1.17	11.3	19.1
phf	0.02	0.34	0.1	0.08	0.19	0.04	0.6	1.07	11.8	18.7
File Modification	0.07	0.05	0.17	0.39	0.96	0.07	1.15	1.97	11.8	20.2
System Policy Violation	0.01	0.16	0.16	0.37	0.28	0.21	1.19	1.1	12.6	19.1
Correlated Alerts	0.11	0.06	0.04	0.17	0.21	0.4	1.21	1.65	11.1	19.1
Analyzer Assessments	0.22	0.03	0.46	0.11	0.6	0	1.21	1.35	11.8	19.7
Heartbeat	0.18	0.25	0.11	0.64	0.01	0.01	0.03	0.11	11.2	17.9

TABLE V. EXPERIMENT 2 MESSAGE SIZE IN BYTES.

Message Type	GZipMinJSON	GZipXML	XMill	MinJSON	XML
Empty Alert	345	420	436	452	574
Complete Alert	888	1387	1659	2594	5488
Tear Drop	620	784	822	1098	1532
Ping Of Death	633	772	774	1276	1471
Port Scanning 1	633	832	834	1063	1719
Port Scanning 2	601	757	757	959	1379
Load Module 1	566	714	724	896	1136
Load Module 2	626	800	802	1094	1681
phf	601	787	794	998	1536
File Modification	677	953	980	1452	2501
System Policy Violation	614	835	855	1068	1713
Correlated Alerts	650	815	814	1187	1766
Analyzer Assessments	658	949	958	1209	1888
Heartbeat	415	559	556	406	772

XMill takes the longest execution time. However, other techniques take much shorter execution time between 0.6 and 1.6 milliseconds. Figure 10 shows detailed average Execution Time for different techniques. The prefix "De" signifies Decoding/Decompression Times, while the un-prefixed techniques signify Encoding/Compression Times.

Figure 11 shows Average Message Size for different techniques. Compared to XML, GZipped Minified JSON Messages are 66.1% shorter. GZipping the original XML files produces 54.8% shorter messages. The time-consuming specialized XMill compressor produces 53.22% shorter messages. Raw minified JSON messages are 37.37% shorter, without applying any compression. Figure 12 shows the detailed message size comparison for all experiment techniques.

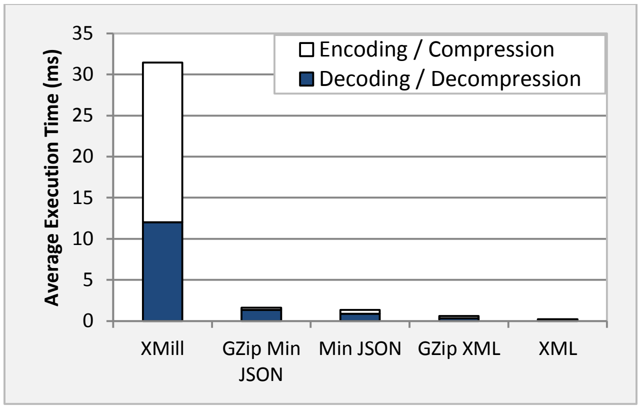


Fig. 9. Experiment 2: Results Comparison of Average Execution Time in milliseconds of Messages Encoding and Decoding using different techniques.

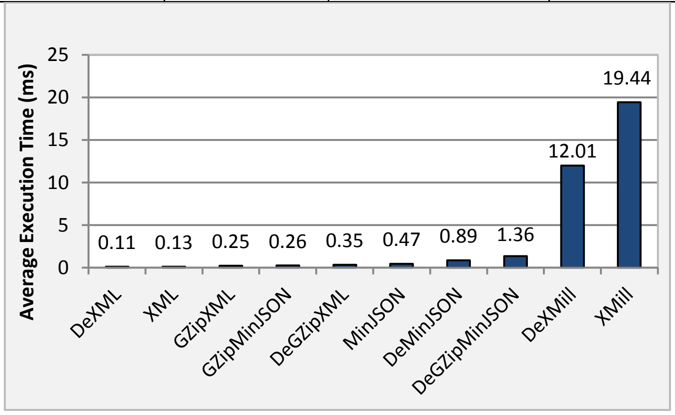


Fig. 10. Experiment 2: Results Comparison of Average Execution Time for Different Techniques in milliseconds.

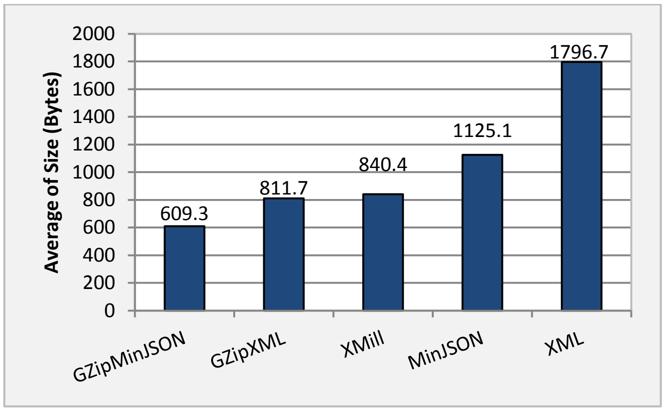


Fig. 11. Experiment 2: Results Comparison of Average Message Size for Different Techniques in Bytes.

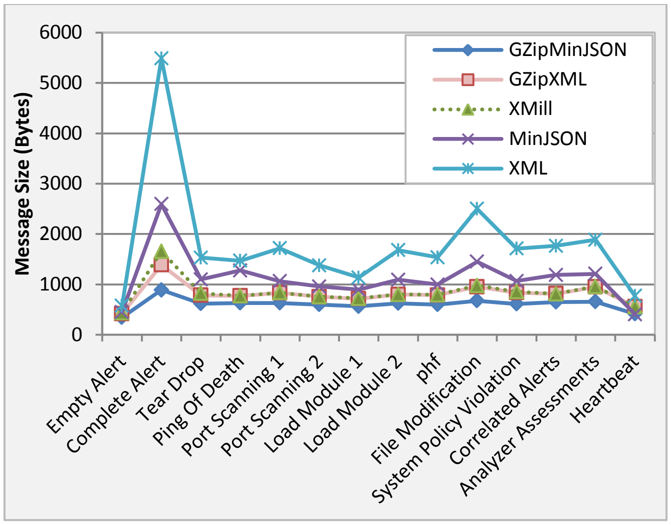


Fig. 12. Experiment 2: Results Comparison of Message Size in Bytes for different message types, for all techniques.

VI. Discussion and Conclusion

In this paper, we introduced a new XML Schema-based minification and communication technique in JSON message format. XML Schemas are minified using XSDMinify tool. This process is required only once per XML schema change. Then, the generated minified XML schema is processed with customized XSD code generation tool (XSD2Code). The code generation step generates code that sends and receives shorter minified messages. Based on the serialization type, communication can occur using shorter XML messages, or even shorter JSON messages. We performed our tests on Microsoft Azure Cloud platform, using different IDMEF messages types. Experiment 1 results show message size reduction ranging from 8.15% to 50.34% compared to raw XML messages. Performance is enhanced by 35 to 342 milliseconds. This technique is applied to raw messages, without applying any compression techniques (like those techniques introduced in section III). Compression techniques yield better results because of the similarities found in the new message structure (e.g. the minified names alphabets (a, b, c, …, and 1, 2, 3, …) instead of the full meaningful names). Experiment 2 applies both XML Compression Technique, and the general purpose GZip compression technique. As average results for all message types, XMill compression produces 53.2% shorter message, but XMill is very expensive in Execution Time (takes 31.45 extra milliseconds). Applying the proposed Minified JSON technique yields 37.37% shorter message compared to original XML messages. Minified JSON technique has extremely low execution time, reaching 1.36 milliseconds only. Adding GZip compression to Minified JSON technique produces 66.1% shorter message size compared to original XML messages (with 12.9% shorter size compared to XMill). GZipping Minified JSON technique takes 1.62 milliseconds only (94.85% faster than XMill).

To conclude, the proposed technique "Minified JSON messages" is a better alternative to using traditional XML messages, or specialized XMill compression. This technique produces a reasonable message size reduction, with almost no performance overhead. To achieve the best results, incorporation of GZip compression and Minified JSON technique is recommended. This produces the ultimate compression ratio, with a tiny negligible performance overhead. A separation between the names of the object oriented classes' members and the underlying transmission representation is well-established to maintain code readability. For future work, well-defined procedure for incorporating XML extensions will be studied. Data visualization tools may be considered for adopting the generated dictionary file (resulting from the minification process), in order to visualize minified data for user viewing.

References

[1] J. Simposon, *Just XML*, 2nd ed. US: Prentice Hall.

[2] H. Debar, D. Curry, and B. Feinstein. *The Intrusion Detection Message Exchange Format (IDMEF)*. [Online]. http://tools.ietf.org/html/rfc4765

[3] K. Ballinger, E. Christensen, and S. Pharies, "XML Serialization and Deserialization," US 6,898,604 B1, May 24, 2005.

[4] R. Krutz, and R. Vines, *Cloud Security: A Comprehensive Guide to Secure Cloud Computing*. US: Wiley Publishing, 2010.

[5] Cloud Security Alliance. *Cloud Security Alliance - Defined Categories of Service 2011*. [Online]. https://cloudsecurityalliance.org/download/defined-categories-of-service-2011/

[6] D. Miller, S. Harris, A. Harper, S. VanDyke, and C. Blask, *Security Information and Event Management (SIEM) Implementation*. US: McGraw-Hill, 2011.

[7] A. Madani, S. Rezayi, and H. Gharaee, "Log management comprehensive architecture in Security Operation Center (SOC)," in *International Conference on Computational Aspects of Social Networks (CASoN)*, Salamanca, 2011, pp. 284 - 289.

[8] R. Gerhards. *The Syslog Protocol*. [Online]. http://tools.ietf.org/html/rfc5424

[9] Mitre. *Common Event Expression*. [Online]. https://cee.mitre.org/about

[10] SNMP Research International. *SNMP RFCs*. [Online]. http://www.snmp.com/protocol/snmp_rfcs.shtml

[11] K. Hameseder, S. Fowler, and A. Peterson, "Performance analysis of ubiquitous web systems for SmartPhones," in *Performance Evaluation of Computer & Telecommunication Systems (SPECTS)*, 2011, pp. 84-89.

[12] G. Wang, "Improving Data Transmission in Web Applications via the Translation between XML and JSON," in *Communications and Mobile Computing (CMC), 2011 Third International Conference*, Qingdao, 2011, pp. 182-185.

[13] Z. Nagy, "Improved Speed on Intelligent Web Sites," *Recent Advances in Computer Science*, vol. 1, no. 14, pp. 215-220, 2013.

[14] H. Liefke, and D. Suciu, "XMill: an efficient compressor for XML data," in *The 2000 ACM SIGMOD international conference on Management of data*, vol. 29, 2000, pp. 153-164.

[15] D. Zhou, "Exploiting structure recurrence in XML processing," in *Web Engineering, 2008. ICWE'08. Eighth International Conference on*, 2008, pp. 311-324.

[16] S. Sakr, "XML compression techniques: A survey and comparison," *Journal of Computer and System Sciences*, vol. 75, no. 5, pp. 303–322, August 2009.

[17] C. Augeri, D. Bulutoglu, B. Mullins, R. Baldwin, and L. Baird, "An analysis of XML compression efficiency," in *The 2007 workshop on Experimental computer science*, 2007.

[18] T. Szalapski, S. Madria, and M., Linderman, "TinyPack XML: Real time XML compression for wireless sensor networks ," in *Wireless Communications and Networking Conference (WCNC)*, Shanghai , 2012, pp. 3165 - 3170.

[19] L. Deutsch. (1996, May). *DEFLATE Compressed Data Format Specification*. [Online]. http://www.ietf.org/rfc/rfc1951.txt

[20] L. Deutsch. (1996, May) *GZIP file format specification*. [Online]. http://tools.ietf.org/html/rfc1952

[21] P. Katz, "DEFLATE: String searcher, and compressor using same," US5051745 A, September 24, 1991.

Designing and Building a Framework for DNA Sequence Alignment Using Grid Computing

EL-Sayed Orabi
IS Dept. MTI University
Cairo, Egypt.

Mohamed A. Assal
Dean of IT Center. MTI Univ.
Cairo, Egypt.

Mustafa Abdel Azim
Dean of CS Faculty. AASTMT
Cairo, Egypt.

Yasser Kamal
CS Dept. AASTMT
Cairo, Egypt.

***Abstract*—Deoxyribonucleic acid (DNA) is a molecule that encodes unique genetic instructions used in the development and functioning of all known living organisms and many viruses. This Genetic information is encoded as a sequence of nucleotides (adenine, cytosine, guanine, and thymine) recorded using the letters A, C, G, and T.DNA querying or alignment of these sequences required dynamic programming tools and very complex matrices and some heuristic methods like FASTA and BLAST that use massive force of processing and highly time consuming. We present a parallel solution to reduce the processing time. Smith waterman algorithm, Needleman-Wunsch, some weighting matrices and a grid of computers are used to find field of similarity between these sequences in large DNA datasets. This grid consists of master computer and unlimited number of agents. The master computer is the user interface for insert the queried sequence and coordinates the processing between the grid agents.**

Keywords—DNA fingerprint; Smith waterman algorithm; Needleman-Wunsch; Grid computing; Coordinator and Agent computers

I. Introduction

DNA sequences are a string of characters (A, C, G and T) representing the genetic information of a living organisms (humans, animals, birds, bacteria, planets, etc.) and many known viruses. Every living culture has its own unique nucleotide code. Based on this fact the government agency all over the world use the DNA sequence to identify persons (criminals, army and police solider, terrorism, etc.) and can also be used to determine a child's paternity (genetic father) or a person's ancestry. The way of differentiates these sequences is called sequence alignment. Sequence alignment is also used to identify the breed (homologies) of unknown protein or nucleotide sequences. This can be solved by using dynamic programming in time proportional to the product of the length of the two sequences being compared, as in [1].

Sequence alignment is a tool used to compare the sequence to find a similarity between them based on complex algorithms and matrixes as in [2]. The Smith Waterman and Needleman Wunsch are used for local and global sequence alignment.

To solve the delay time of the comparison scientists all over the world proposed different models and techniques including hardware improving (sequencer machines)to reduce the length of the sequence using microarray as in [3].

II. DNA Fingerprint

DNA fingerprinting is a test to identify and evaluate the genetic information in a person's cells. It is called a "fingerprint" because it is very unlikely that any two people would have exactly the same DNA information, in the same way that it is very unlikely that any two people would have exactly the same physical fingerprint.

The test is used to determine whether a family relationship exists between two people, to identify organisms causing a disease, and to solve crimes. [4] With different DNA datasets contains millions of records, it may be impossible to find the target sequence (person) at the right time. So the need to get information fast from the large databases was raised rabidly. The parallel computing is a solution to reduce the time of querying and retrieving information from these databases by distributes the processing over numbers of devices, as in [2].

III. Smith Waterman Algorithm

It performs a local alignment over two sequences. It is an example of dynamic programming. This algorithm is useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context or sequences with the same length. Initialization, Scoring and Trace back (Alignment) are three steps to find the best alignment over the conserved domain of two sequences.

The complexity of this algorithm is $O(N*M)$ where N is the length of queried sequence and M is the length of target sequence. Smith waterman is a pair wise sequence alignment on other word it is 1 to 1 alignment. Example of local alignment 2 sequence N and M with match= 4, mismatch = -1 and gap = -2 is illustrated in the following example. The First step is the initialization of matrix N*M as illustrates in table 1.

TABLE I. FILLING THE MATRIX WITH THE GIVEN SCORES

Initiates the matrix with gap		Sequence M						
			G	A	T	T	G	A
		0	0	0	0	0	0	0
Sequence N	A	0	0	4	2	0	0	4
	C	0	0	2	3	1	0	2
	G	0	4	2	1	2	5	3
	C	0	2	3	1	0	3	4

The second step is the trace back (local alignment) as shown in table 2. The trace back starts with the maximum value in the matrix then goes left or up or diagonal according to values next to the start point.

TABLE II. THE TRACING BACK (LOCAL ALIGNMENT)

Trace back from the Max. Value		Sequence M						
			G	A	T	T	G	A
		0	0	0	0	0	0	0
Sequence N	A	0	0	4	2	0	0	4
	C	0	0	2	3	1	0	2
	G	0	4	2	1	2	5	3
	C	0	2	3	1	0	3	4

The last step is the alignment of the two sequences as illustrates in figure 1.

Seq. M: A T T G

Seq. N: A C - G

Fig. 1. The Final Local alignment of the given sequences using smith waterman algorithm

IV. NEEDLEMAN WUNSCH

The Needleman–Wunsch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequences. It was published in 1970 by Saul B. Needleman and Christian D. Wunsch. It uses dynamic programming, and was the first application of dynamic programming to biological sequence comparison. It is sometimes referred to as the optimal matching algorithm. Initialization, Scoring and Trace back (Alignment) are three steps to find the best alignment over the entire length two sequences. The following example shows a simple alignment between two sequences N and M with match score = 4, mismatch = -1 and gap = -2. The First step is the initialization of matrix N*M as illustrates in table 3.

TABLE III. THE SCORING MATRIX OF NEEDLEMAN WUNSCH ALGORITHM

Initiates the matrix with gap		Sequence M						
			G	A	T	T	G	A
		0	-1	-2	-3	-4	-5	-6
Sequence N	A	-1	-1	3	-3	-4	-5	-1
	C	-2	-2	2	2	1	0	-1
	G	-3	2	1	1	1	5	4
	C	-4	1	1	0	0	4	3

The second step is the trace back (Global alignment) as shown in table 4. The trace back starts with the last right point in the matrix then goes left or up or diagonal according to values next to the start point.

TABLE IV. THE TRACING BACK (GLOBAL ALIGNMENT)

Trace back from the last point		Sequence M						
			G	A	T	T	G	A
		0	-1	-2	-3	-4	-5	-6
Sequence N	A	-1	-1	3	-3	-4	-5	-1
	C	-2	-2	2	2	1	0	-1
	G	-3	2	1	1	1	5	4
	C	-4	1	1	0	0	4	3

The last step is the alignment of the two sequences as illustrates in figure 2.

Seq. M: G A T T G A

Seq. N: - A - C G C

Fig. 2. The Final Global alignment of the given sequences using Needleman Wunsch algorithm

V. GRID COMPUTING

Grid computing is basically a paradigm that aims to enable access to high performance distributed resources in a simple and standard way. A grid is defined as a type of parallel and distributed autonomous resources dynamically at runtime depending on their availability, capability and performance as in [5]. The aim of grid computing is to enable coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organization as in [6][7][8].Grid computing is known as a distributed system that connects many computer systems having different hardware and platforms (operating system). It allows applications to run in parallel on multiple machines, clusters, or systems (Virtual Organization). The system is suitable for solving the problems that require a large amount of computation as well as storage capacity. In our research, the Grid system is used for solving the delay time of the global & local alignment. Figure 3 illustrated the proposed grid components.

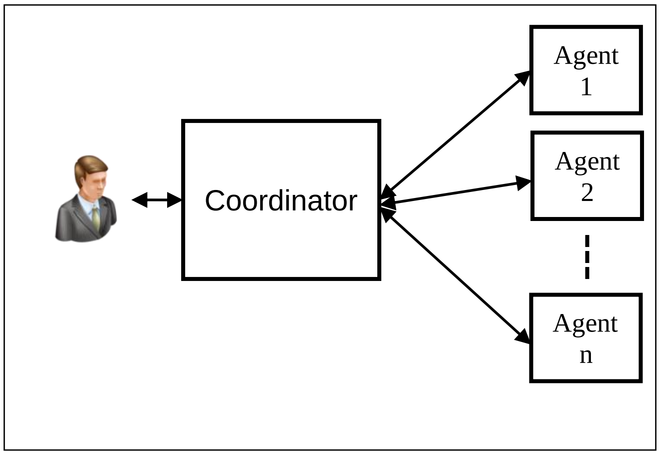


Fig. 3. The grid components

a) The coordinator

The coordinator sends the tasks to agents so agents are fully utilized as possible. The coordinator considers each physical core of each agent as a separate execution unit as in [8]. The user inters the queried sequence and XML file contains the dataset that will be searched through the coordinator application. Then the coordinator calculates the total number of sequences in the dataset counts the connected agents and divides the task equally for each agent. Each task consists of two sequences the first is the queried sequence and the second is a sequence from the dataset file. Example, dataset contains 1024 sequences and the grid consists of 8 agents so each agent receives 128 tasks. Figure 4 shows the components of the coordinator computer.

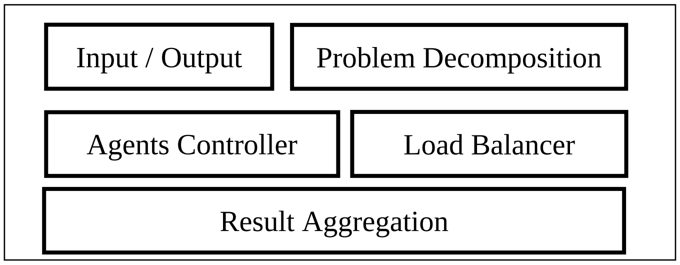


Fig. 4. Coordinator components

b) Agent

The agent registers itself with the coordinator and waits to receive grid tasks from coordinator as in [8]. The agent receives the sequence alignment tasks from the coordinator and executes them using Smith waterman algorithm. An agent is configured to be dedicated which mean that agent resources are centrally managed by the coordinator. Then each agent sends results of alignment back to the coordinator which selects the most similar sequence to the queried sequence. Figure 5 shows the main function of the agent computer.

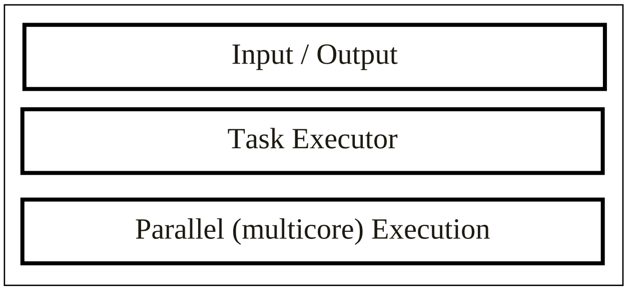


Fig. 5. The Agent functions

V. RELATED WORK

In many published paper the researchers all over the world proposed a lot of models to make local alignment using parallel computing. The following section demonstrates some examples of those models.

- Propose parallel processing of optimal alignment between two sequences by exploiting parallel MPI/FORTRAN 90. The algorithm for optimal alignment is based on dynamic programming techniques. Two versions of algorithms have been developed: one versus one sequence alignment and one versus many sequence alignment. The second algorithm used "block" parallel dynamic programming algorithm and this technique will increase the amount of workloads done by each processor as in [9].

- DNA sequence alignment model under this hierarchical grid architecture. They used dynamic programming algorithm with linear space parallelism and is separated into two parts: parallelization of the similarity matrix and parallelization of the divide-and-conquer algorithm. Three clusters have been setup where each cluster has eight nodes. The clusters are connected by Ethernet switch where the bandwidth is about 8 MByte/s. Meanwhile the bandwidth between nodes in each cluster is about 190 Mbyte/s. The architecture of the software is based on two layers; upper layer uses MPICH-G2 and lower layer employs MPICH as a communication interface protocol as in [10].

- FASTA is a heuristic based technique in sequence similarity search. Parallelization of FASTA has been implemented in the Grid Application Development Software (GrADS) project as in [11]. The GrADS adapts the master-worker paradigm, scheduling and rescheduling the tasks on an appropriate set of resources, launching and monitoring the execution. The GrADSoft scheduler makes a static schedule for its application where the whole or a portion of sequence databases are replicated on some or all of the grid nodes. The master will inform each worker which portions of database should be loaded into memory. The master also sends the input query sequence to each worker and collects the results from the workers.

VI. THE IMPLEMENTED FRAMEWORK

By using the previous grid topology and Smith Waterman, Needleman-Wunsch basic implementation the researcher proposes a DNA multiple sequences alignment framework. This framework is been based on pairwise comparison to query large databases by distribute one to one tasks to find the maximum complete or partial alignment over the grid computing. Those tasks are generated from user interface (coordinator). The coordinator counts the number of sequences of the database. Then calculates the number of connected agents and counts the number of available cores. Then the coordinator divides the tasks equally with load balance through the agent to execute the Smith waterman and Needleman-Wunsch algorithms. Then the agents start to execute the tasks one by one and send the results to the coordinator. If any failure is found in any agent, the coordinator reassigned the task for another available agent considering the load balance of each agent in the grid. At last the coordinator selects the sequence with maximum matching score. The second scenario is distributing the dataset as clustered sub datasets on each agent. Then the coordinator defines the cluster of the queried sequence with codon cluster technique. To increase the accuracy of the alignment to cover the analysis requirement, the researcher combined the original implementation of Smith waterman algorithm with some weighting matrixes. The first will be blosum 62, the second is PAM 250 and the third is Gonnet160.Those weighting matrices are used in many DNA analysis applications. So the researcher thought it will be very helpful feature in the proposed model.

VII. Experiments and Result

The experiments were carried out in a computer laboratory contains 16 connected personal computer. One PC used as coordinator and the rest used as agents. The configuration of each PC is shown in table 5.

TABLE V. DEMONSTRATES THE GRID NODES CONFIGURATIONS

#	The nodes configuration	
1	OS	Microsoft Windows 7 Enterprise Service Pack 1 (64 bits)
2	Processor	Intel® Core™ i7-2600 CPU @ 3.40Ghz
3	Number of Cores	4 P / 4 V
4	Clock Speed	3.701 Ghz
5	Memory	4 GB

The first iteration of the experiment was for test the performance of the grid through sending some tasks (sequences to be aligned) with different lengths. These tasks and the time to carry out them over number of nodes are illustrated in table 6 and figure 6.

TABLE VI. ILLUSTRATES FIRST ITERATION WITH DIFFERENT SEQUENCE LENGTH

Seq. Length	No. Of Seq.	No. Of Cores (time per minute)				
		1	4	8	16	32
1000	128	0.3178	0.2347	0.1237	0.06727	0.0343
2000	128	1.2622	0.2409	0.1804	0.10385	0.0579
4000	128	5.0105	0.6445	0.6333	0.32808	0.1801
8000	128	9.6267	0.2641	0.2232	0.11123	0.0946

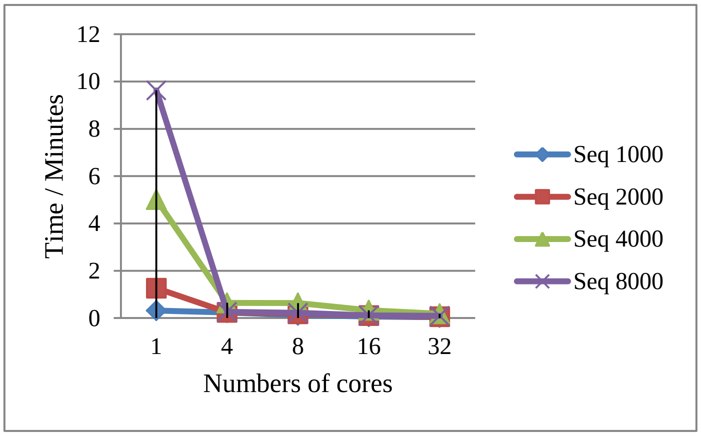


Fig. 6. Demonstrates the result of first iteration

After the first iteration it was found that the processing power of each agent using multi core processing is fully utilized when using sequence length larger than 2000 nucleotides. It was found that the sequence of length 4000 nucleotides takes more than double time of the sequence 2000 nucleotides length, so the second iteration is querying sequences with length 4000 nucleotides against datasets contain different number of sequences. The results of this iteration are demonstrated in table number 7 and figure number 7.

TABLE VII. DEMONSTRATES THE RESULT OF SECOND ITERATION (SAME SEQUENCE LENGTH AGAINST DIFFERENT DATASETS)

Seq. lengt h	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	5.01052	0.6445	0.6333	0.3281	0.1801
4000	256	4.85454	1.1837	0.5745	0.3216	0.1718
4000	512	19.8005	3.2172	1.0123	0.5622	0.3454
4000	1024	39.6572	9.7471	2.0293	1.2582	0.6607
4000	2048	79.3692	19.494	9.7472	4.4924	2.4490

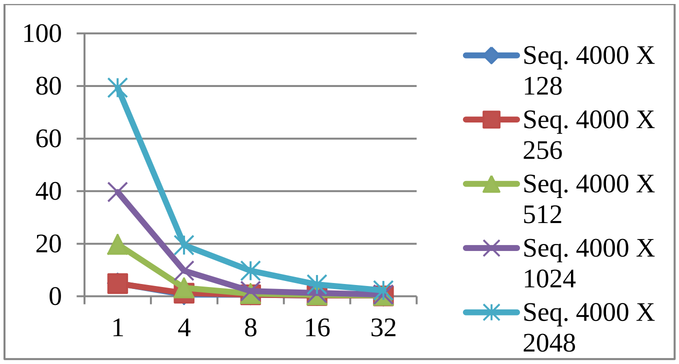


Fig. 7. Shows the relation between the sequences with the same length and different numbers of sequence to be aligned with

For more alignment details the researcher combined the original Smith Waterman & Needleman Wunsch algorithms with BLOSUM, PAM and Gonnet matrices. To calculate the effect of adding the blosum62 and PAM250 weight matrices to the original implementation the researcher run three more iterations. The results of those iterations are illustrated in tables and figures number 8, 9 and 10.

TABLE VIII. DEMONSTRATES THE RESULT OF BLOSUM62 WEIGHTING MATRIX

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	7.3378	1.8514	0.9181	0.3148	0.24600
4000	256	14.676	3.6027	1.8361	0.6296	0.49200
4000	512	29.351	7.6054	3.6593	1.8726	0.94385
4000	1024	58.702	14.677	7.3186	3.5989	1.86914
4000	2048	117.97	29.255	14.627	7.0823	2.62731

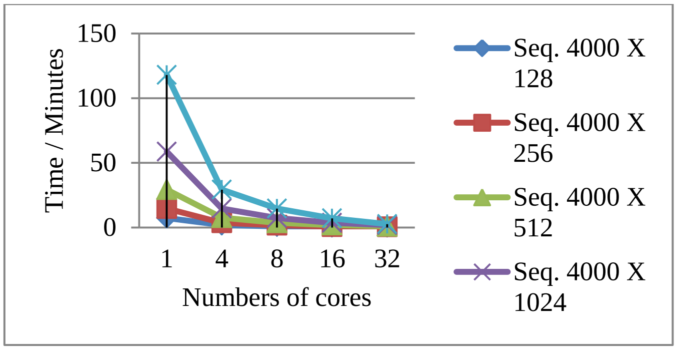


Fig. 8. Demonstrates the result of Blosum62 matrix

TABLE IX. THE RESULT OF PAM250 MATRIX ITERATION

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	7.51120	1.8779	0.9459	0.4748	0.2473
4000	256	15.0224	3.7557	1.8918	0.9497	0.4947
4000	512	30.0448	7.5112	3.6923	1.8517	0.9657
4000	1024	60.0896	15.022	7.5142	3.7202	1.8580
4000	2048	120.179	30.145	15.123	7.5231	3.7302

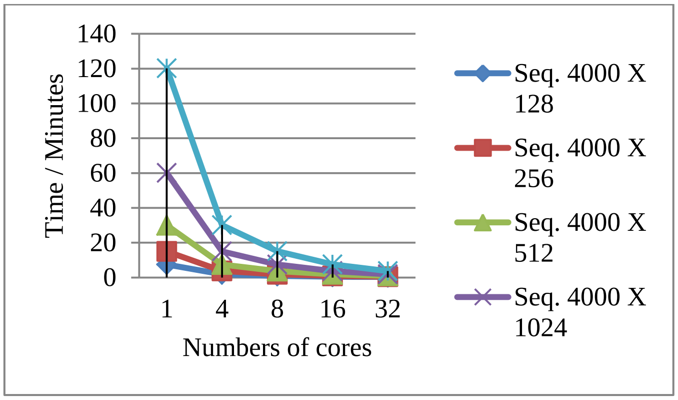


Fig. 9. Shows the result of PAM250 iteration

TABLE X. DEMONSTRATES THE RESULT OF GONNET160 WEIGHTING MATRIX

Seq. length	No. of Seq.	Number of cores (time per minutes)				
		1	4	8	16	32
4000	128	9.2305	3.1070	1.5535	0.7581	0.4259
4000	256	18.461	6.2142	3.1162	1.5161	0.8517
4000	512	36.922	12.428	6.2142	3.0321	1.7034
4000	1024	73.844	24.857	12.428	6.0642	3.4069
4000	2048	147.69	49.713	24.857	12.128	6.8138

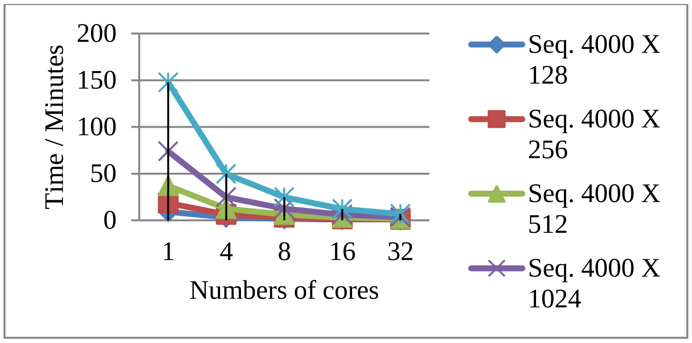


Fig. 10. Shows the result of Gonnet160 iteration

From previous experiments it is found that the time of processing decreased when the number of nodes (cores) increased. These tests make the complexity of Smith waterman and Needleman Wunsch algorithms for finding target sequence in multi sequence as O(N*M)S/C where S is number of sequences and C is number of cores connected to the grid.

VIII. CONCLUSION

The implementation of Smith waterman and Needleman Wunsch algorithms over a grid of computers decreases the time of processing. Grid computing makes querying large DNA datasets fast enough and with affordable cost. The use of grid and smith waterman algorithm to find a match between a sample and multi sequences is a negative relation as more cores in the grid the less computational time and vice versa. The complexity of the proposed model for Multi sequence alignment is O(N*M)S/C where N and M is the pairwise sequences, S is the number of sequences in the dataset file and C is number of cores in the grid agents.

Fig. 11. References

[1] “Algorithms in Bioinformatics”,.I,WS06/07, C.Dieterich.

[2] “Multiple Sequence Alignment on the Grid Computing using Cache Technique, ” International Journal of Computer Science and Telecommunications Volume 3, Issue 7, July 2012.

[3] “A Novel Algorithm for Fast Synthesis of DNA Probes on Microarrays” ACM Journal on Emerging Technologies in Computing Systems, Vol. 9, No. 1, February 2013.

[4] “DNA Testing in Criminal Justice: Background, Current Law, Grants, and Issues,“Congressional Research Service 7-5700.

[5] “Evolution of Cloud Computing and Enable Technologies,” International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol.1, no.4, pp.182-198, October 2012.

[6] PerfCloud: Grid Services For Performance-Oriented Development of Cloud Computing Application,” 18th IEEE International Workshop on Enabling Technologies: Infrastructures for Collaborative Enterprise (WETICE 09), pp. 201-206,2009.

[7] Ahmed Said Abo El-Ala, Mohamed Anwar Assal, and Mohamed Bakr,” On the Design of a framework for Grid computing Developing System(GDS) “, In Managerial Research Journal, Consultancy Research & Development Centre, Sadat Academy for Management Sciences, July 2012.

[8] Ahmed Said Abo El-Ala, Mohamed Anwar Assal, and Mohamed Bakr," An Enhanced framework for Grid computing Developing System(EGDS) ", In Managerial Research Journal, Consultancy Research & Development Centre, Sadat Academy for Management Sciences, July 2012.

[9] Nguyen, E. N. D., Nguyen, D. N., Nguyen, D. T. and Tungkahotara,"Comparing DNA Sequences by Dynamic Programming in Sequential and Parallel Computer Environments", Proc. of the 2006 WSEAS International Conference on Mathematical Biology and Ecology, 2006, 146 – 153.

[10] Chen, C. and Schmidt, B. "An Adaptive Grid Implementation of DNA Sequence Alignment", Future Generation Computer Systems,2005, 988 – 1003.

[11] YarKhan, A. and Dongarra, J. J. "Biological Sequence Alignment on the Computational Grid Using the GrADS Framework", Future Generation Computer Systems,2005, 980 – 986.

A High Performance Biometric System Based on Image Morphological Analysis

Marco Augusto Rocchietti, Alejandro Luis Angel Scerbo, Silvia M. Ojeda
Facultad de Matemática Astronomía y Física
Universidad Nacional de Córdoba
Córdoba, Argentina

***Abstract*—At present, many of the algorithms used and proposed for digital imaging biometric systems are based on mathematical complex models, and this fact is directly related to the performance of any computer implementation of these algorithms. On the other hand, as they are conceived for general purpose digital imaging, these algorithms do not take advantage of any common morphological features from its given domains.**

In this paper we developed a novel algorithm for the segmentation of the pupil and iris in human eye images, whose improvement's hope lies in the use of morphological features of the images of the human eye. Based on the basic structure of a standard biometric system we developed and implemented an innovation for each phase of the system, avoiding the use of mathematical complex models and exploiting some common features in any digital image of the human eye from the dataset that we used. Finally, we compared the testing results against other known state of the art works developed over the same dataset.

***Keywords*—Biometric System; Digital Image Processing; Pupil and Iris Segmentation; Iris Matching**

I. Introduction

The purpose of Biometrics is the research of features that enable the univocal identification for each human being [1].

Since some time ago, the human iris is considered as a biometric because it has some special features against other biometrics [2] [3]. For example, the iris biometrical features are present in the human from the 3rd month of gestation, and it remains almost identical until individual's death. On other hand, any physical contact is not necessary to take an iris sample and the sample's forgery is practically remote (or at least too troublesome). In the iris visual pattern there is more biometrical information (for univocal human identification) than in a fingerprint [2] [3] [4]. Besides, the human iris diameter is very regular, varying between 11.5mm and 12mm from one individual to another, although, by the lens effect caused by the cornea we could measure 13mm of horizontal length [5]. This fact is really important, because it gives us an anatomical max value of 6.5mm for the iris radius in any eye image. Fig. 1 shows iris, pupil and sclera of the human eye.

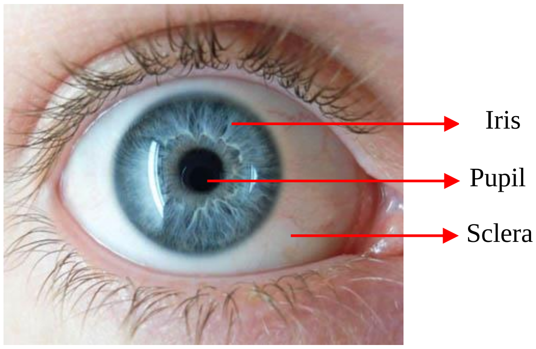


Fig. 1. Human eye's basic sections.

A complete standard biometric system based on iris recognition (BSI) consists of four phases [6]:

A. Segmentation.

B. Normalization.

C. Extraction.

D. Matching.

Almost every latest technique used for iris recognition and human identification shares a common origin, related from the beginning to the statistical analysis of digital imaging [7]. Many of the best algorithms for digital images treatment, filtering, and compression owe their success to this statistical approach, because, among other reasons, by modeling images as mathematical objects, theoretical developments, such as Fourier, Wavelet, Hugh transformations and Gabor filters could then be applied, obtaining in many cases excellent results [8] [9] [10] [11] [12] [13].

Thus, most of the latest Iris biometric systems currently use these complex mathematical tools to accurately obtain the edges of the iris (Segmentation phase), and employ generic entropies (like Hamming Distance [14]) later, during the Matching phase. Aversely, the application of these statistical model-based algorithms in biometric systems may depend on filtering, pre-treatments, matrix calculus and other operations that could be costly in terms of computer implementations.

In this paper we developed a novel algorithm for the segmentation of the pupil and iris in human eye images, whose improvement's hope lies in the use of morphological features of the images of the human eye. Considering an image as a data structure and not as a mathematical object only, we proposed an original alternative for each one of the four phases of a biometric system based on the iris recognition. Our proposal introduces improvements in the performance of the system by drastically decreasing the complexity of the segmentation, in order to get lower computational cost compared to other similar algorithms. We implemented the system in an appropriate imaging framework in order to compare our results with other actual developments.

The rest of the paper is organized as follows: Section II, presents a brief description of the Database, equipment and basic definitions and notations used in this work. In Section III we explain our iris segmentation algorithm in detail. Section IV and V are about the improvement of the Normalization and Extraction phases respectively. Section VI shows the results of our study. Conclusions and future scopes will appear in sections VI and VII respectively.

II. Materials, Definitions and Notations

We used CASIA iris image database version 1.0 [15]; because it is used today in most of the developing works in the area of Iris Recognition [8] [16] [17] [18]. This database includes 7 different samples per individual and includes data of 108 individuals; 4 for testing stage and 3 for training stage. The images of the Database have some common features, which became morphological invariants for our development:

1) In every sample (image) of the dataset, there is exactly one eye.

2) The pupil in the sample will look like a regular dark discoid.

3) In each sample, the pupil represents the biggest dark region of the image.

4) All the samples of the dataset are taken maintaining the same distance between sensor (camera) and target (individual's eye); therefore, all the images share approximately the same spatial resolution.

5) We know the spatial resolution parameter of the dataset; then, we can estimate in pixels the max value for the iris diameter (remember the 13mm max value, section I).

In this work we used the following equipment:

- Hardware: Sony(R) Vaio(TM) notebook VPCEB13-EL model. (CPU Intel(R) Core(TM) i3 M330 @2.13GHz, 2GB RAM).
- Software: RSI-IDL(R) y RSI-ENVI(R) 4.7 suite [19].

Finally, in this section, we present some definitions and notations used in this work.

- We represent an ocular image of n x m pixels as a brightness value matrix I of n x m dimension (n columns and m rows). To refer to the element in column i, and row j of matrix I, we use the standard notation: $I[i,j]$, where i and j are natural numbers, with $\leq i \leq n-1$, and $0 \leq j \leq m-1$. Lower values in the matrix will be related to darker pixels in the image.
- We use "cell" or "pixel" to denote an element of matrix I, which is associated to a brightness value and its coordinate pair in I.
- The pupil segmentation of the ocular image I is the smallest circle that contains the pupil in I, which is represented with the pair (c,r), where c is the coordinate pair of its center and r its radius measured by adjacent pixels.
- The iris segmentation of the ocular image I is the geometric circular crown $S = (c,r,R)$ such that: the pair (c,r) is the pupil segmentation of I, and R is the distance (in straight line pixels) from c to a pixel from the limit iris-sclera on I.

III. Iris Segmentation. a Novel Proposal

To delimit the iris in an ocular image, we need to find the appropriate c, r, and R parameters. First, let us note that R parameter is a constant for every sample (ocular image) in a dataset since we choose the anatomic max value of 13mm for the iris diameter as an outside diameter for all our iris segmentations. Then, we calculated R using the spatial resolution parameter of the dataset (see v in Section I). Thus, R was defined indeed as the necessary amount of adjacent pixels to cover 6.5mm in the image.

To determine c and r parameters, which are not constants at all, note that c will vary from a sample to another according to the eye's position in the image, and r will be determined by the pupil dilatation in each sample. As we defined in the previous section, c and r parameters are obtained by calculating the pupil segmentation of the image. So, at this point, we reduced the problem of iris segmentation to obtain the pupil segmentation.

A. *Locating the pupil. "CRUZ" algorithm*

Let $\mathbf{i} = (x_i, y_i)$ the coordinate pair of a pixel $\mathbf{p_i}$ located inside the pupil of image I. Let us trace four paths from $\mathbf{i}$. Two will draw up the vertical trace (north and south paths), and the other two will compose the horizontal trace (east and west paths). Each path will end when the difference between the brightness value of the next pixel and value of the current pixel $\mathbf{p_i}$ is greater than τ (tolerance).

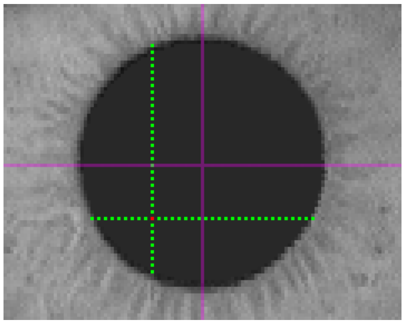


Fig. 2. In red, the coordinate pair **i**. In green, the traces made by CRUZ algorithm. In violet, the horizontal and vertical traces.

Our idea was to approximate the pupil by mean of a perfect circle. To estimate the center *c* of the circle, the horizontal and vertical traces, and some geometrical principles were used:

Consider the perpendicular line that passes through the middle point of horizontal trace (vertical violet line in Fig.2) and the perpendicular line that passes through the middle point of vertical trace. As you can see in Fig.2, the intersection of both lines approximates the pupil center. Now formally: let T_N, T_S, T_E, T_W be the lengths (in pixels) of the north, south, east and west paths respectively obtained by CRUZ algorithm initialized with $\mathbf{i} = (x_i, y_i)$.

Let us define the center $c = (x_c, y_c)$ as follows:

$$x_c = \frac{(x_i + T_E) + (x_i - T_W)}{2} \quad (1)$$

$$y_c = \frac{(y_i + T_S) + (y_i - T_N)}{2} \quad (2)$$

Let I be the matrix of a given ocular image, and $\mathbf{i} = (x_i, y_i)$ be the coordinate pair of some pixel located inside the pupil of that image. A pseudo-code for CRUZ algorithm from i could be:

```
$T_N, T_S, T_E, T_W \leftarrow 0$
while $|I[x_i, y_i] - I[x_i, y_i - T_N]| \le \tau$ and $T_N < y_i$
do $T_N \leftarrow T_N + 1$
while $|I[x_i, y_i] - I[x_i, y_i + T_S]| \le \tau$ and $y_i + T_S < M-1$
do $T_S \leftarrow T_S + 1$
while $|I[x_i, y_i] - I[x_i + T_E, y_i]| \le \tau$ and $x_i + T_E < N-1$
do $T_E \leftarrow T_E + 1$
while $|I[x_i, y_i] - I[x_i - T_W, y_i]| \le \tau$ and $T_W < x_i$
do $T_W \leftarrow T_W + 1$
Calculate $x_c$ given by (1)
Calculate $y_c$ given by (2)
$c \leftarrow (x_c, y_c)$
return $c, T_N, T_S, T_E, T_W$.
```

The CRUZ algorithm outputs are the center $\mathbf{c} = (x_c, y_c)$ and the lengths (in pixels) of the four paths. Thus, after CRUZ algorithm running, we have a possible center for the circle; but we still need to determine the radius *r* to complete the pupil segmentation. At first we could approximate *r* as half trace (whatever vertical or horizontal trace), but there is a detail we have to consider: The real pupil in the sample is not a perfect disc, since it has irregular edges, sometimes depending on light conditions, other times varying from an individual to another. This fact determines at least the next two issues:

1) After CRUZ algorithm running, the lengths of the traces could be different, so we would have to establish criteria to obtain r from the traces.

2) The farther from the real center of the pupil the initial coordinate pair **i** *is, the worse the estimation by c of the real center will be.*

If the pupil were a perfect disc, both traces would measure the same; and a half-length trace would measure the radius *r* of our interest. In this case, CRUZ algorithm would obtain the exact center of the disk starting from any coordinate pair of the pupil, no matter how far from the center it is.

The second issue suggests that we will obtain a better center *c* (closer to the real center) if the starting position **i** is already near the pupil center. Therefore, let **'c*'** be the center calculated by CRUZ algorithm running. Therefore, to fix the second issue we will run CRUZ algorithm again, just from **'c*'** as initial position. In other words, the first execution will get us closer to the real center, and the second will give us a very good estimation of it.

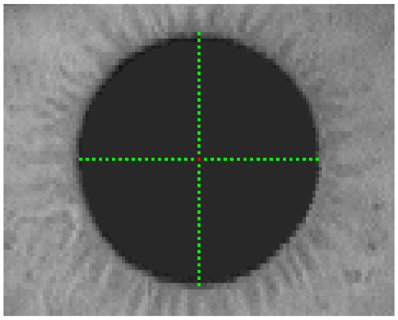


Fig. 3. Graphical CRUZ algorithm execution from the estimated center.

Finally, to attack the first issue, we decided to choose as radius *r* the minimum from the four paths given by a third CRUZ execution from the last center obtained (see Fig.3). This will guarantee that our pupil segmentation does not include iris pixels. Summary: To obtain our pupil segmentation *(c,r)*, we will use CRUZ algorithm (double run) to obtain *c*, and again to obtain *r*, always assuming that the first run starts from a coordinate pair inside the pupil (the second and third run start from the center calculated by the earlier run).

B. Improving CRUZ algorithm

Since the pupil represents the biggest dark area in any sample (see iii, Section I), if we choose a set of regular spaced positions from the image, most of the darker pixels will be from the pupil. Other dark pixels could come from eyelashes or some kind of noise. Therefore, we used that fact as follows: If CRUZ algorithm is executed from a dark pixel that is not in the pupil (eyelashes, noise, etc.), the horizontal and vertical traces will be too tiny or too different from each other.

Therefore, we redefined the criterion to determine if a trace is adequate with **ρ** parameter, which specifies a minimum value for pupil radius. On the other hand, to determine if both traces are alike enough, we fixed a criterion of "expected circularity" introducing **σ** parameter as a percent of desired similarity. We applied these rejection criteria over an ordered list *C* of coordinate pairs (candidates to be from pupil). If a candidate is rejected, we pass to the next in *C* list. If not, we assumed that we have found a coordinate pair **i** as we needed to apply CRUZ algorithm. To minimize the number of comparisons, we could define *C* as a selection of regular spaced pixels (fig. 4) and then order them by bright level, starting from the darkest one.

We introduced the following definitions:

Let $C = \{c_i\}$ be the ordered finite list of pixels defined above such that: c_o is the darkest one and c_i is darker than c_{i+1}. We will use the symbol "CRUZ[x]" to denote the execution of CRUZ algorithm initialized with the sample's pixel 'x'.

$$T_{\min} = \min\{T_N, T_S, T_E, T_W\} \tag{3}$$

$$T_{Max} = \max\{T_N, T_S, T_E, T_W\} \tag{4}$$

$$\overline{T} = \frac{T_N + T_S + T_E + T_W}{4} \tag{5}$$

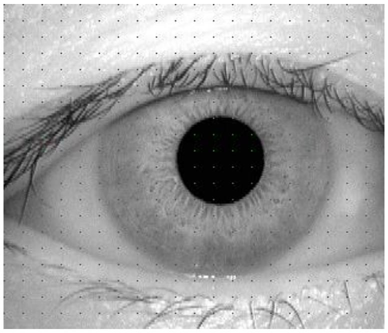


Fig. 4. Regular spaced pixels on a sample.

Finally, the pupil segmentation algorithm is defined by the next nine steps:

```
1. $i \leftarrow 0$ ; $T_N, T_S, T_E, T_W \leftarrow 0$ ; $c \leftarrow (0,0)$
2. CRUZ[$c_i$] (1st CRUZ running: each execution
   updates $Ti's$ and $c$)
3. while $\overline{T} < \rho$ ($\rho$-rejection criterion)
4. do $i \leftarrow i+1$ ;CRUZ[$c_i$]     (try next candidate)
5. CRUZ[c]  (2nd CRUZ running: to improve $c$ ($c^*$)).
6. CRUZ[c]  (3rd CRUZ running: to recalculate $r$)
7. If $T_{\min} < T_{Max} \sigma/100$   ($\sigma$-rejection criterion)
8. Then $i \leftarrow i+1$ ; go to 2 (try next candidate).
9. else return ( $c$, $T_{\min}$ ).
```

IV. Normalization: Improvement of the Second Phase

The purpose of this section is to provide some kind of standardization of samples, in pursuit of obtaining improvements to further stages of a BSI.

Most of the normalization methods consist in obtaining a feature matrix smaller than the original I sample [5] [6] [8] [20] [21] [22]. Dougman [5] proposed to build a normalized matrix, *N*, based on an iris sub-circumferences selection. To build the matrix *N* this methodology uses polar representation with **θ** and **r** parameters, where:

- **r** (radial resolution) is the amount of regular spaced sub-circumferences to take from the circular crown given by the previous phase. These circumferences will be the rows of matrix *N*.
- θ (angular resolution) is the amount of regular spaced radios to take from the iris segmentation; these radios will be the columns of matrix *N*.

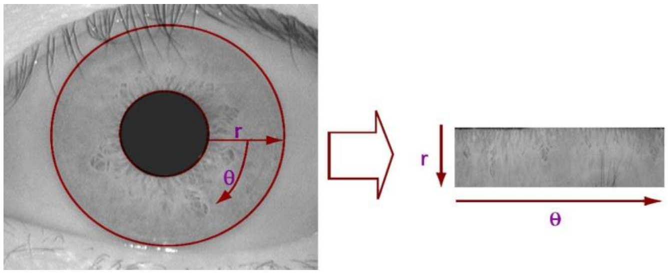


Fig. 5. Graphic representation of Daugman method. Left: original sample. Right: normalization of the original sample.

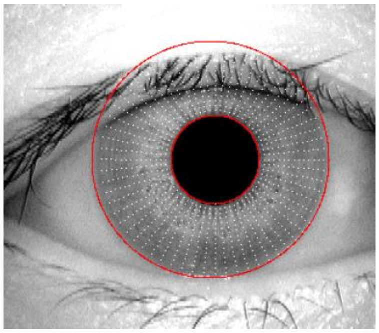


Fig. 6. Selected pixels by normalization with parameters **'θ'**=65 and 'r'=15.

Daugman's normalization is based on an arbitrary selection of regular spaced pixels. The risk of this method is that it may produce a weak representative selection of iris texture. In addition, attacking this issue by **θ** and **r** increasing, will result in a bigger *N* matrix, and so, in a higher computational cost. In this work we proposed an improvement of Daugman's normalization. This new normalization procedure is based on the use of the sample mean and involves all of the pixels inside the iris. We decided to use the sample mean, because after several proofs we obtained better results than using other statistic functions for example median or standard deviation. Our normalization method produces the *M* normalized matrix as follows:

1) Apply Daugman's method with appropriate parameters in order to produce an output matrix N that contains the whole of pixels of iris image. Note that matrix N will be redundant.

2) Divide N on n x m sub-matrices according to a grid of n x m blocks.

3) Define the M normalized matrix as M[i,j] equal to the sample mean of the values in the N sub-matrix of block (i,j), with $i \in \{0,1,..n-1\}$ *and* $j \in \{0,1,..,m-1\}$.

Figure 7 shows schematically the method for obtaining the *M* normalized matrix.

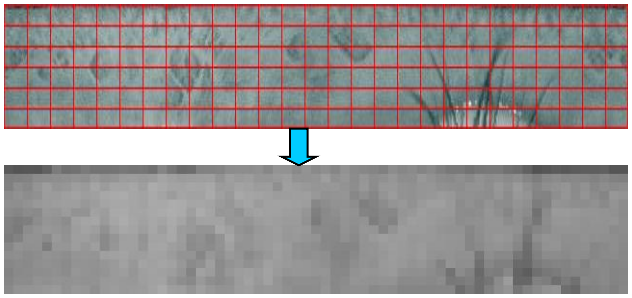


Fig. 7. Above, the matrix *N* from step 1, which contains all the pixels from the original sample. Overlapped in red, the grid of *n* x *m* blocks. Below, the final normalized matrix *M* obtained by computing the sample mean of each block.

Formally: Let θ_N and r_N be the Daugman's input parameters such that applying unwrapping over an iris image, the output matrix *N* contains all the pixels of the image. Let ℓ_H, ℓ_V be the horizontal and vertical lengths of every block respectively, defined by:

$$\ell_H = \left[\frac{\theta_N}{n}\right] \quad \text{and} \quad \ell_V = \left[\frac{r_N}{m}\right] \qquad (6)$$

Our normalization matrix *M* is defined by:

$$M[i,j] = \overline{B}_{i,j} \qquad (7)$$

where $i \in \{0,1,..\text{n-1}\}$ and $j \in \{0,1,,\text{m-1}\}$ and:

$$B_{i,j} = N[x,y] \qquad (8)$$

with:

$$i \cdot \ell_H \le x < (i+1) < \ell_H \cdot \min \theta_N$$

$$j \cdot \ell_V \le y < (i+1) < \ell_V \cdot \min r_N$$

V. Improving the Extraction Phase

The target of the Extraction phase in BSI is to find the biometric interest feature inside the normalized sample, and save it to build the so called "biometric code". This step is not really essential, since the normalized samples of the previous phase could be already used to identify persons. Anyway, we developed an additional new improvement of the BSI taking advantage of this phase and inspired by the following issue: If two samples have different base light level, the matching phase could fail even when the samples came from the same individual.

We defined our feature vector (biometric code) attempting to keep the change relation from one pixel to another as follows:

$$D[i,j] = M[i+1,j] - M[i,j] \qquad (9)$$

where *M* is the normalized matrix obtained after applying (7), and *D* is a matrix of dimension [(*n-1*) x *m*], called differential matrix.

Now, let us suppose that M_1 and M_2 are two normalized samples to be compared; suppose that M_2 is the same that M_1 but adding a constant *k* to every element of M_1. Note that we will get the same feature vector *D* for both M_1 and M_2 . This means that *D* can fix the base light issue successfully, and the matching phase will take advantage of this improvement.

VI. Results and Validation

We ran CRUZ algorithm in CASIA database version 1.0. The success-failure criterion in pupil segmentation stage was based on geometrical circle properties. Additionally, the success in the segmentation process was visually verified. We obtained a 100% effectiveness in 4.45 seconds (approximately 5 milliseconds per sample). We calculated an "average stepping" of 1.067. This parameter measures the number of pixels that CRUZ algorithm discarded before obtaining the center and radius of the circle in the pupil segmentation process.

The Matching phase results were evaluated according to the following two criteria:

- Individual's Matching criterion: We counted one success every time our method matched an individual with any of its testing samples. (Success over number of individuals).
- Sample's Matching criterion: We counted one success every time our method matched a sample with the right individual. (Success over number of samples).

Matching phase task is to measure the level of similarity between two biometric codes to establish whether these came from the same individual or not. As in most of the works in this area ([6] [8] [21]), we used Hamming distance to compare two iris codes. Given D_1 and D_2, two differential matrices of [(*n-1*) x *m*] dimension, the Hamming distance between D_1 and D_2 is defined as follows:

$$d(D_1,D_2) = \sum_{j=0}^{m-1}\sum_{i=0}^{n-1} \frac{\left|D_1[i,j] \oplus D_2[i,j]\right|_1}{nmB} \qquad (10)$$

where *B* is a constant that indicates the amount of bits necessary to represent every D_1 (or D_2) element, $\oplus$ corresponds to XOR operator and "$|\ |_1$" counts no-nulls bits in binary representation.

Our Matching results compared to another development presented in [6] are summarized in Table 1.

TABLE I. Matching Results

Methods	Parameters	
	Computational Cost (milliseconds)	*Matching Effectiveness*
Daugman [5]	285	*99,90%*
LiMa and Tan [21]	95	*99,23%*
Boles and Boashashe [22]	55	*93,20%*
Wavelet Multiscale [8]	81	*99,60%*
Our Method	**73**	**95.8% ~ 100%**$^{(a)}$

[a] 95.8% effectiveness for "sample's matching" criterion, and 100% for "individual's matching" criterion.

VII. Conclusions

A. Segmentation

Our results are highly positive because the analysis and segmentation of 756 images in the entire dataset only takes 4.45 seconds in a low profile standard laptop. The value obtained for average stepping shows that the first pixel selected by the CRUZ algorithm was already inside the pupil in most of the cases.

B. Matching

The computational cost of our methodology is lower than in most of the methods presented in [8]. The Matching effectiveness in CASIA database Version 1.0 is superior or equal to 95.8%. When we used the individual's matching criterion, our method reached the 100% of matching effectiveness. In consequence, we were capable of identifying all the individual of the CASIA database version 1.0 [15].

C. Globals

In general we verified informal ideas such as:

- "It was possible to resolve the segmentation problem in a very much simpler way and without complex mathematical models"
- "Most of the existing methods put a big effort and high complexity into taking the most accurate segmentation, we committed to improving the other steps of the system awaiting for competent results"

Indeed, we obtained competitive results spending fewer computing resources on the segmentation step (resigning perhaps some accuracy), and proposing then some prior improvements before comparison by Hamming.

In other words, we achieved a drastic complexity reduction by segmenting the inner pupil edge with our CRUZ algorithm and taking advantage of anatomical standards for the outer edge, opposing the possible accuracy loss in segmentation through improvements implemented in the normalization step and the use of differential matrix before comparison.

VIII. Future Scopes

At present the problem of matching based on iris segmentation is a matter of big interest in the forensic and security areas. The behavior of our proposal using color images of faces is still being a pending matter. Likewise, the study of the effectiveness of our matching method, considering other measures of similarity between biometric codes, is matter an interesting open problem to be addressed in the future. The analysis of CRUZ algorithm performance using another database is also a pending issue.

References

[1] Anil K. Poor, Arun Ross, Salil Prabhakar "An introduction to Biometrics Recognition". IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 14, NO. 1, JANUARY 2004.

[2] J. G. Daugman. "High confidence visual recognition of persons by a test of statistical independence". IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 15, nº 11. Noviembre 1993.pp. 1148-1161.

[3] S. Sanderson, J. Erbetta. "Authentication for secure environments based on iris scanning technology". IEE Colloquium on Visual Biometrics, 2000.

[4] Zohaib Bukhari, Bilal Shams "Iris biometrics recognition system and comparison of biometrics techniques", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 7– No.9, September 2014 – www.ijais.org

[5] J. Daugman, "How iris recognition works", IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, pp. 21–30, 2004

[6] R. Wildes, "Iris recognition: an emerging biometric technology", Proceedings of the IEEE,vol. 85, pp. 1348–1363, 1997.

[7] Bernd Jähne. (2004) "Practical handbook on image processing for scientific and technical applications", 2nd Ed. CRC Press LLC.

a. *Bouridane, Imaging for forensics and security, signals and communicationtechnology, DOI 10.1007/978-0-387-09532-5 1.*

[8] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Efficient iris recognition by characterizing key local variations," IEEE Transactions on Image Processing, vol. 13, pp. 739–750, June 2004.

[9] L. Masek, "Recognition of human iris patterns for biometric identification." http://www.csse.uwa.edu.au/ pk/studentprojects/libor, 2003.

[10] C.L.Tisse, L.Martin, L.Torres, and M.Robert, "Person identification technique using human iris recognition," in Proceedings of Vision Interface, (Canada) pp. 294–299, 2002.

[11] H. Sung, J.Lim, J.Park, and Y.Lee, "Iris recognition using collarette boundary localization," in Proceedings of 17th International Conference on Pattern Recognition (ICPR'04), vol. 4, pp. 857–860, August 2004.

[12] J. Cui, Y. Wang, T. Tan, L. Ma, and Z. Sun, "An iris recognition algorithm using local extreme points," in First International Conference Biometric Authentication (D. Zhang and A. K. Jain, eds.), vol. 3072 of Lecture Notes in Computer Science, pp. 442–449, Springer, 2004.

[13] Jesse Russell, Ronald Cohn "Hamming distance" Book on Demand, 2012 ISBN 5512146971, 9785512146972.

[14] Chinese Academy of Sciences, CASIA-IrisV1, http://biometrics.idealtest.org/

[15] Farouk, R. 2011. "Iris recognition based on elastic graph matching and Gabor wavelets". Computer Vision and Image Understanding 115 (2011) 1239-1244. Elsevier

[16] Somnath Dey, Debasis Samanta. "A novel approach to iris localization for iris biometric processing." World Academy of Science, Engineering and Technology International Journal of Medical, Health, Pharmaceutical and Biomedical Engineering Vol:1 No:5, 2007

[17] Wenbo Dong, Zhenan Sun, Tieniu Tan "Iris matching based on personalized Weight Map" IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 9, September 2011

[18] http://www.exelisvis.com/ProductsServices/IDL.aspx

[19] R. Sanchez-Reillo and C. Sanchez-Avila, "Iris recognition with low template size," Proceedings of the International Conference Audio and Video-Based Biometric Person Authentication, pp. 324–329, 2001.

[20] L. Ma, T. Tan, et al. "Efficient iris recognition by characterizing key local variations," IEEE Transactions on Image Processing, vol. 13, pp. 739–750, 2004.

[21] W. Boles and B. Boashash, "A human identification technique using images of the iris and wavelet transform", IEEE Transactions on Signal Processing, vol. 46, pp. 1185–1188, 1998.

A Text Classifier Model for Categorizing Feed Contents Consumed by a Web Aggregator

H.O.D. Longe
Department of Computer Science,
University of Lagos

Fatai Salami
Department of Computer Science
Bells University of Technology, Ota, Nigeria

***Abstract*—This paper presents a method of using a Text Classifier to automatically categorize the content of web feeds consumed by a web aggregator. The pre-defined category of the feed to be consumed by the aggregator does not always match the content being consumed and categorizing the content using the pre-defined category of the feed curtails user experience as users would not see all the contents belonging to their category of interest. A web aggregator was developed and this was integrated with the SVM classifier to automatically categorize feed content being consumed. The experimental results showed that the text classifier performs well in categorizing the content of feed being consumed and it also affirmed the disparity in the pre-defined category of the source feed and appropriate category of the consumed content.**

Keywords—feed; aggregator; text classifier; svm

I. Introduction

Web feeds provide a way for websites especially those that are frequently updated to provide up to date information to their users. Feeds are provided in either RSS or Atom format.

Users who are interested in consuming the content of feeds use an aggregator software called feed reader. Aggregator software can either be a windows or a web application and it collects feed contents from various sources in one view. With a feed reader, a user can have the latest content of his/her favourite website in one place; thereby reducing time spent checking different websites. A spin-off of feed readers is web aggregation sites. A web aggregation site is a website that has content from various feeds in one place. This makes it easier for users to view contents from various websites at once. It also removes the overhead of having to build the content of a feed aggregator by the user. Popular aggregation websites include newsnow.com, kicknews.com.

When aggregators have to categorize the content consumed from feeds, they either use a predefined category that has been registered for the source of the feed or try to get the category from the meta-data supplied with the feed content. Using the predefined category of the source brings up scenario in which the category does not match the actual content being consumed. In some cases also, the category supplied in the meta-data would not match any of the categories set up in the aggregator.

The categorization of content from feeds can be achieved via the use of Text Classifiers. Text Classifiers are algorithms that are used to carry out Text Categorization (TC). In formal terms, taking a document di from a set of documents D and categories {c1, c2, c3}, text categorization is assigning a category ci to document di [11]. Example of text categorization algorithms include; K Nearest Neighbour (KNN), Naïve Bayes (NB), Support Vector Machines (SVM).

In TC, documents may be classified in such a way that it can only belong to one category (single-label categorization) or can belong to multiple categories (multi-label categorization) [15]. Multi-label categorization is better suited to aggregators because the content consumed from a feed can belong to multiple categories. Example, a story about a Nigerian footballer getting married to a Nollywood (Nigerian movie industry) actress can rightly belong to categories about sports, gossip and entertainment.

The paper is organized as follows. Section II contains a review of existing literatures in the field of Text Categorization. It is followed by system architecture and software design in Section III. The categorization process is discussed in Section IV and implementation and evaluation of the system is in Section V. Conclusion is made in Section VI.

II. Related Work

There are two main approaches to building text classifiers – Knowledge Engineering (KE) approach and Machine Learning (ML). Knowledge Engineering (KE) used to be very popular. It involves manually defining a set of rules encoding knowledge from experts to place texts in specified categories. KE gradually lost its popularity in the 1990's to Machine Learning (ML) approach which involves building automatic text classifier by learning the characteristics of the categories of interest from a set of pre-classified texts [18].

In deciding whether to use Machine learning or Knowledge Engineering approach to text classification, sentences in Dutch Law were classified using both Machine Learning technique and Knowledge engineering approach [7]. SVM and pattern based KE were implemented and was found that SVM attained accuracy of up to 90%.

A Scientific News Aggregator that gathered news from both Atom and RSS feeds of about 1000 web journals was developed in [19]. NB classifier was used to classify the news coming from the different sources into stipulated categories of interest. Since a relatively large part of the RSS/Atom feed was already manually classified from the originating news source, the key idea implemented for classifying was to use the classifier in a mixed mode: as soon as already classified scientific news by a scientific news source was seen, the classifier switched to training mode; the remaining unclassified scientific news was categorized with the classifier in categorizing mode.

Multi-label classification was implemented by [4]. A ranking function was used to compute the relevancy of all predefined categories to the news item. The contents of <title>, <description> and <link> elements were retrieved and used as features. Normalized term frequency method was used to determine the weight of individual feature in the vector space.

SVM was used by [12] to classify news articles into three categories; Sports, Business and Entertainment. The vector representation of features serves as entry point into the SVM classifier. The SVM classifier was implemented using LIBSVM - an integrated software for support vector classification, regression and distribution estimation [one-class SVM] with the support for multiclass classification.

Categorization of news text using SVM and ANN was carried out in [2]. In the overall comparison of SVM and ANN algorithms for the data set that was used, the results for both recall and precision over all conditions indicate significantly differences in the performance of the SVM algorithm over the ANN algorithm and since SVM is a less (computationally) complex algorithm than the ANN, they concluded that SVM is preferable at least for the type of data examined, i.e., many short text documents in a relatively few well populated categories.

A method of Text Categorization on web documents using text mining and information extraction based on the classical summarization techniques was proposed in [9]. First, web documents are pre-processed by removing the html tags, meta-data, comment information, images, bullets, buttons, graphics, links and all other hyper data in order to establish an organized data file, by recognizing feature terms like term frequency count and weight percentage of each term. Experimental results showed that this approach of Text Categorization is more suitable for Informal English language based web content where there is vast amount of data built in informal terms. The method significantly reduced the query response time, improved the accuracy and degrees of relevancy.

In [16], rough set theory was used to automatically classify text documents. After pre-processing text documents and stemming the features, they used specific thresholds of 10%, 8%, 6% and 4% to reduce the size of the feature space based on the frequency of each feature in that text document. Thereafter, their model used a pair of precise concepts from the rough set theory that are called the lower and upper approximations to classify any test text document into one or more of main categories and sub-categories of interest. The rough set theory produced accuracy of up to 96%.

SVM and NB classifiers were used to categorize Arabic texts in [1]. In the Arabic dataset that was used, each document was first processed to remove digits and punctuation marks and then some letters were normalized after which stop words were removed. They used three parameters for their evaluation – precision, recall and F1 and SVM outperformed NB with respect to all the evaluation parameters.

The combination of SVM and Elitist Genetic Algorithm (EGA) was applied to the classification of Chinese text by [10]. Genetic Algorithms (GA) are used to determine the values of parameters such as the regularization parameter (C) when used in combination with SVM. However, it is possible that some better solution found in previous steps may be lost because of the genetic operation in traditional GA. This led to introduction of memory to keep track of the better solutions that would have otherwise been discarded. Elite survival strategy is employed in combing algorithm, EGA-SVM. The results obtained in their evaluation showed that the EGA-SVM outperformed GA-SVM and ordinary SVM.

Ttext categorization was used to detect intrusion by [9]. KNN classifier was used for the classification. System processes were taken as documents to be classified and system calls were taken as distinct words. The tf-idf text categorization weighting technique was adopted to transform each process into a vector. Their preliminary result showed that the text categorization approach is effective in the detection of intrusive program behaviour.

SVM was used as the classification algorithm in this paper because it has high dimensional input space, understands that there are few irrelevant features and tries to use as many features as possible, the documents' vectors are sparse and most text categorization problems are linearly separable [6].

III. Proposed Web Aggregator System Architecture

The architecture as shown in Fig. 1 consists of a user that makes request to view information from the aggregator, an application server which serves the pages and connects the system to the internet, a feed database that contains the information about registered feeds, training data for the Categorization engine and retrieved contents by the Content Retrieval engine. It also includes a Content Retrieval Engine which retrieves new contents from the registered feeds and a Categorization Engine which carries out the categorization process.

A. The Feed Database

This consists of six entities. The Entity Relationship Diagram (ERD) presented in Fig. 2 shows all the entities in the Feed Database and the relationship between them. The entities in the Feed database are: Category – contains the categories used in the aggregator, Feed – registered feeds to retrieve contents from, Post – contents retrieved from registered feeds, PostCategory – categories assigned to the retrieved content by the Categorization engine, PostView – a count of the number of times a particular post has been viewed and TrainingPost – retrieved contents that would be used to train the categorization engine.

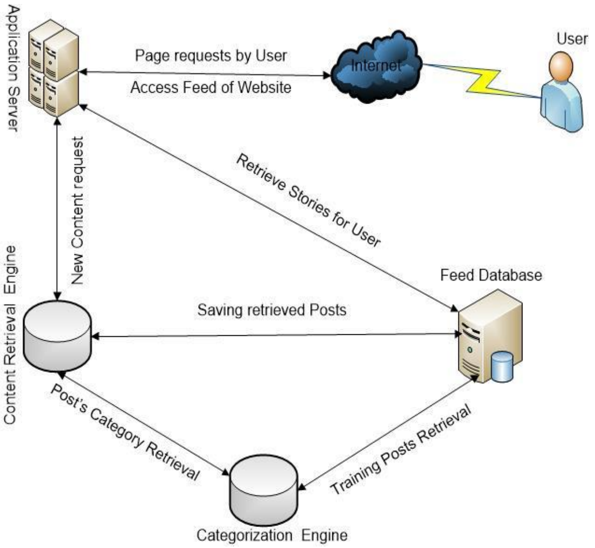


Fig. 1. Architecture for Web Aggregator

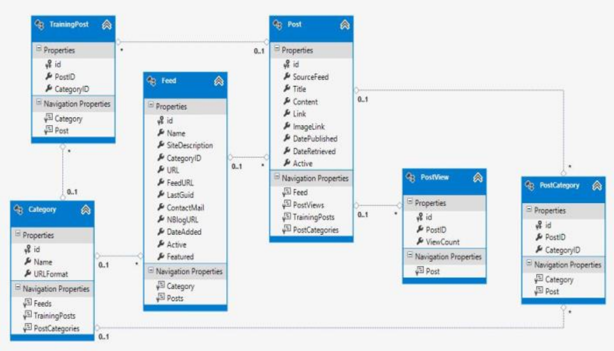


Fig. 2. Entity Relationship Diagram of Feed database

B. The Content Retrieval Engine

It retrieves most recent yet to be retrieved contents from the registered feeds.

The steps to retrieve new content are as follows:

1) Retrieve all Feeds to be polled for content from Feed database and store as ListFeeds.

2) Set ListPost as list of posts to be added to database, ListUpdate as list of Feeds to update their LastGuid and ListPostCategory as Categories determined for the Contents.

3) For each Feed in ListFeeds repeat steps 4 to 16.

4) Download XML of Feed.

5) Determine type of Feed and adjust tags to examine appropriately.

6) Set LatestGuid = Guid of the most recently published content in the feed, usually the first.

7) If LatestGuid = LastGuid of the Feed, Go to next Feed in ListFeeds else continue to 8.

8) Set count = 0, maxCount = maximum number of posts to retrieve and PostGuid = null.

9) If PostGuid = LastGuid of the Feed or count >= maxCount; Add LatestGuid and Feed to ListUpdate then Go to next Feed in ListFeeds ELSE select content as Post.

10) Process Post to remove all unnecessary HTML tags.

11) Add processed Post to ListPost.

12) Set ListCat as categories determined for the Post by the Categorization Engine.

13) Add ListCat to ListPostCategory.

14) Set count = count + 1.

15) Set PostGuid = Guid of Post.

16) Go to 9.

17) Save ListPost and ListPostCategory to Feed database and update Feed table using ListUpdate.

C. The Categorization Engine

This makes use of SVM classifier to classify contents. The literatures reviewed showed that the SVM is one of the best classifiers available hence its choice for this paper. The Categorization Engine builds SVM model which is required for categorization using the Posts saved to the TrainingPost table in Feed Database. The TrainingPost table had 1020 manually categorized posts which were retrieved from some Nigerian blogs and websites. The spread of the training posts among the various categories is presented in Table I.

The Categorization Engine also determines the categories that best fits a post retrieved by the Content Retrieval Engine. Since the project looks at the possibility of placing a retrieved content in more than one category, SVM multi-label classification class is employed. The result returns a list of possible categories for the retrieved content.

TABLE I. TRAINING POST SPREAD AMONGST CATEGORIES.

Category	Number of Training Data
Business	104
Current Affairs	130
Education	92
Entertainment	124
Gossip	134
Jobs	109
Personal	80
Politics	65
Science & Technology	80
Sports	102

IV. OVERVIEW OF THE CATEGORIZATION PROCESS

The text categorization process can be divided into seven sub processes – Read document, Tokenize text, Stemming, Stop words removal, Vector representation of text, Feature Selection and/or Feature Transformation (Dimensionaliity Reduction) and Learning Algorithm. The Feature Selection and/or Feature Transformation phase was not used in this paper because the contents of Feeds are usually a summary and often times already have few features. A diagrammatic representation of the categorization process is shown in Fig. 3.

The Read Document phase was achieved by supplying the categorization engine with string representation of content to be categorized. Tokenization of Text removed punctuation marks and separated the text into individual words.

Stop Words removal involved removing words with little semantic meaning from the tokens. The list of stopwords used in this paper was gotten from [17].

The stemming process involves getting the stem terms for words. This is done by removing the suffix from words. The Porter Stemmer is a conflation Stemmer developed by Martin Porter and it is based on the idea that the suffixes in the English language are majorly made up of a combination of smaller and simpler suffixes. The Porter Stemmer Algorithm is widely used and it is probably the stemmer most widely used in IR research [8].

The vector space representation involves converting the words in the text to be categorized into SVM matrix representation of words. The general format of the vector space representation for SVM is:

<label> <index>:<value> <index>:<value>

<label> is the number representation of the category of the text to be classified. A random category amongst the legal categories can be selected. The id value in the Category table of Feed Database is used to represent the categories. <index> is the number representing the stemmed word and <value> is the tf.idf value of the word. The <index> values are arranged in alphabetical order.

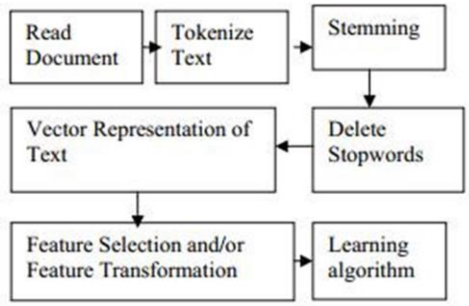

Read Document
Tokenize Text
Stemming
Vector Representation of Text
Delete Stopwords
Feature Selection and/or Feature Transformation
Learning algorithm

Fig. 3. Text Categorization Process (Source: [5]).

The learning algorithm that was used in this work is the SVM algorithm. There are several implementations of the SVM algorithm. LibSVM.Net which is the .Net implementation of LibSVM [3] was used in this project. Modification was made to LibSVM.Net to allow it accept string inputs instead of the default text document. LibSVM first builds a Model using the vector space representation of the training data along with a set of parameters.

A. Vector Space Representation Process

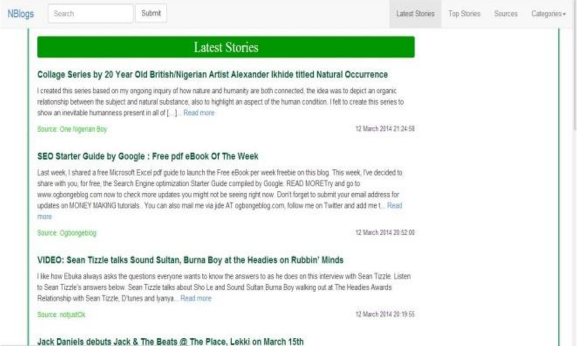
The algorithm used to carry out the vector space representation process is as follows:

```
1) Initialize BagOfWords = combination of all ListStemWords for all training data arranged alphabetically.
2) Initialize ListCategorizingWords = ListStemWords for the text that is to be categorized arranged alphabetically.
3) Initialize string VSR which would hold the vector space.
4) For each word W in ListCategorizingWords.
5) If W exists in BagOfWords go to 6 ELSE go to next W.
6) Set string S = W's index in BagOfWords + ":".
7) Calculate the tf.idf frequency of W as ti.
8) Set string S = S + ti + " ".
9) Set VSR = VSR + S.
10) Go to next W.
11) Return VSR.
```

V. Implementation and Evaluation

A. Web Aggregator User Interface

The web aggregator developed called "NBlogs" was based on the concept of responsive design. A responsive website is a website that automatically adjusts the screen size to fit the size of the screen from which it is being viewed from whether a desktop, a tablet pc or a smart phone. Twitter bootstrap package was used in the design to achieve responsiveness. Fig. 4 shows what the home page of NBlogs looks like in a desktop browser while Fig. 5 shows the same home page on a smaller screen. C# programming language was used in coding the business logic. NBlogs runs on .Net's MVC framework. MSSQL server was used to house the Feed Database.



Latest Stories

Collage Series by 20 Year Old British/Nigerian Artist Alexander Ikhide titled Natural Occurrence

I created this series based on my ongoing inquiry of how nature and humanity are both connected, the idea was to depict an organic relationship between the subject and natural substance, also to highlight an aspect of the human condition. I felt to create this series to show an inevitable humanness present in all of [...]... Read more

Source: One Nigerian Boy 12 March 2014 21:24:58

SEO Starter Guide by Google : Free pdf eBook Of The Week

Last week, I shared a free Microsoft Excel pdf guide to launch the Free eBook per week freebie on this blog. This week, I've decided to share with you, for free, the Search Engine optimization Starter Guide compiled by Google. READ MORETry and go to www.ogbongeblog.com now to check more updates you might not be seeing right now. Don't forget to submit your email address for updates on MONEY MAKING tutorials . You can also mail me via jide AT ogbongeblog.com, follow me on Twitter and add me t... Read more

Source: Ogbongeblog 12 March 2014 20:52:00

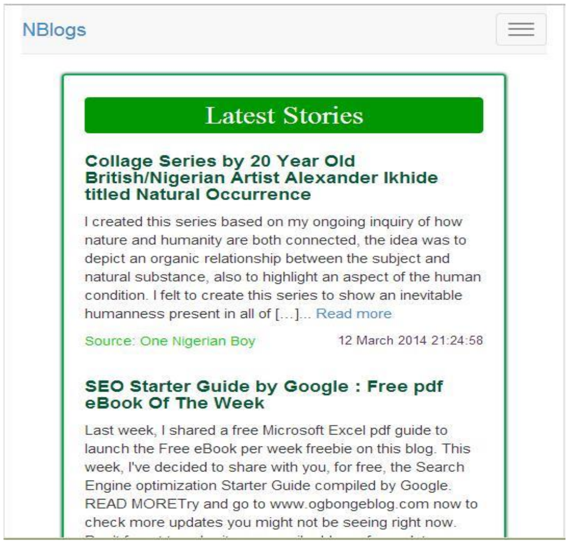
VIDEO: Sean Tizzle talks Sound Sultan, Burna Boy at the Headies on Rubbin' Minds

I like how Ebuka always asks the questions everyone wants to know the answers to as he does on this interview with Sean Tizzle. Listen to Sean Tizzle's answers below. Sean Tizzle talks about Sho Le and Sound Sultan Burna Boy walking out at The Headies Awards Relationship with Sean Tizzle, D'tunes and Iyanya... Read more

Source: notjustOk 12 March 2014 20:19:55

Jack Daniels debuts Jack & The Beats @ The Place, Lekki on March 15th

Fig. 4. Web Aggregator Home Page on Desktop browser.



Latest Stories

Collage Series by 20 Year Old British/Nigerian Artist Alexander Ikhide titled Natural Occurrence

I created this series based on my ongoing inquiry of how nature and humanity are both connected, the idea was to depict an organic relationship between the subject and natural substance, also to highlight an aspect of the human condition. I felt to create this series to show an inevitable humanness present in all of [...]... Read more

Source: One Nigerian Boy 12 March 2014 21:24:58

SEO Starter Guide by Google : Free pdf eBook Of The Week

Last week, I shared a free Microsoft Excel pdf guide to launch the Free eBook per week freebie on this blog. This week, I've decided to share with you, for free, the Search Engine optimization Starter Guide compiled by Google. READ MORETry and go to www.ogbongeblog.com now to check more updates you might not be seeing right now.

Fig. 5. Web Aggregator Home Page on smaller screen

B. Performance Evaluation of Categorization Algorithm

The evaluation of classifiers can be carried out using metrics such as precision, recall and F-Measure.

Recall is the proportion of real positive cases that are actually predicted as positive while Precision is the proportion of Predicted Positive cases that are correctly Real Positives (Powers, 2011).

Recall = r = $\frac{TP}{TP + FP}$

Precision = p = $\frac{TP}{TP + FN}$

Where:

TP = True Positive – predicted the right category for a story.

FP = False Positive – predicted category is wrong category for a story.

FN = False Negative – category was not rightly predicted for a story.

A total of one hundred and fifty (150) stories were retrieved from feeds to test the Categorization Engine. The stories were categorized into one hundred and ninety six (196) categories. The result of categorization including the TP, FN, FP, p and r is presented in Table 2. The bar graph of p and r is presented in Fig. 6.

F-Measure is the harmonic mean of the recall and precision with interval between 0 and 1 with a high F-Measure indicating a high quality classifier. The micro-averaged F-Measure is computed over all categories and it is achieved by summing the individual precision and recall scores for the categories. The macro F-Measure score is first computed over the individual categories before an average is taken (Ozgur, Ozgur, and Gungor, 2005).

Micro-Averaged F-Measure can be calculated as:

$$\frac{2(sr * sp)}{sr + sp}$$

Where:

sr =

$$\frac{\sum_{i=1}^{N} TPi}{\sum_{i=1}^{N}(TPi + FPi)}$$

sp =

$$\frac{\sum_{i=1}^{N} TPi}{\sum_{i=1}^{N}(TPi + FNi)}$$

N = number of categories.

Macro-Averaged F-Measure can be calculated as:

$$\frac{\sum_{i=1}^{N} FMi}{N}$$

Where

N = number of categories

FM_i =

$$\frac{2(ri * pi)}{ri + pi}$$

ri = recall of category i.

pi = precision of category i.

The Micro-Average F-Measure computed from Table 2 above is *0.731457801* while the Macro-Average F-Measure computed from the same table is *0.721934751*. The F-Measure values indicate a high quality classifier.

C. Effect of Text Classifier on Post Categories

Table 3 presents the distribution of posts after categorization has been carried out. PC is the number of posts that were categorized to be in the same category as the category registered for the feed while PD is the number of posts that were categorized in a different category to the category of the feed. %PD is the percentage of posts for that category that were placed in a different category. Overall, 68% of retrieved feed content were placed in a different category compared to the category of the source feed.

TABLE II. CATEGORIZATION RESULT

Category	True Positive (TP)	False Positive (FP)	False Negative (FN)	Precision (p)	Recall (r)
Business	8	9	3	0.73	0.47
Current Affairs	6	6	1	0.86	0.50
Education	5	0	2	0.71	1.00
Entertainment	25	8	6	0.81	0.76
Gossip	6	7	7	0.46	0.46
Jobs	13	0	4	0.76	1.00
Personal	27	11	11	0.70	0.71
Politics	3	1	0	1.00	0.75
Science & Technology	38	5	14	0.73	0.88
Sports	12	6	4	0.75	0.67

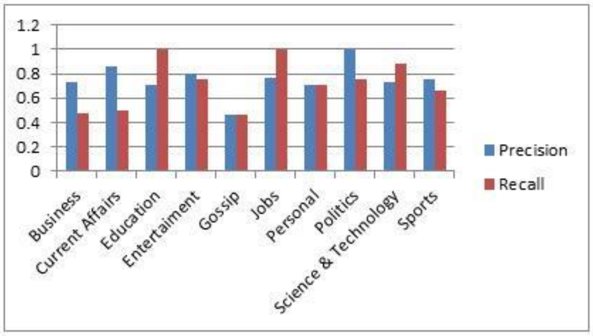


Fig. 6. Precision and Recall Grapgh

TABLE III. CATEGORIZING USING FEED CATEGORY AGAINST CATEGORIZATION ALGORITHM

Feed Category	PC	PD	% PD
Business	3	18	86
Current Affairs	1	22	96
Education	5	1	16
Entertainment	21	38	64
Gossip	2	13	87
Jobs	8	12	60
Personal	4	44	91
Politics	0	3	100
Science & Technology	34	24	41
Sports	12	22	64

VI. CONCLUSION

In this paper, text categorization algorithm was used to categorize the contents of feed consumed by a web aggregator. With training data obtained from the feeds of Nigerian websites, a SVM model was constructed to carry out the categorization.

The result obtained showed that the categorizer is of a high quality with a Micro-Average F1 measure of 0.731457801 and Macro-Average F1 measure of 0.721934751 and it further showed that it is not reliable to categorize contents consumed from a feed using the pre-defined category of the Feed as 68% of the feed content retrieved was placed in a different category by the SVM classifier.

The use of text categorization algorithm in web aggregators would improve user experience as they would be able to more easily access stories of interest to them.

REFERENCES

[1] Alsaleem, S., 2011. Automated Arabic Text Categorization Using SVM and NB. International Arab Journal of e-Technology, Volume 2, No. 2, pp. 124-128.

[2] Basu, A., Watters, C. and Shepherd, M., 2003. Support Vector Machines for Text Categorization.

[3] Chih-Chung Chang and Chih-Jen Lin, 2011. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.

[4] Darabi, M., Adeli, H. and Tabrizi, N., 2012. Automatic Multi-Label Categorization of News Feeds.

[5] Ikonomakis, M., Kotsiantis, S. and Tampakas, V., 2005. Text Classification Using Machine Learning Techniques. Wseas Transactions on Computers, Volume 4, Issue 8, pp. 966-974.

[6] Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. University of Dortmund Computer Science Department.

[7] Krabben, K., 2010. Machine Learning vs. Knowlegde Engineering in Classification of Sentences in Dutch Law. BSc. Universiteit Van Amsterdam

[8] Lancaster University, 2014. What is Porter Stemming?.

[9] Liao, Y. and Vemuri, R.V., 2002. Using Text Categorization Techniques for Intrusion Detection. In: USENIX Association, 11th USENIX Security Symposium. San Francisco, California, USA August 5-9, 2002.

[10] Liu, X. and Fu, H., 2012. A Hybrid Algorithm for Text Classification Problem.

[11] Manne, S. and Sameen, F.S., 2011. A Novel Approach for Text Categorization of Unorganized data based with Information Extraction. International Journal on Computer Science and Engineering (IJCSE), Volume 3, Issue 7, pp. 2846-2854.

[12] Mayor, S. and Pant, B., 2012. Document Classification Using Support Vector Machine. International Journal of Engineering Science and Technology (IJEST), Volume 4, No.04, pp. 1741-1745.

[13] Ozgur, A., Ozgur, L. and Gungor, T., 2005. Text Categorization with Class-Based and Corpus-Based Keyword Selection. P. Yolum et al.(Eds.): ISCIS 2005, LNCS 3733, pp. 606–615, 2005.

[14] Powers, D.M.W., 2011. Evaluation: From Precision, Recall and F-Measure to ROC,Informedness, Markedness & Correlation. Journal of Machine Learning Technologies ISSN: 2229-3981 & ISSN: 2229-399X, Volume 2, Issue 1, 2011, pp-37-63.

[15] Rafi, M., Hassan, S. and Shaikh, M.S., 2011. Content-based Text Categorization using Wikitology. National University of Computer and Emerging Sciences (NU-FAST) Karachi, Sindh, Pakistan.

[16] Sadiq, A.T. and Abdullah, S.M., 2013. Hybrid Intelligent Techniques for Text Categorization. International Journal of Advanced Computer Science and Information Technology (IJACSIT), Volume 2, No. 2, pp. 23-40.

[17] Savoy, J., 2005. IR Multilingual Resources at UniNE. Universite de Neuchatel.

[18] Sebastiani, F., 2001. Machine Learning in Automated Text Categorization.

[19] Shaikh, F., and Rajawat, A., 2012. Approach for Developing Scientific News Aggregators Using ATOM Feeds. International Journal of Electronics and Computer Science Engineering (IJECSE), Volume1, Number 4, pp. 2279-2284.

Frequency Estimation of Single-Tone Sinusoids Under Additive and Phase Noise

Asmaa Nazar Almoosawy
MSc Candidate
Department of Physics
Faculty of Education for Women
University of Kufa, Najaf, Iraq

Zahir M. Hussain
Professor; Dept. of Computer Science
University of Kufa
P.O.Box 21, Kufa, Najaf, Iraq
Professor (Adjunct), ECU, Australia

Fadel A. Murad
Assistant Professor
Department of Physics
Faculty of Education for Women
University of Kufa, Najaf, Iraq

***Abstract*—We investigate the performance of main frequency estimation methods for a single-component complex sinusoid under complex additive white Gaussian noise (AWGN) as well as phase noise (PN). Two methods are under test: Maximum Likelihood (ML) method using Fast Fourier Transform (FFT), and the autocorrelation method (Corr). Simulation results showed that FFT-method has superior performance as compared to the Corr-method in the presence of additive white Gaussian noise (affecting the amplitude) and phase noise, with almost 20dB difference.**

Keywords—Frequency Estimation; Correlation; Cramer-Rao Bound; Phase Noise; Maximum Likelihood Estimator

I. INTRODUCTION

The frequency estimation (IF) of a complex sinusoidal signal in white Gaussian noise is one of the major problems in the literature. This is so because IF has been applied widely in many areas such as radar, sonar, communications and image analysis [1-5]. There is a variety of approaches to the frequency and phase estimation problem, with differences in performance as regards frequency estimation accuracy and computational complexity [5]. In many applications, it is necessary to detect the frequency of a single tone in a noisy environment. Taking the Discrete Fourier Transform (DFT) using FFT algorithm of the collected samples is the most common method of making such a frequency estimate. Practical limitations like the computational complexity can restrict the number of samples under processing (hence, the amount of signal information), a factor that will restrict the resolution of the estimate provided by the DFT [6]. The maximum likelihood estimator (MLE) to estimate the frequency of a sinusoid damaged by additive Gaussian noise was thoroughly studied by Rife and Boorstyn [7]. Quinn [8] developed a simple and efficient method to estimate the frequency of a single-tone sinusoidal signal based on the three samples around the DFT maximum (peak). A similar method was developed by Grandke [9]; this method uses the DFT maximum point (in the frequency domain) along with only one adjacent frequency. Both of the above methods are efficient in frequency estimation in terms of good performance (accuracy of frequency estimation) at higher noise powers (i.e., low SNRs that may reach 0dB). However, neither of these two methods can directly give a good magnitude estimate, also, both methods require division operation [6].

In this paper we will estimate the frequency of a single-tone sinusoid under AWGN and phase noise (PN) using two most popular methods: MLE method through using Fourier Transform (FT) (calculated by Fast Fourier Transform algorithm, FFT), and the Correlation method (Corr). The latter has been traditionally preferred to MLE for being computationally less intensive than FFT. Frequency estimation based on Fourier transformation is explained in Section 2, while in Section 3 we explain the autocorrelation method. Section 4 provides simulation results and performance comparison between the two methods.

II. FREQUENCY ESTIMATION BASED ON FOURIER TRANSFORM

Let the signal to be a single-tone sinusoid as follows:

$$x(t) = A \cdot \sin(\omega_o t + \emptyset_o) + \epsilon(t) \qquad (1)$$

where, A is the signal amplitude, ω_o is the frequency of the signal, $\emptyset_o$ is the initial phase and $\epsilon(t)$ is an additive noise process. Noise is assumed to be Gaussian white noise process with $\mathcal{E}[\epsilon]=0$ ($\mathcal{E}$ being the expectation functional) and var $[\epsilon] = \sigma^2$.

Assuming that all the above parameters are unknown, we try to get an estimate for the frequency ω_o as $\widehat{\omega}$. The estimate should be as accurate as possible, also, it should not be computationally intensive [10].

Two important quantities associated with any estimate is the bias, $b[\widehat{\omega}] = \mathcal{E}[\widehat{\omega}] - \omega$, and the variance, given by $\mathrm{var}[\widehat{\omega}] = \mathcal{E}[\left(\widehat{\omega} - b(\widehat{\omega})\right)^2]$.

For unbiased estimators (bias=0), an important performance measure is the Cramér-Rao bound (CRLB), which represents the minimum possible variance for the unbiased estimator when noise effect decreases or the *Signal-to-Noise Ratio* (SNR) increases. The CRLB of the unbiased frequency estimator has been formulated as follows [11]:

$$\mathrm{CRB}_\omega = \frac{1}{\mathrm{SNR}} \frac{6}{N(N^2-1)} \qquad (2)$$

where N is the number of signal samples and SNR is the signal - to - noise ratio ($= A^2/(2\sigma^2)$).

We know that FT method estimates the frequency by the peak of the Fourier Spectrum $X(f)$ of the sinusoidal signal

$x(t)$, computed from the sampled signal $x(n)$ by the DFT as $X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) \exp(-\frac{2\pi nk}{N})$.

However, the actual frequency of a signal may not fall on one of the above frequencies of the DFT bins, hence; we use the magnitudes of the nearby bins to determine the actual signal frequency through the process of interpolation. There are several interpolation methods as follows.

A. *Quadratic Interpolation:*

This method finds a quadratic fit $y = a + bx + cx^2$ in the neighborhood of the maximum $\max\{X(f)\}$ with the three points [5]:

$(K-1, u_1 = |X_{K-1}|)$,
$(K, u_2 = |X_K|)$,
and $(K+1, u_3 = |X_{K+1}|)$,

where $K = \arg\{\max_k [X(k)]\} =$ index of the absolute maximum magnitude of the DFT, which refers to the actual frequency $F = K f_s / N$, f_s being the sampling frequency.

Now the actual maximum given by the quadratic formula above will be at the point $y = -b/(2c)$ as follows:

$u = K + d$;
where $d = (u_3 - u_1) / [2 * (2 * u_2 - u_1 - u_3)]$.
The estimated frequency is $F_o = \frac{u f_s}{N}$.

The **Barycentric method** is similar, with $u = K + d$; where $d = (u_3 - u_1) / (u_1 + u_2 + u_3)$.

B. *Quinn's First Estimator [8]:*

Taking the three DFT points:

$(K-1, v_1 = X_{K-1} = r_1 + i \cdot s_1)$,
$(K, v_2 = X_K = r_2 + i \cdot s_2)$,
and $(K+1, v_3 = X_{K+1} = r_3 + i \cdot s_3)$,

we perform the following calculations:

$R = r_2^2 + s_2^2$
$p = (r_3 \cdot r_2 + s_3 \cdot s_2) / R$;
$g = -p/(1.0 - p)$;
$q = (r_1 \cdot r_2 + s_1 \cdot s_2) / R$;
$e = q/(1.0 - q)$;
If $(p > 0)$ and $(q > 0)$ then , $d = p$, else , $d = q$,
Now: $u = K + d$.

C. *Quinn's Second Estimator [12]:*

Using the above three points with other quantities, we have:

$d = \frac{p+q}{2} + h(p^2) + h(q^2)$; $u = K + d$,

where $h(x) = \frac{\frac{1}{4}\ln(3x^2+6x+1) - \frac{\sqrt{6}}{24}\ln(x+1-\sqrt{\frac{2}{3}})}{x+1+\sqrt{\frac{2}{3}}}$.

Estimating the frequency $f_o = \omega_o / 2\pi$ using Quinn's second estimator has the least RMS error; however, in our simulation we used the Quadratic Method with frequency compensation:

$$d = \frac{(u_3 - u_1)}{(4 * u_2 - 2u_1 - 2u_3)} \qquad (3)$$

III. Frequency Estimation Based on Autocorrelation

The autocorrelation algorithms are to extract the frequency from the phase of the available signal's autocorrelation with fixed lags.

The periodogram-based estimators use the Discrete Fourier Transform (DFT) for a coarse search and an interpolation technique for a fine search [13]. In correlation-based single-tone frequency estimation, consider the single-tone model as per Equation (1). For correlation-based estimators, an estimate of the frequency is obtained by the information of one or several estimated entries of the auto-correlation sequence of $x(n)$:

$$r(m) = \mathcal{E}[x(n)\, x^*(n-m)] = |A|^2\, e^{i 2\pi f_o m} + \sigma^2\, \delta_{m,0} \qquad (4)$$

where $(\mathcal{E}[\cdot])$ denotes statistical expectation, $\delta_{m,0}$ is the Kronecker delta, σ^2 is the noise variance as defined in Equation (1), and $(*)$ denotes complex conjugation. Note that since noise is uncorrelated with itself, its autocorrelation is a delta function (exists at lag $m = 0$ only).

We can find the autocorrelation sequence $\{r(m)\}$ from the data sequence as follows:

$$r(m) = \frac{1}{N-m} \sum_{n=0}^{N-1} x(n) x^*(n-m) \qquad (5)$$

Note that $r(m) = r(-m)$.

From Equation (5), we may have close information about the frequency f_o from the phase angle of $\{r(m)\}$, that is, if we exclude the case $m = 0$ in order not to interfere with the noise effect, we have:

$m \cdot \omega_o =$ phase $[r(m)]$ ⟨ mod $[2\pi]$⟩
$m \cdot \omega_o =$ phase $[r(m)] + 2\pi k$; $k =$ integer   (6)

The integer k satisfies $0 \le k < m$. As we want positive results for frequency, the angle and mod $[2\pi]$ operation are restricted to the interval [0,2π). Also, only positive values of m are considered in our simulations.

The first possible frequency estimate from Equation (7) is obtained by putting $k = 0$; hence, if we choose the first autocorrelation sample at $m = 1$, we have:

$\omega_o =$ phase $[r(1)]$

This estimator is known as the minimal order linear predictor [14]. It is also a special case of the Pisarenko harmonic decomposer frequency estimator [15]. It is shown that the performance of this linear predictor can be improved by using a different correlation lag [16]. In [17] - [18], it was shown that the estimator based on a single correlation coefficient can be made more efficient.

A disadvantage with the above estimators (other than the fundamental estimator) is the ambiguity to the frequency estimate [19], [20]. It is shown in [21] that the frequency ambiguity could be resolved using two correlations with relatively prime correlation lags; this is further explained in [22], [23].

IV. Frequency Estimation Under Gaussian and Phase Noise

The works of frequency estimation in the literature have tested the above algorithms only under additive Gaussian noise (AWGN), however, no test has been performed under phase noise (PN).

The main source of noise in electronic and communication systems is the thermal noise. This noise process (which is normally additive) is generated due to the random thermal agitation of free electrons as an electrical current passes through a conductor. This type of noise is white, i.e. it is composed of all frequencies. Another form of noise affecting communication systems is called phase noise [24]. This noise is created during the process of combination and recombination of charge carriers inside the molecular structure of the semiconductor. Hence, the sinusoidal signal with a fundamental frequency f_o is disturbed by noise in the phase part, leading to a slight fluctuation in the instantaneous frequency. This is so because the instantaneous frequency $f(t)$ and phase $\varphi(t)$ are related by the instantaneous formula [4]:

$$f(t) = \frac{1}{2\pi}\frac{d\varphi(t)}{dt} \qquad (7)$$

In this work, we consider phase noise (PN) affecting the phase of a single-tone sinusoid as follows:

$$x(t) = A \cdot \cos\big(\omega_o\, t + \emptyset_o + \rho(t)\big) + \epsilon(t) \qquad (8)$$

where A is the signal amplitude, ω_o is angler frequency, $(\emptyset_o)$ Initial phase, $\rho(t)$ is the phase noise and $\epsilon(t)$ is the additive white Gaussian noise. This is just an extension to Equation (1) above. The above parameters are assumed to be *unknown*. We formulated PN as Gaussian noise added to the phase of the signal. This is the simplest model for phase noise.

V. Simulation Results

We simulated the above algorithms with signal model with AWGN and PN as per Equation (8) using MATLAB. The simulated signal has time length $L = 70$s, sampling interval $T_s = 0.001s$, $f_s = 100\,Hz$, and a number of samples $N = [\frac{L}{T_s}]$. The signal amplitude is $A = 1$ volt, ω_o is angler frequency $\omega_o = 2\pi f_o$, where $f_o =$ 23 Hz. We modeled PN as zero-mean Gaussian noise. Monte Carlo simulations were performed with $M = 100$ realizations. We used the quadratic frequency compensation as per Equation (4):

$u = K + d$; with $d = (u_3 - u_1)\,/[2 * (2 * u_2 - u_1 - u_3)]$, and estimated frequency $F_o = \frac{uf_s}{N}$.

The signal-to-noise ratio (SNR) is still defined as before, i.e., using the AWGN power only. This is so because the phase noise power is affecting the phase only but not the amplitude of the signal.

Finally, we calculate the relative squared-error under each SNR and PN power as follows:

$$e = |((F_o - f_o)/f_o)|^2$$

As for the frequency estimated by correlation, we do not calculate all the correlation coefficients of the signal to get the estimate, but only the 2nd coefficient was considered. Note that we used Hilbert transformation (HT) to get the analytic signal $z(t)$ associated with the original signal $x(t)$ before estimation. This is to remove the negative part of the signal spectrum $X(f)$, where:

$$z(t) = x(t) + j \cdot \mathcal{H}[x(t)]$$

$$\mathcal{H}[x(t)] = [\frac{1}{\pi t}] *_t [x(t)]$$

noting that $*_t$ denotes time-convolution, and $\mathcal{H}$ denotes HT [25]. Hence:

$$Z(f) = X(f)[1 - j^2 \cdot \text{sgn}(f)] = X(f)[1 + \text{sgn}(f)]$$

$$\therefore \quad Z(f) = \begin{Bmatrix} 2X(f); & f \geq 0 \\ 0; & f < 0 \end{Bmatrix}$$

Therefore, using HT will not affect the frequency estimation.

After estimation, we calculate relative squared-error for each SNR as follows:

$$\mathbb{e} = \left|\frac{(\mathcal{F}_o - f_o)}{f_o}\right|^2$$

Finally, we draw our results as shown in Figures (1) and (2).

Note that taking more correlation coefficients (hence, more estimations for the frequency) will give more accurate results, but this is not recommended for real-time applications.

Figure (1) shows the estimated frequency versus SNR using interpolated FT peak and correlation methods for various powers of phase noise (PN). Numbers 1, 2, 3 correspond to PN powers of -50, 1, 5 dB, respectively. Note that FT hold in a high SNR less -30dB, as for to the correlation method holds to -15dB , It is clear that PN does not affect CRB, as all curves converge to the same asymptote for large SNR. For all PN powers, FT peak outperforms correlation by almost 15 dB. Also, it is clear that FT and correlation have the same CRB [as per Equation (2)], since both estimates have the same asymptote.

Figure (2) shows the frequency estimation mean-squared error (MSE) versus SNR using interpolated FT peak and correlation methods for various powers of phase noise (PN). Numbers 1, 2, 3 correspond to PN powers of -50, 1, 5 dB, respectively. It is clear that FT peak is more robust under very low SNR; however, it is more computationally expensive. This is not a surprise because correlation is highly dependent on phase.

VI. Conclusions

We tested two popular frequency estimation algorithms, MLE through FFT and Correlation, using complex single-tone sinusoid affected by additive Gaussian and phase noise. Results of implementing these methods in MATLAB helped in comparing between them as follows:

- Fourier Transform (FT) approach is more efficient than the correlation approach (Corr) for frequency estimation. This is so because FT can work under low SNRs (as low as -30 dB), while the lowest SNR for the correlation method is (-15dB), hence there is about (-15dB) difference between the two approaches.

- FT outperforms Corr under phase noise, as it gives better f_0 estimation (lower error) at higher PN power values. This is so because Corr method is dependent on phase, so it will be more sensitive to phase noise.
- It is clear that PN does not affect CRB, as all error curves converge to the same asymptote for large SNR. Hence, both FT and Corr approaches have the same CRB.
- Despite the superiority of FT in frequency estimation as compared with Corr, the FT approach is computationally expensive. This so because FT requires the whole signal and estimates the frequency from the peak of FT, while in Corr approach we can take one correlation coefficient to estimate the frequency.

REFERENCES

[1] ZHANG Gang-bing, LIU Yu, XU Jia-jia, HU Guo-bing, “Frequency Estimation Based on Discrete Fourier Transform and Least Squares,” IEEE International Conference on Wireless Communications & Signal Processing (WCSP), 2009.

[2] Zahir M. Hussain and Boualem Boashash, “Multi-component IF estimation,” Proceedings of the IEEE Signal Processing Workshop on Statistical Signal and array Processing (SSAP'2000), Pocono Manor, Pennsylvania, USA, pp. 559-563, 14-16 Aug. 2000.

[3] Zahir M. Hussain and Boualem Boashash, “Design of time-frequency distributions for amplitude and IF estimation of multicomponent signals,” invited paper for the Statistical Time-Frequency Special Session in the Sixth International Symposium on Signal Processing and Its Applications (ISSPA'2001), vol. 1, pp. 339-342, Aug. 2001.

[4] Zahir M. Hussain and Boualem Boashash, “Adaptive instantaneous frequency estimation of multi-component FM signals using quadratic time-frequency distributions,” IEEE Transactions on Signal Processing, vol. 50, no. 8, pp. 1866 –1876, August 2002.

[5] Yizheng Liao, Phase and Frequency Estimation: High-Accuracy and Low-Complexity Techniques, M.Sc. Thesis, Worcester Polytechnic Institute, 2011.

[6] E. Jacobsen, “On Local Interpolation of DFT Outputs,” EF Data Corporation Report [Online, 1994]. Available: http://www.ericjacobsen.org/FTinterp.pdf.

[7] D. C. Rife, R. R. Boorstyn, “Single-Tone Parameter Estimation from Discrete-Time Observations,” IEEE Trans. on Information Theory, v. 20, n. 5, 1974.

[8] B. G. Quinn, "Estimating Frequency by Interpolation Using Fourier Coefficients," IEEE Trans. Signal Processing, Vol. 42, no. 5, 1994.

[9] T. Grandke, "Interpolation Algorithms for Discrete Fourier Transforms of Weighted Signals," IEEE Trans. Instrumentation and Measurement, Vol. IM-32, no.7, 1983.

[10] B. Bischl, U. Ligges, C. Weihs, "Frequency Estimation by DFT Interpolation: A Comparison of Methods,” Technical Report, Technische Universität Dortmund, 2009.

[11] V. Clarkson, “Efficient Single Frequency Estimators,” International Symposium on Signal Processing and Its Applications (ISSPA), 1992.

[12] B. G. Quinn, "Estimation of Frequency, Amplitude, and Phase from the DFT of a Time Series," IEEE Trans. Signal Processing, Vol. 45, no. 3, 1997.

[13] Cui Yang, Gang Wei, and Fang-jiong Chen, “An Estimation-Range Extended Autocorrelation-Based Frequency Estimator,” EURASIP Journal on Advances in Signal Processing, Volume 2009.

[14] L. B. Jackson and D.W. Tufts, “Frequency Estimation by linear prediction,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’78), USA, 1978.

[15] P. Händel, “Markov-Based Single-Tone Frequency Estimation,” IEEE Trans. Circuits Syst. II, vol. 45, no. 1, 1998.

[16] G. W. Lank, I. S. Reed, and G. E. Pollon, “A Semi-Coherent Detection and Doppler Estimation Statistic,” IEEE Trans. Aerosp. Electron. Syst., vol. AES-9, 1973.

[17] S. Kay, “A Fast and Accurate Single Frequency Estimator,” IEEE Trans. Acoustics, Speech, and Signal Processing, no. 12, 1989.

[18] E. Jacobsen and P. J. Kootsookos, "Fast, Accurate Frequency Estimators," IEEE Signal Processing Magazine, May 2007.

[19] P. Händel, A. Eriksson, and T. Wigren, “Performance Analysis of a Correlation Based Single Tone Frequency Estimator,” Signal Processing, vol. 44, no. 2, no. 6, 1995.

[20] M. P. Fitz, “Further Results in the Fast Estimation of a Single Frequency,” IEEE Trans. Coinniun.1994.

[21] D. W. Tufts and P. D. Fiore, “Simple, Effective Estimation of Frequency Based on Pony’s Method,” IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 5, 1996.

[22] P. Händel, B. Völcker, and B. Göransson, “Analysis of a Simple, Effective Frequency Estimator Based on Prony’s Method,” IEEE Signal Process. Workshop Stat. Signal Array Process., Sept. 1998.

[23] B. Völcker, P. Händel, “Frequency Estimation From Proper Sets of Correlations,” IEEE Trans. Signal Process., vol. 50, no. 4, April 2002.

[24] R. Corvaja and S. Pupolin, “Phase noise effects in QAM systems,” IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun., vol. 2, no. 2, Sep. 1997.

[25] Zahir M. Hussain and Boualem Boashash, “Hilbert transformer and time-delay: statistical comparison in the presence of Gaussian noise,” IEEE Transactions on Signal Processing, vol. 50, no. 3, pp. 501-508, March 2002

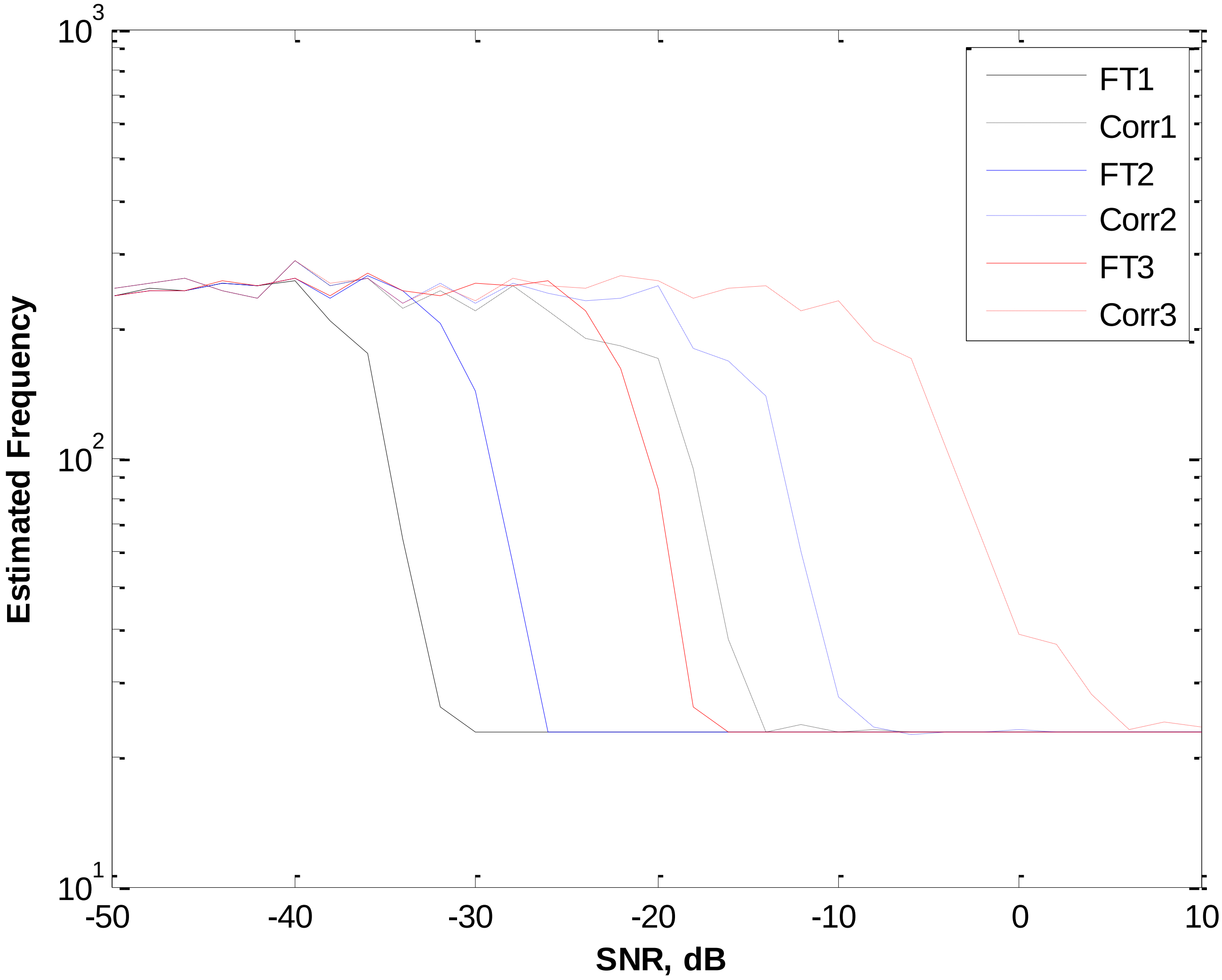


Fig. 1. Estimated frequency versus SNR using interpolated FT peak and correlation methods for various powers of phase noise (PN). Numbers 1, 2, 3 correspond to PN powers of -50 (no noise), 1, 5 dB, respectively. Note that SNR is only considered for AWGN.

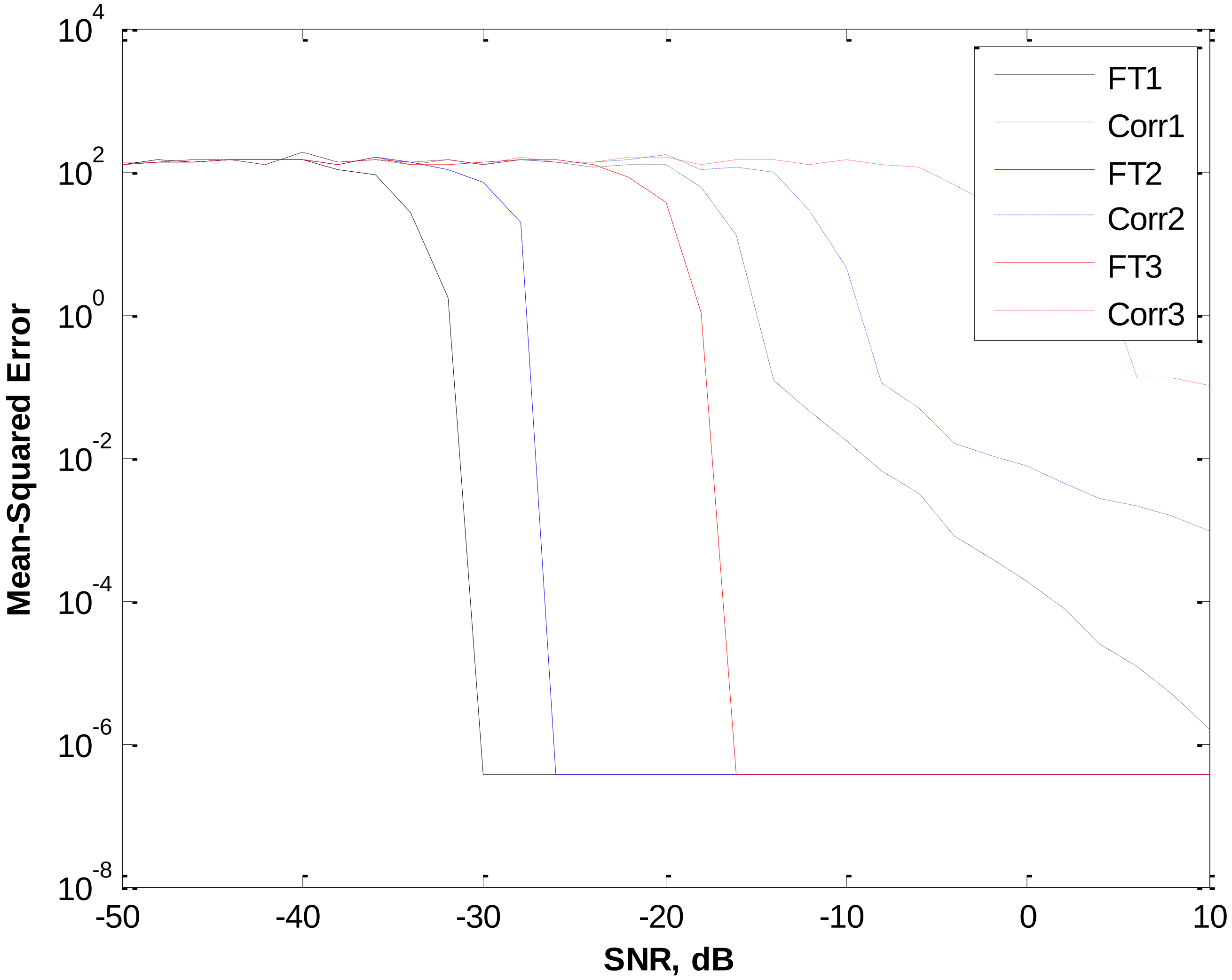


Fig. 2. Frequency estimation mean-squared error (MSE) versus SNR using interpolated FT peak and correlation methods for various powers of phase noise (PN). Numbers 1, 2, 3 correspond to PN powers of -50, 1, 5 dB, respectively. It is clear that FT peak is more robust under very low SNR.

Literature Survey of previous research work in Models and Methodologies in Project Management

Ravinder Singh
AVP, JP Morgan Chase & Co
Research Scholar,
Department of Informatics,
King's College, London, UK

Dr. Kevin Lano
Reader
Department of Informatics,
King's College, London, UK

***Abstract*—This paper provides a survey of the existing literature and research carried out in the area of project management using different models, methodologies, and frameworks. Project Management (PM) broadly means programme management, portfolio management, practice management, project management office, etc. A project management system has a set of processes, procedures, framework, methods, tools, methodologies, techniques, resources, etc. which are used to manage the full life cycle of projects. This also means to create risk, quality, performance, and other management plans to monitor and manage the projects efficiently and effectively.**

Keywords—Programme/ Program Management, Project Management, Maturity Models, Onshore-offshore Management, Leadership and Management, Global Distributed Projects

I. Introduction

Projects are considered at all levels of organisation which may involve one or many business units and they can involve one or 100s of persons. The duration of projects can vary from a few weeks to many years. Projects can be simple to highly complex projects, which may be implemented at one location or multiple locations across multiple countries.

There are two main PM standards i.e. Project Management Body of Knowledge (PMBOK) by PMI, USA and PRINCE-2 by APMG, UK.

Considering the PMBOK and PRINCE-2 standard as the base, models may be defined for the process and knowledge management areas of PM. The benefits of having models includes support for defining tools for project management, deliverables and documentation, and to improve consistency between different areas and processes of PMBOK. It may be able to benefit other projects as the tools, documents and deliverables can be reused. The use of modelling will also bring consistency and reliability in the management of various projects. This will make project monitoring and controlling easier during the execution and whole life cycle of the project. The final advantage will come in the form of delivering good quality projects on time, budget and scope with improved quality. This will lead to overall customer satisfaction as risk and issues have been managed efficiently and effectively.

According to an independent study, undertaken and conducted by Loudhouse Research, on behalf of CA in 2007 on the "The changing face of project management" studying the Project panorama in UK corporates and observing the project failure as well as excellence. Study redefines the failure, emphasising the strategic value the projects deliver rather than exclusively on delivery within budget and time.

- 26% of surveyed companies are spending more than half of IT budget on IT projects
- Average company is managing 29 projects at once and 1 in 10 companies is managing more than 100 projects at a time
- More than half of IT Directors (52%) think projects are more complex than 5 years ago, fuelled by increased demand from business (62%) and a more pressurised regulatory environment (55%)
- For 59%, less than half of initiatives are strategic.
- 53 % projects are based on business processes improvement,
- 36% on growing revenues
- 32% on bringing new products/ services

Survey also covered various issues in managing projects effectively. Some of these are as follows:

- Typical budget over-runs 30%; 1 in 6 projects going more than 50 % over budget; 10 out of 29 projects on the go at one time will come in over budget
- Inaccurate scoping and forecasting is blamed for budget over-runs in half of cases (50%). With scope creep responsible in 4 out of 10 cases.
- Only a 3rd (35%) of businesses check that initiatives are aligned with business objectives with only 1 in 8 companies basing this decision on real time, accurate information.
- 74% struggle to access critical skills
- 39% of IT Directors don't have complete visibility over all IT initiatives in progress
- 42% of IT directors know within a day if an initiative is off-course

Therefore a standard methodology is required to deliver the projects efficiently and effectively within time and budget.

There are two main standards for project management i.e. Project Management Body of Knowledge (PMBOK) by PMI, USA and PRINCE-2 by APMG, UK. PMBOK is more dominant standard as this is used in more than 75% of the projects around the world.

Software/ IT projects use various methodologies such as SSADM, RUP, Spiral Model, Scrum, Extreme Programming, etc. and management methodologies such as PMBOK, PRINCE2.

According to Project Management Body of Knowledge (PMBOK) [1], a project is a temporary endeavour undertaken to create a unique product, service or result. According to PMBOK project management is realised through the combination and practice of five project management processes: Initiating, Planning, Executing, Monitoring and Controlling, and Closing. PMBOK divides project management into ten knowledge areas of Integration, Scope, Time, Cost, Quality, Human Resource, Communication, Risk, Procurement, and Stakeholder Management. Project management is the effective use of processes, procedures, tools, techniques along with knowledge, and skills to meet project objectives. Project managers have to manage the traditional Triple constraints – "Scope, Time, and Cost", which have been enhanced in recent time with three more constraints called "Quality, Risk and Customer Satisfaction".

PRINCE2 [2] describes a project as "A management environment that is created for the purpose of delivering one or more business products according to specified business needs". The PRINCE2 process model comprises of eight distinctive management processes namely Starting up a Project (SU), Directing a Project (DP), Initiating A Project (IP), Planning (PL), Managing Stage Boundaries (SB), Controlling a Stage (CS), Managing Product Delivery (MP), Closing a Project (CP) which covers the full life cycle of a project. PRINCE2 is a de facto standard used extensively in UK government.

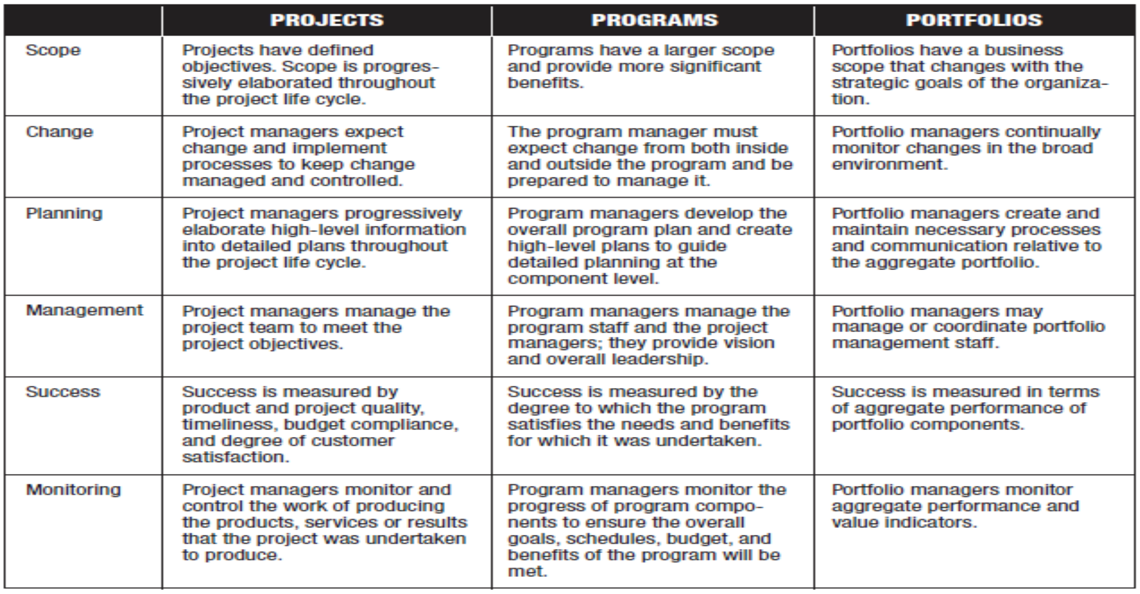
A key standard amongst maturity models for Portfolio, Programme, and Project Management is P3M3 [3]. This offers a framework for organizations with which they can assess their current performance and develop/ implement improvement plans with measurable outcomes based on industry best practice

II. PROJECT MANAGEMENT BODY OF KNOWLEDGE (PMBOK)

PMI (USA) developed a Project Management Body of Knowledge (PMBOK) [1], which can be described as collection of project management knowledge. Similar to other professions like engineering, medicine, accounting etc., this body of knowledge rests with professionals in the project management field who uses this knowledge and also propose advancements in it. PMBOK provides good practices, knowledge, tools, skills, and techniques for better, efficient and effective project management. It does not mean that all the things described can be applied to all projects, but has to be modelled and tailored depending on the various constraints of the project like size, budget, time, location etc. PMBOK also describes the common terminology and language for project documentation, reports, writing etc. This makes it easier for all to understand and work on the project by reducing the communication gap.

According to PMBOK [1], differences among project, program and portfolio management is as follows in Table 1.1:

Table 1-1. Comparative Overview of Project, Program, and Portfolio Management



	PROJECTS	PROGRAMS	PORTFOLIOS
Scope	Projects have defined objectives. Scope is progressively elaborated throughout the project life cycle.	Programs have a larger scope and provide more significant benefits.	Portfolios have a business scope that changes with the strategic goals of the organization.
Change	Project managers expect change and implement processes to keep change managed and controlled.	The program manager must expect change from both inside and outside the program and be prepared to manage it.	Portfolio managers continually monitor changes in the broad environment.
Planning	Project managers progressively elaborate high-level information into detailed plans throughout the project life cycle.	Program managers develop the overall program plan and create high-level plans to guide detailed planning at the component level.	Portfolio managers create and maintain necessary processes and communication relative to the aggregate portfolio.
Management	Project managers manage the project team to meet the project objectives.	Program managers manage the program staff and the project managers; they provide vision and overall leadership.	Portfolio managers may manage or coordinate portfolio management staff.
Success	Success is measured by product and project quality, timeliness, budget compliance, and degree of customer satisfaction.	Success is measured by the degree to which the program satisfies the needs and benefits for which it was undertaken.	Success is measured in terms of aggregate performance of portfolio components.
Monitoring	Project managers monitor and control the work of producing the products, services or results that the project was undertaken to produce.	Program managers monitor the progress of program components to ensure the overall goals, schedules, budget, and benefits of the program will be met.	Portfolio managers monitor aggregate performance and value indicators.

PMBOK [1] defines a project as follows:

- ✓ A temporary endeavour to create a unique product, service or a result.
- ✓ Creates a unique product, service or result.
- ✓ Is progressively elaborated – distinguishing characteristics of each unique project will be progressively detailed as the project is better understood.

According to PMBOK [1]:

- Temporary means that every project has a definite beginning and a definite end. Projects are not on-going efforts.
- Unique Product means a product or an artifact that is produced, is quantifiable, and can be either an end item in itself or component item.
- A capability to perform a service, such as business functions supporting production or distribution.
- A result such as outcomes or documents e.g. research project.

Progressive Elaboration is a characteristic of projects that accompanies the concepts of temporary and unique. This means developing in steps and continuing by increments. Progressive elaboration is not scope creep.

Management by objectives (MBO) has three steps:

- ✓ Establish unambiguous and realistic objectives.
- ✓ Periodically evaluate if objectives are being met.
- ✓ Implement corrective action.

Project managers have to manage the traditional triple constraints of Cost, Time, and Scope. The new demands in the project management emphasises three more constraints of Quality, Risk, and Customer Satisfaction which the project manager should manage. Prioritisation of the constraints may be managed by the management directly or indirectly.

The stakeholder in the project is someone whose interests and whose influence may be positively or negatively impacted by the projected. Negative stakeholders are often overlooked by the project team at the risk of failing to bring their project to successful end. The key stakeholders in the project are project manager, customer, performing organisation, users, project team, sponsor, and project management office (PMO). All of the stakeholders must be identified, documenting their requirements, communicating and managing their expectations and influences on the project. Project management team has professional responsibility towards all of its stakeholders and public in general. PMI members must adhere to the “Code of Ethics” and Project Management Professionals (PMP)” certification should adhere to a “Code of Professional Conduct”.

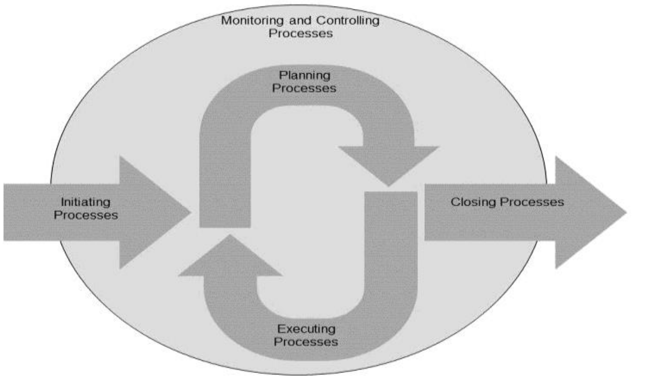
Projects are approved depending on the strategic objectives of the organisation such as Market demand, organisational need, customer request, technology advance, legal requirement, etc.

Project management system is described as a set of tools, techniques, methodologies, resources and procedures. PMI divides project management into professional and social responsibility knowledge areas and process groups. The PMBOK describes the integration among project management processes, the relations among them and function they serve.

PMBOK combines these processes into five process groups which are defined as Project Management Process Groups which are Initiating Process Group, Planning Process Group, Executing Process Group, Monitoring and Controlling Process Group, and Closing Process Group.

Even though the project management processes are presented as discrete elements, but they overlap and project management professionals can me manage the project in different ways. The objectives of the projects can be defined based on the complexity, risk, size, time, resources, documents, deliverables, application area, geographic spread, experience, and maturity of team and organisation. The concept of interaction and overlapping of PM processes can be traced to Plan-Do-Check-Act (PDCA) cycle.

The Planning Process Group corresponds to “Plan” component of PDCA, the Executing Process Group corresponds to “Do” component of PDCA, the Monitoring and Controlling Process Group corresponds to “Check and Act” component of PDCA as shown in the Figure 1 below:



Project Management Process Groups mapped to the PDCA Cycle (Plan-Do-Check-Act

Fig.1. Project Management Processes as PDCA cycle

PMI further divides the project management into ten Knowledge Areas which are Project Integration Management, Project Scope Management, Project Time Management, Project Cost Management, Project Quality Management, Project Human Resources Management, Project Communications Management, Project Risk Management, Project Procurement Management, and Project Stakeholder Management.

Project Integration Management: it describes the processes and activities which integrate various components of the project.

Project Scope Management: This describes the processes which are used to calculate the scope, and only the work required, for successful completion of the project.

Project Time Management: defines the processes for completing the project on time.

Project Cost Management: describes process for estimating, planning, budgeting and controlling the cost of the project.

Project Quality Management: defines the processes for assuring that the project is delivered as per the required standard and objectives.

Project Human Resource Management: defines the processes for managing the human resources for the project.

Project Communication Management: describes the collecting, processes for processing, sending and receiving the information to the appropriate channels.

Project Risk Management: describes the processes for managing, prioritising, and mitigating risk for the project.

Project Procurement Management: describes the processes for contract management for purchasing any services, results or products.

Project Stakeholder Management: discusses the processes for managing stakeholders and understanding their expectations and managing them effectively.

III. PRINCE-2

PRINCE (**PR**ojects **IN** **C**ontrolled **E**nvironments) [2] is a structured method for effective project management. The method was developed first in 1989 by CCTA (The Central Computer and Telecommunication Agency). The method was adopted from PROMPTII, a project management method developed by Simpact Systems Ltd in 1975. Office of Government (earlier CCTA) further enhanced the method continuously and PRINCE-2 was launched in 1996. PRINCE-2 is based on the information and experiences shared by various experts, professionals in project management field.

PRINCE-2 is a de-facto standard used extensively by UK government. This is also used widely in private sector, but more in UK than internationally.

PRINCE-2 defines a project as, "A management environment that is created for the purpose of delivering one or more business products according to a specified business case". Another definition of project according to PRINCE-2 is, "A temporary organisation that is needed to produce a unique and predefined outcome or result at a pre-specified time using predetermined resources".

According to Prince-2, project has the following characteristics:

- A finite and defined life cycle.
- Defined and measurable business products.
- A corresponding set of activities to achieve the business products.
- A defined amount of resources.
- An organisation structure, with defined responsibilities, to manage the project.

The PRINCE-2 process model is shown in the Figure 2 This model consists of eight distinctive management processes for the full life cycle of the project. The Planning process is used by four of the other processes.

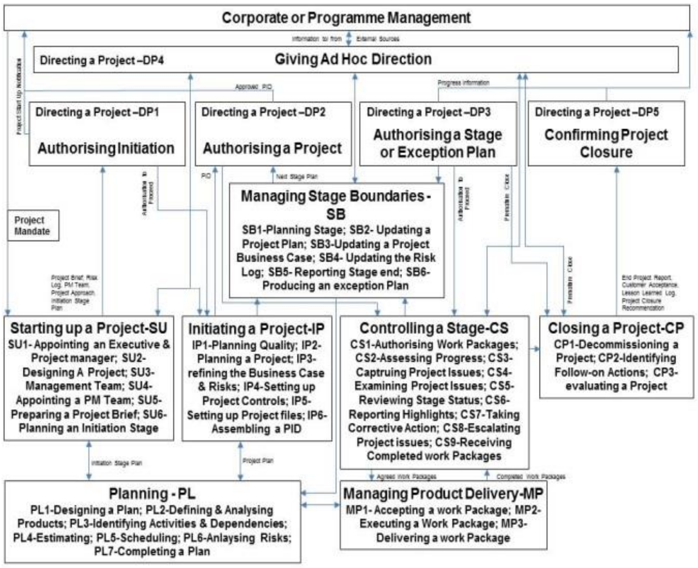

Corporate or Programme Management
Information to from External Sources
Directing a Project –DP4
Giving Ad Hoc Direction
Project Start Up Notification
Approved PID
Progress information
Directing a Project –DP1
Authorising Initiation
Directing a Project –DP2
Authorising a Project
Directing a Project –DP3
Authorising a Stage or Exception Plan
Directing a Project –DP5
Confirming Project Closure
Authorisation to Proceed
PID
Next Stage Plan
Project Mandate
Managing Stage Boundaries - SB
SB1-Planning Stage; SB2- Updating a Project Plan; SB3-Updating a Project Business Case; SB4- Updating the Risk Log; SB5- Reporting Stage end; SB6- Producing an exception Plan
Authorisation to Proceed
Premature Close
Premature Close
Project Brief, Risk Log, PM Team, Project Approach, Initiation Stage Plan
End Project Report, Customer Acceptance, Lesson Learned Log, Project Closure Recommendation
Starting up a Project-SU
SU1- Appointing an Executive & Project manager; SU2- Designing A Project; SU3- Management Team; SU4- Appointing a PM Team; SU5- Preparing a Project Brief; SU6- Planning an Initiation Stage
Initiating a Project-IP
IP1-Planning Quality; IP2- Planning a Project; IP3- refining the Business Case & Risks; IP4-Setting up Project Controls; IP5- Setting up Project files; IP6- Assembling a PID
Controlling a Stage-CS
CS1-Authorising Work Packages; CS2-Assessing Progress; CS3- Captruing Project Issues; CS4- Examining Project Issues; CS5- Reviewing Stage Status; CS6- Reporting Highlights; CS7-Taking Corrective Action; CS8-Escalating Project issues; CS9-Receiving Completed work Packages
Closing a Project-CP
CP1-Decommissioning a Project; CP2-Identifying Follow-on Actions; CP3- evaluating a Project
Initiation Stage Plan
Project Plan
Agreed Work Packages
Completed Work Packages
Planning - PL
PL1-Designing a Plan; PL2-Defining & Analysing Products; PL3-Identifying Activities & Dependencies; PL4-Estimating; PL5-Scheduling; PL6-Anlaysing Risks; PL7-Completing a Plan
Managing Product Delivery-MP
MP1- Accepting a work Package; MP2- Executing a Work Package; MP3- Delivering a work Package

Fig.2. PRINCE-2 Process Flow

A project must be able to make use of each of these processes in some form, and this may require tailoring of the processes for the needs of the individual project. Each process should be followed by asking a question about the relevance of the process for the particular project.

IV. PRINCE-2 STRENGTHS

- Organization (Project Boards; defined roles and responsibilities; ownership & accountability)
- Business case–based; on-going assessment of project viability by project owners (Board)
- Product-Based Planning (strictly deliverable oriented); Product Flow; Product Descriptions
- Integrated process structure: clear statement of how to manage the project ("How do I get started? What do I do first?")
- Clear quality management points (esp. Quality Control), and Quality Assurance roles and responsibilities
- Defined and orderly handling of Work Packages (Managing Product Delivery)

V. PMBOK AND PRINCE-2 COMPARISON

Basic contrast between PMBOK and PRINCE-2 is as follows in Table-2:

TABLE.II. PMBOK Vs PRINCE-2

PMBOK	PRINCE-2
Comprehensive	Focusses on key risk areas only; does not claim to be complete
Largely descriptive, prescriptive on a high level	Highly prescriptive, especially on Process Structure, but adaptable to any size project
Core and facilitating processes; need to be scaled to the needs of the project	All processes should be considered; also need to be scaled
Customer requirements driven	Business case driven
Sponsor and stakeholders	Clear project ownership and direction by senior management
International/ UK standard	UK standard

PRINCE2 is built on seven elements, or Themes: Business Case, Organization, Plans, Progress, Risk, Quality, and Change (comprising configuration management and change control). They roughly map against the nine PMBOK Knowledge areas as follows in Table-3:

TABLE.III. Knowledge Areas PMBOK Vs PRINCE-2

PMBOK Knowledge areas	PRINCE-2 components/ Themes
Integration	Combined Processes and Components/ Themes, Change Control
Scope, Time Cost	Plans, Business Case
Quality	Quality, Configuration Management (Change)
Risk	Risk
Communications	Controls
HR	Organisation (limited)
Procurement	Not Covered
Stakeholder	Not Covered

Five processes groups of PMBOK map against the PRINCE-2 processes as follows in Table-4:

TABLE.IV. Process Areas PMBOK Vs PRINCE-2

PMBOK	PRINCE-2 (Project Level)	PRINCE-2 (Stage Level)
Initiating	Starting Up; Directing	Managing Stage Boundaries; Directing
Planning	Initiating, Planning	Managing Stage Boundaries; Planning
Executing/ Controlling	Managed on a stage-by-stage basis	Controlling a Stage; Managing Product Delivery; Directing
Closing	Closing a Project	Managing Stage Boundaries

VI. Literature Survey

Previous research had been focusing on different aspects of the program and project management such as study of models and framework, empirical, and statistical studies. The studies had been conducted in different industry sectors but most of the research has been in the software and IT industry as given below in Table-5.

TABLE.V. Literature Survey

Ref. No.	Category/ Topic	Study Description/ Method/ Argument/ Theoretical Approach	Results, gaps, and Conclusions
	Project Management Models/ frameworks	The following papers [4-8] have used existing project management models and/ or framework for the studies and proposed a few modifications and amendments to make it more effective and efficient.	
4.		This paper discussed and compared project management / frameworks for managing software projects. Authors compared five different methodologies with PMBOK. This research can be of useful help for projects managers to decide on particular methodology for different projects.	This research suggests that project manager must have skills and knowledge of various methodologies used for successful completion of projects. The research highlighted that organisation adhere to one methodology due to cost involved, risks associated with other, and training required for its staff. The paper proposed a generic approach to project management.
5.		This research paper discussed three concepts for project approach. The first step is to develop global project framework for defining objectives, project life cycle and possible steps for projects repeatability. The 2nd step is to analyse and build project specifications, plans using transition graphs, and performance measurement baselines. Final step is to describe organisation design, roles, job specifications, deliverables and timelines, etc. for project manager and teams.	The paper proposed to follow one specific model but at the same time suggests changing the model dynamically with the project needs using modular and flexible approach. The results showed that projects can be managed in a better manner with clear deliverables and milestones using three step approach. The dynamic change in the deliverables/ milestones helps to manage risks and issues.
6.		The research decomposed the Industrial System development into number of phases. Defined the organisation process with results and deliverables i.e. documents or system components which are to be	The research proposed that a bigger system can be managed effectively and efficiently by converting it into smaller phases of planning, analysis, functional and technical design, build, test (unit, assembly, UAT, performance etc.), implementation and maintenance. Furthermore each phase can have transitions to effectively move from one

		transitioned from one phase to another. Studied different systems and proposed a new mode which was evaluated on another system showing better results.	phase to next. The transitions can be managed effectively with some people moving from one phase to next phase so that knowledge can be transferred fully.
7.		This research used object based models for better co-ordination of complex projects. Models used on chip design, identifying problems/ issues with PM tools. Object based models were created to overcome the above issues and summarise the stages/ phases of the project in better way for the project manager.	The object based model provided value addition by increasing the stability of the projects, co-ordination between PM tools and structure approach, and hence reducing uncertainty. Model used properties of inheritance from one phase to next, modularity to break project into small pieces of work, using relation to define interactivity. Data hiding is used to provide information/ access to the authorised people.
8.		This research highlighted that standard project management methodologies should be used by project managers. Project manager should create his/her own set of required documents, deliverables, processes etc. for more efficient approach and better synergy among various project activities and resources.	The paper proposed that project managers should tailor the standard project management processes as per their requirements so as to enhance the efficiency and synergy among project activities and resources. The tailoring could be in terms of deliverables, documents, tools, methods, etc. to be used depending upon the complexity of the project.
	Use of Modelling in Project Management	The following papers [9-12] have tried to implement the system modelling, estimation and object oriented concepts for better project management.	
9.		This research paper proposed a methodology to integrate the use of system dynamics approach to manage risks within the existing project management processes. Project risks change dynamically as the projects are being run and are difficult to predict and manage, this paper suggested the use of system dynamic modelling and framework for better monitoring and control within the PMBOK risk management processes.	Paper suggested the use of System dynamic modelling and framework within PMBOK processes for better management and control of risks in projects. Dynamic modelling provides better control of risks and project managers will be able to forecast and manage project milestones more effectively.
10.		This research suggested the use of system modelling as communication tool for evaluating and gathering stakeholder expectations in a better and efficient manner. Since communication levels and skills are different for project managers and business stakeholders, and hence there is a gap among customer expectations. This results in scope changes, time and budget changes, and quality. System modelling was used at MIT-Portugal Green Islands project. The results show that the communication gap can be bridged and stakeholders' expectations would be satisfied more effectively.	The research showed that System modelling could bridge the communication gap between clients and project managers. PMs would be able to define models to define communications channels, means, processes, and manage the objectives, requirements, and expectations effectively and efficiently. The research shows that model would be able to provide framework to reduce communication gap and understand stakeholder expectation, which resulted in better scope, time and budget control.
11.		This paper proposed the use of object oriented model for project management. The hierarchical structure is used to represent the full development life cycle of the software project and provide a view of various parts to the project manager. The interactivity among various components of the projects is defined as relations among objects.	Project activities are represented in text and graphical form. Paper demonstrated the used of various object oriented properties like inheritance, relations, modular, and data encapsulation for sharing information through different phases of the life cycle of the project. Various objects could be reused in different projects for better control.
12.		This paper proposed an estimation model for determining the project effort based on use-case. Relationship between actual and estimated data is developed to be used for better estimates in future. The model works better in iterative environments as it allows comparing of successive developments.	The research suggested an estimation model based on actual and estimated data to predict better results in future. A mathematical model is created between actual and estimated data for more accurate forecasting. This model is more suitable for iterative development environment.
	Project Management in IT/ Software Projects	The research on software and IT projects had been discussed in the research papers [13-22]. Authors had explored not only the standard project management methodologies like PMBOK, Prince2, RUP but also the new practices of Agile, JAD, RAD, extreme programming. Researchers had also suggested adaptations of project management models for managing complex projects in the software and IT industry.	
13.		This research paper discussed how the traditional process models in software engineering are being modified/ replaced with new web based models/ agile development due to more dynamic nature of projects and fast changing requirements.	The research paper discusses how the traditional processes have given way to new web based, iterative, and incremental models which can cater to the new demands and requirements in the fast and quick changing world. The changes are expected in terms of look and feel, market demand, technological advances, and providing more efficient services.

		The changes in the client expectations and requirements are quick and frequent and hence requires new models which can be iterative, incremental like Agile, JAD, RAD, Extreme programming.	It discusses various models like Agile, JAD, RAD, extreme programming etc. These new models are capable of responding rapidly and meet stakeholders' expectations.
14.		This paper suggested that customisation and tailoring of regular project management methodologies is required to increase efficiency of processes in large scale complex software projects. The research provides guidelines and techniques for developing generic models for process improvement. The customised model would be more productive and having better analysing capabilities as project manager would be able to focus on the necessary requirements only.	This research emphasised that regular project management methods/ processes must be customised/ tailored in order to make it more efficient and avoid unnecessary details. The research found that customised models are better, more efficient and effective due to better alignment with current scenarios and objectives. This also provides project manager with better analytical skills to focus on the necessary details.
15.		This research suggested a simple and straightforward Project Matrix model for technical project management of software projects. The model requires no special training or resources.	The model is found to be efficient and highly effective in coordinating resources monitoring and controlling software projects. The paper demonstrated that model would be able to produce high quality software projects with ease.
16.		This paper proposed the integration of RUP and PMBOK for managing technical software development process of product lifecycle and management of project lifecycle respectively for efficient and effective management of projects and delivering high quality products.	The paper suggested that organisation would be able to automate various activities/ tasks in software development processes and project management with the integration of RUP and PMBOK methodologies. Using the capabilities of both RUP and PMBOK project manager would be able to manage projects more efficiently. This would in turn lead to better quality products.
17.		This research paper proposed the integration of standard software engineering practices with standard PM methodologies and framework to create a new framework for software development projects.	The results showed that new integrated framework would help to develop quality software project with better efficiency. Paper used standard software engineering framework from Spiral model, Waterfall model to create new framework integrated with PM methodology. This provided better planning, monitor and control, and execution of projects.
18.		This research paper proposed to make optimum utilisation of triple constraints of Time, Cost, and Scope as functions of project's high-level requirements/ business objectives. The new framework proposed extending the benefits of polarity management to triple constraints of PMBOK. Polarity management involves moving from focusing on one pole as the problem and the other as the solution (either/or thinking), to valuing both poles (both/and thinking). Good polarity management gets the best of both poles while avoiding the limits of either.	This paper suggested that integration of new framework which uses triple constraints as functions of business requirements/ objectives for optimum utilisation. It also suggested that polarity management could be used to manage the triple constraints of PMBOK more effectively and efficiently, since these constraints of time, cost, scope are opposite of each other.
19.		Government IT projects were analysed critically by this research. Authors tried to identify various challenges/ issues faced by e-government projects. The paper discussed the gaps in implementation of e-government projects, their monitoring and control as well as execution.	This research studied e-government projects and challenges faced by these projects. The study proposed that appraisal should be more systematic and periodic for successful completion. The appraisals have to be done in timely manner, risks and issues highlighted and thus controlling the projects efficiently.
20.		This research suggested the use of T-cube and Metromap visual tools for managing software development projects. These techniques use graphical and metaphor presentations to show various project management tasks. As the name suggests these techniques use Rubik Cube and Metro map metaphors.	The research proposed two techniques i.e. T-cube based on Rubik Cube and Metromap and tested them on real project data and found that these tools are effective and provide positive results for project management. The method provides graphical presentation which is easy to view/ track/ monitor the milestones and deliverables. This results in better project management with improved visibility.
21.		This paper suggested the application of software agent framework to help in managing various processes of software projects. Two comprehensive frameworks were developed to assist all the core and facilitating processes and functions of software project management.	The paper demonstrated that the use of software agent framework could significantly help in meeting market expectations and deliver projects as per the stakeholders' expectations adhering to schedule and budget. The framework assigns each agent to manage a small and manageable task, hence reducing complexity. The smaller tasks/ activities are easier to manage and thus providing better monitoring and control of project.
22.		This research discussed whether the existing project management methodologies like PRINCE2 had been useful in the management of software projects and information systems. Organisations have to develop full scale quality management	The study suggested that the project managers may have to use some other tools for effort and budget estimation like function point, empirical etc. and performance measurement techniques along with standard project management methodology. The research showed that organisations may have to use CMMI, or other

		plan to ensure delivering a good quality software product efficiently and effectively.	quality management plans to manage the projects in a better controlled manner.
	Managing Projects in R&D organisations	Differences between managing projects in R&D organisations and other projects had been examined in papers [23-24]. Since R&D projects are knowledge intensive, hence gives rise to different set of requirements and expectations form stakeholders.	
23.		This research discussed an approach that emphasises on building relationship between project management and engineering processes for better management and improved performance of research projects and programmes. This had been used at systems management office at NASA Langley Research Centre.	This approach suggested implementation of good project management practices and processes with consultation of teams from the early stages and improving them with tailor made training modules as and when required using just-in-time approach. This methodology also takes care of improvements in processes and policies. The approach also emphasised on conducting reviews and assessments independently for value addition.
24.		This research paper discussed and analysed the differences among R&D enterprises and other organisations. R&D enterprises being knowledge intensive, therefore more emphasis has to be on knowledge management. Therefore a culture of sharing knowledge by use of documentations, templates, and shared information systems has to be created.	This research confirmed that knowledge management system is needed to strengthen the R&D enterprise information system. The sharing of knowledge through shared workspace, intranet sites, documents, research papers, discussion and chat groups along with blogs would provide easier sharing of information and knowledge in the organisation.
	Empirical and Statistical Analysis	A number of empirical and statistical studies [25-33] had been conducted to investigate the success of various project management methodologies and frameworks. Data has been collected and analysed using various techniques in order to understand the effectiveness of diverse project management methodologies.	
25.		This study conducted a survey to find the experiences of people in project management. Survey finding were highlighted in terms of various methods, tools, techniques used and their effectiveness. Performance, project success factors and common criteria for managing successful projects were also studied.	The survey found that people are using different tools and techniques for similar kind of projects as per their skills, comfort level, and experience. The tools are also sometimes prescribed by the organisation, limited by legacy systems and availability. Some PMs have also tailored the standard methodologies to suit their needs. This provides them to model the system as per their needs and requirements and hence reduce the cost.
26.		This research studied various factors that influence the execution of project management and measure the performance of project management methodologies.	The study found that the 6 factors which have greater impact on execution of project management are Management commitment, financial constraints, organisational structure, reward system, education and training of project teams.
27.		This research explained the importance of scope management for successfully managing ICT projects. Standish group's CHAOS report states that only 1/3 of ICT projects are successful. Therefore project managers need to manage scope and changes more efficiently and effectively. This paper describes various approaches and techniques used by the trained scope managers in USA, Europe and Australia which have increased the success rate of the ICT project.	This paper suggested that with proper scope management process in place and with good training of PMs on scope and change management, ICT projects can be completed more successfully. Paper highlighted the use of standard methodologies like PMI, Prince-2, and other PPM tools have helped the project managers to a great extent in managing the scope and deliverables. It also showed that different methodologies are popular in different countries.
28.		This research conducted an empirical study to identify the important factors in the success of software process management in software projects. The study was to find out the importance and prioritisation of other factors like baselining, user involvement, management commitment, change management and documentation etc. for successful completion of projects.	Study found out that synchronisation/ synergy of activities/ tasks/ processes is more important than other processes for the successful management of software projects. The study also found out that other factors for process management in order of ranking were baselining, user involvement, management commitment, change management, and documentation.
29.		This study tried to understand the effect of change on the methods, practices, and performance of ICT projects. The paper discussed the impact and challenges due to the dynamic nature of work and environment in the current scenario. The research focussed on finding the reasons for failure, and reviewed the tools, techniques, practices, standards. The also tried to analyse if these failure are due to changing, dynamic or unstable nature of ICT projects and environment.	The results showed that complexity of the projects had increased due to the dynamic nature of projects and quick and fast changing market demand and requirements. The research highlighted the use of new tools, methods, frameworks, and methodologies like Agile, RAD and software as a service, helps to deliver quicker solutions at less cost.
30.		This paper discussed comparison had been done between modern and traditional project management. Author also discussed the key issues in the modern project management and challenges faced in current environment. It also explored the use or preference of one or other project management methodology by the project manager due to which there will be difficulty in managing projects which use different methodology or framework.	The paper suggested that in modern environment of ever changing requirements, new methods, tools and techniques are required to face the challenges and manage the projects effectively and efficiently. Study highlighted that client methodology could be different to the service provider and hence would create new challenges for project managers.
31.		This study explored the relation between project management assets and performance by considering project management assets as independent variables and project management performance as dependent variables. To analyse the	Seven factors were characterised for project management assets, three factors for organisational support and two factors for project management performance. The findings of the survey showed that project management assets give a competitive advantage to the organisations.

		advantage offered by project management assets, analysis of online survey by 198 PMI members was done.	
32.		This research studied the application of PMBOK 2008 standard processes to manage Enterprise Project Management (EPM) system in one of the organisations. The authors reviewed how the EPM project was implemented and its status based on the data collected by them. They proposed number of concepts to reduce the time and budget and enhance the system for better efficiency.	The study proposed that critical path management and PERT techniques could be used along with customisation of processes for the needs of the organisation. These methods help to track, monitor and control the projects in a better manner with increased visibility of the critical tasks, activities and parameters.
	Use of Alternative Methodologies	Some researchers [33-38] had used alternative methodologies and statistical techniques in the project management area. The investigations had used approach-avoidance theory, Fuzzy logic, NPV etc. for project management.	
33.		The paper described the use of approach-avoidance theory to develop integrated process model for managing escalation and de-escalation process in IT projects. Authors had used a process model to identify conditions, critical incidents, sequence of actions, and consequences over the life cycle of the project. The study had been done at various levels of project, organisation and environments to get an insight into the approach-avoidance decision of escalation or de-escalation and the stakeholders' response to the same.	The study suggested that approach avoidance theory may be useful in some cases, depending upon the critical path activities, incidents and when to escalate or de-escalate the decisions/ situations to the stakeholders. Various actions can be categorised into critical to low priority. This would give a view to the project managers for escalation or de-escalation of the decisions to the stakeholders.
34.		This research proposed a conceptual model to represent the actual project performance. Author had done the feasibility study by way of questionnaire and interview data. Paper highlights the management skills and factors required for managing the project performance and difficulties.	The factors of technical knowledge, estimating skills, critical path methods, and management skills are required in both the technical as well as management areas of cost, schedule etc.
35.		This study focused on the use of Fuzzy logic for critical path analysis and other PM activities. The approach also determines the significance of activity and path along with critical path. This is very useful in risk management and the method is able to manage the uncertainties.	The study suggested that the fuzzy logic would be quite useful in determining the activity, its significance along the critical path, and managing the risk more efficiently. The project manager would have better view of the critical path and risks and the actions they need to take for managing the projects efficiently.
36.		This research paper described that the project success and performance would be measured in terms of output benefits realised. Authors explained a technique to establish a cause and effect relation between output utilisation and target outcomes.	Paper presented a concept of managing project scope through utilisation map of the outputs. The fishbone or cause and effect diagrams, decision tree analysis are effective tools for mapping & managing the projects effectively.
37.		This study discussed the utilisation of Net Present Value (NPV) as a tool to better project management. Author highlighted the fact for successful monitoring and controlling of the project, NPV should be used. This can be the most important tool for finding the suitable solution.	The author demonstrated that NPV could be one of the most efficient tools for decision analysis and resolution for successful monitoring and control of the project. NPV can provide better budget control and managing the cost and in the process managing the schedule and scope efficiently.
38.		This study described the relationship of project management with other allied disciplines of in the field of management. The research is based on the study of 18 top management and business journals, publications and then divides them into 8 categories. The categories are (1) Strategy/Portfolio Management; (2) Operations Research/ Decision Sciences; (3) Organizational Behaviour/ Human Resources Management; (4) Information Technology/Information Systems; (5) Technology Applications/Innovation; (6) Performance Management/Earned Value Management; (7) Engineering and Construction; and (8) Quality Management/Six Sigma.	The study suggested that close relationship among various categories and disciplines. It requires good coordination among various processes and disciplines for successful completion of projects. PMs have to use somewhat different skills, techniques, tools to execute, monitor and control different category of projects. Therefore for web projects management technique is different form COTS projects, which is different from infrastructure projects and so on.
	Effects of Leadership and Management Qualities of Project manager	Effects of leadership and management qualities of project manager along with communication channels for managing projects had been explored by some researchers [39-42]. The importance of IQ, EQ, and MQ had also been studied. Teaching of proper project management tools and software is also stressed.	
39.		This research study discussed the impact of personality and behaviour of project manager on the project types and the success of the projects. A questionnaire was developed to evaluate the relationship of project manager's behaviour and personality with project type and success rate. The four types of projects considered were Urgent, Complex, Novel, and Normal and they are judged with three constraints cost, time and quality of projects.	The study showed that managers with team leading and management skills and who had better communication skills were more successful than others. These managers were found to be building an environment of trust and loyalty with good understanding of human needs. The motivational techniques are different for different level of people in the organisation and project managers must be able to address these issues and the strategy would have to change appropriately for different complexity of the projects.
40.		This research paper discussed various issues related to IT projects such as uncertain, unique, fast changing, short term, etc. The study emphasised that the project manager play a key role in successful delivery and implementation of IT projects.	Project managers must explore various communication channels with all stakeholders so that they can get and circulate proper support and information from all sides. This would ensure building long term relations and synergy with clients, project teams and various stakeholders.

41.		This paper studied the effects of leadership quality on the success of different type of projects. Authors studied the impacts of IQ, EQ, and MQ of project manager/ leader on the success of the projects with different level of complexities. Study used factor analysis and moderated hierarchical regression analysis to analyse various responses and data gathered. Authors also did variance and non-parametric tests to see the means and medians of EQ, IQ, MQ, complexity of faith, fact, and interaction.	Analysis showed that there is a relation between EQ, MQ and project success but are moderated differently by the complexity of the projects. EQ and Project success relationship is moderated by the complexity of faith, whereas MQ and project success relation is moderated by both complexity fact and faith. Project success is directly affected by the interaction and its complexity.
42.		This research paper discussed the issues of teaching project management as part of system analysis. Most of the training/ teaching programmes just put emphasis only on drawing Gantt chart or PERT chart for planning without using proper PM tools and skills. Therefore, the project plan could not give a full picture of the project and issues.	Authors suggest that proper tools, skills and techniques must be given to the aspirants along with proper assignments and tasks. PMs should use good project and/or portfolio management software for completing the assignments and tasks.
	Project Management in Global Distributed Environment	Managing projects in global distributed has its own challenges [43-66]. Researchers had explored use of different methodologies, techniques, tools for managing distributed projects from standard processes to incentive based approaches.	
43.		With the exponential growth of communication technologies and information systems, the globalisation of the commercial world has also increased significantly.	This research paper highlighted that in order to increase efficiency, productivity, quality and cost effectiveness, organisations are going for outsourcing and distributing their wok globally.
44.		This research study described the importance of software requirement specification (SRS) document to the success of global software projects. The authors discussed various difficulties in creating a standard SRS as companies have their own methods of creating such documents.	The authors studied how Capgemini overcame the issue of creating standard SRS by using specification patterns so as to create synergy among the global teams.
45.		The significance of knowledge sharing among global teams and stakeholders and how it can be addressed by mature processes and tools is highlighted in this study. There will be lesser readjustment required if the processes, methods and tools are used enterprise wide.	The authors proposed that enterprise wide software should be used for project assurance, quality and knowledge sharing. The software would help provide timely information, data and visibility for the preventive and corrective actions to be taken for better execution of the project.
46.		This study described the team structure for successful completion of offshore projects. The authors studied two types of structures for offshore teams and highlighted the problems faced by managers for changing the team structure and organisation model.	The paper proposed that changes have to be done to the existing structure for successful global operations. The team structures for managing offshore teams for various phases of the project and the reporting structure has to be managed keeping tin to account various time zone issues, cultural issues and skills availability.
47.		A framework for managing risks in global software projects is proposed in this research paper. The integrated framework had been created for distributed projects based on various parameters and requirements of global environment.	The framework proposed the use of various communication channels, different set of development environment for different needs/ requirements of the stakeholders and projects. The flow chart could also help to provide better information across the organisation.
48.		This research studied the impact of communication media like email, messaging, phone etc. on the conflict resolution in global teams. The authors tried to evaluate which could be the best sequence or combination of media tools for communication for resolving the conflicts.	The study showed how the cross cultural issues, different communication channels, time zone management had to be taken into account for managing global teams/ people effectively. The process for conflict management has to be robust and transparent so that the conflicts can be controlled/ resolved in an efficient manner.
49.		In this study, authors tried to analyse the global development projects using framework so as to overcome various issues in the distributed projects. The authors tried to study the processes used by various organisations to manage the distributed projects efficiently and effectively, and maximise the benefits of onshore-offshore delivery.	The paper showed different models and frameworks used by global organisations to manage the distributed projects successfully. Various activities can be distributed offshore/ near-shore or onshore and also the life cycle divided among them for maximising the benefits.
50.		This research studied different communicating media and its application the global agile software development projects.	The authors found that instant messaging is a good substitute tool for face to face communication and email is good tool for wider and enterprise wide information sharing.
51.		This research paper proposed predicting the outcome of global software development projects with the application of analytical modelling. The analytical models are parameterised to accommodate the single-site or multi-sites, team sizes, skills levels, expertise, availability, and support level etc.	The paper suggested various types of models for distributing various phases/ stages between offshore and onshore sites.
52.		This research study described the processes for managing a multi-site software development project is complex and requires a very good collaboration among teams.	The study suggested management of multi-site projects can be improved using networked virtual environment which allows for better communication, familiarity, sharing, mentoring, faith and faster resolution of conflicts.
53.		This research studied the growth of teams in distributed software development projects. The authors had tried to study the growth of teams in terms of expertise, communication skills, economic impact and working conditions.	The study described the communication channels, skills and the impact of virtual communication techniques for successful management of teams and projects in global environment.

			The better the economic and working condition, the better would be the team morale and more successful project management.
54.		This research paper explained that the “Distributed Work” is basically a number of different work provisions. Since the teams are distributed globally, and are separated by time zones, the managers have to rely heavily on the availability and efficiency of communications tools and information systems.	The research highlighted the importance of communication tools and information systems for successful management of global teams and projects.
55.		Use of incentive based theories to the distributed work environments is described in this research paper. The paper endeavours to address two subjects; firstly, to understand the effect of incentives on the worker’s choice for using distributed work environment, and secondly the collaboration of multiple incentives or disincentives across organisation, groups or individuals. This paper also looks into motives as to why people always prefer to take up distributed work environment. The theory of incentive is applied to two organisations to understand the behaviour and pattern.	The research suggested that people prefer distributed work environment because of flexibility, incentives, and availability. The disincentives are managing different time zones and culture. The study showed that incentives highly influence the working of people and opting for distributed working. It also highlighted that work life balance is one of the main criterion for people for remote/ home working.
56.		This research paper studied as why organisations choose for distributed work environment. The research was conducted to understand the use of distributed work environments in terms of costs, efficiency and productivity, motivation of employees, and impact on the group’s outputs.	The research suggested that the use of distributed work environments is to mainly reduce the costs, improve efficiency and productivity, motivate employees, and impacting the group outputs positively.
57-60		Even though there is clear impact on the employees for the work-life balance, more flexibility but there are conflicting observations made which are owing to more distractions at home which results in increased stress.	These papers showed that remote working, home working or flexible working is able to provide better work life balance but at the same time needs more planning as it could also lead to more distractions at home and less work. The employees have to manage themselves more efficiently to be more productive. Organisations provide hot-desk facilities to save on cost of space and also improve its travel carbon footprint.
61.		This research paper defined knowledge intensive firms as those that “offer to the market the use of fairly sophisticated knowledge or knowledge-based products”. Knowledge intensive firms can be divided into professional service, and research & development firms such as engineering and law firms or pharmaceutical companies. Knowledge intensive firms differ from other types of organisations through the organisation’s massive reliance on the intellectual skills of its employees to carry out its core functions.	Although many of the problems and barriers to distributed work are not unique to knowledge intensive firms, the sophisticated nature of the knowledge these firms typically deal in has the potential to magnify these problems. This report focuses on the interaction of individuals and teams within knowledge intensive firms and the ways that they interact and perform under distributed work arrangements.
62.		This research defined a virtual team as “groups of people employed in a shared task while geographically separated and reliant on electronic forms of communication”.	The research paper compared various factors such as telephonic conferences, video conferences, e-mails, time zones, and for managing virtual teams. The virtual communication tools are important and also people should be sensitive to the cultural communication styles and language used in communication to overcome misunderstandings and reduce communication gap.
63.		The paper defines the term remote resourcing as “carrying out work in an office remote from the point where a project is principally delivered”. The report defines remote resourcing when virtual communication tools are used and teams are distributed at one or more sites in different geographical locations.	These terms essentially describe interactions between people separated by physical distance who perform most of their work through communication technology. Within the body of this report the term distributed work is used to represent this concept. The dynamic changes to the project are handled more effectively when the team is at one place and long-term projects can get greater benefits from remote teams or by distributed working.
64.		The research paper discusses that distributed work covers many alternative methods of work which include satellite offices, flexible work arrangements, telecommuting and global collaborative teams.	The paper describes that distributed work could be defined in many different ways. The distributed teams could use different ways of working from flexible home working to offshore, onshore or near-shore arrangements. The paper highlighted that distributed teams and working are often used to reduce overall cost and improve services.
65.		This paper describes various issues and problems faced by distributed work faces which are similar to all the issues and problems that normal collocated group’s face, with the added complexity of workers being based at locations remote from each other, be it in the next room or in another country The inclusion of IT as a required element of many definitions reflects the importance of ICT as a replacement media to mimic the communicative and collaborative qualities inherent in collocated work groups.	This paper highlighted that distributed work faces many more problems in addition to the normal projects at one site. The projects and teams distributed in different locations brings in the importance of good communication media and skills, cross cultural issues and management, time zone management, and clear understanding of the stakeholders’ expectations. The project documentation has to be detailed and shared with all teams highlighting various milestones and deliverables and also giving details of communication requirements.
66.		This research explained that small and medium enterprises (SMEs) are also facing huge competition due to globalisation of economies and easier availability of cheaper and good	This paper highlighted that in order to stay ahead of the competition and technology SMEs should focus on to e-collaborations through project management approach. This will ensure them structured processes, better

		quality products, services across the world.	visibility for managing the full life cycle of the project and giving them better monitoring and control of project execution.
	Project Management Maturity Models	Various maturity models have been developed in the area of project management [67-76]. Maturity models help in assessing the capability and maturity of various processes, tools, techniques and management methodologies in an organisation.	
67.		This research studied the maturity levels of project management in different industries. They surveyed 126 organisations based on 42 components of maturity. The investigation also studied various industries and formed it into four groups i.e. professional, scientific, technical services; information; financial and insurance; and manufacturing.	The research found that maturity level is 2 out of 5 w.r.t. 36 components out of 42 analysed. The results showed that the maturity level of the four groups i.e. professional, scientific, technical services; information; financial and insurance; and manufacturing are similar across industries barring a few exceptions.
68.		The use of maturity models in improving project management practice is discussed in this research paper. The paper analyses the current maturity models for project management.	The analysis showed that the maturity models are mostly used reactively rather than proactively. The study also emphasised the fact that more empirical evidence and research is required to establish relationship of project performance with project management maturity models. The organisations that are using maturity models like CMMI are found to be in greater control of the projects since there processes are mature, well documented and provide good planning.
69.		This research paper presented a Project Management Process Maturity (PM)^2 model to evaluate and compare project management levels of organisations.	PM^2 model provides a well-structured model for evaluating project management maturity level of organisations. The model takes into account PM processes, factors and characteristics and shows the organisation's progress from functional to project driven organisation. The model helps organisations to enhance their project management approaches, create better documentation and plans.
70.		This research paper proposed a maturity model which has thee maturity levels with continuous improvement group of Key Process Areas (KPAs). The paper has taken ISO 9001:2000 as base for quality management. Each KPA is mapped onto plan-do-check-act (PDCA) cycle. Conical structure is developed for displaying the gradual development of KPAs in a better manner. KPAs are developed till they attain a dependable maturity level for project management. These KPAs may have to be improved continuously in order to respond to the changes.	By defining KPAs and then improving those continuously using PDCA cycle would enable the organisations to manage the projects more effectively and efficiently. By clear mapping of the KPAs to PDCA cycle, organisation would be able to improve the management and delivery of projects continuously. Thus organisations can work to optimise the processes for effective and efficient delivery of the projects.
72.		PMI, USA provided a very useful maturity model for organisations OPM3 to reflect their maturity with the best practices achieved within the project, portfolio, and programme domains.	OPM3 tries to link organisational strategy to successful, consistent, and predictable project completion. OPM3 divides the process improvement into four sequential stages i.e. Standardise, Measure, Control and continuously Improve.
73.		IEEE developed a standard for Software Project Management Plans (SPMP), IEEE std. 1058-1998. This applies to all types of software projects and all sizes.	This IEEE standard specifies format and contents of SPMP but does not specify technique to be used or examples. This standard provides information and document for managing a software project The standard defines the technical and managerial processes necessary to deliver the project requirements.
74-76		Capability Maturity Model Integration (CMMI) provides a framework for process improvements across the enterprises. The models have been used in many organisations since 1995 when the models were first created.	This framework gives applications of principles, practices, and objectives to achieve enterprise-wide process improvement. Many organisations have been able to manage budget and times efficiently with enhanced productivity and quality of projects using CMMI. Use of CMMI helps organisation to grow from Level1: Initial, Level2: Managed, Level3: Defined, Level 4: Quantitatively Managed to Level 5: Optimising. CMMI provides two representations: continuous and staged. The continuous representation allows the organisation to focus on the specific processes that are considered important for the organisation's immediate business objectives, or those to which the organization assigns a high degree of risks. The staged representation provides a standard sequence of improvements, and gives a basis for comparing the maturity of different projects and organisations. The staged representation also provides for an easy migration from the SW-CMM to CMMI

VII. Summary

Project management is a very complex field and encompasses technical skills along with man management skills to manage stakeholder expectations. Projects are influenced by both internal and external factors and require various communication channels and techniques to reduce the gap and deliver project effectively and efficiently.

The studies have been conducted to understand the existing standard models [4-8] and adapt them so as to manage projects in a better way. Modifications have been suggested so as to make existing methodologies work in different areas of work. The studied showed that different methodologies could be suitable for different types of projects. It has also been proposed that tailoring of methodologies would be more useful for different scenarios providing efficient and effective project management and control. The studies also showed that large complex system can would be better managed and controlled by dividing the system into smaller modules or phases.

System modelling, dynamic modelling, estimation models and object oriented modelling concepts have been investigated for improving the project management [9-12]. The studies showed that system modelling would be useful to have better communication among different stakeholders by managing various communication means, channels and modes. The studied showed that risks could be forecasted and managed more effectively by the use of dynamic modelling. The studies also showed that object oriented modelling concepts and properties like inheritance, modularity, data-encapsulation, relations could be used to describe the life cycle of the projects and manage it more efficiently. The studies proposed a mathematical model for estimating and forecasting along with the use of iterative development for better monitoring and control of the project.

Software and ICT projects have their own challenges due to rapid technological advances, providing better services, facilities etc. Therefore standard project management methodologies only may not be sufficient for managing project in this area. These methodologies may have to enhance and that's the reason that Agile, JAD, RAD, extreme programming have been developed [13-22]. The studies also showed that project management methodologies/ models should be customised/ tailored for the different projects so as to best fit the scenario and hence avoid unnecessary details and reduce cost and time. The studies also tried to combine different methodologies and framework like RUP, PMBOK, agent framework, metaphors and graphical presentations for creating new frameworks/ models to suit the requirements for managing, controlling, and execution of the project successfully within time and budget. The studies demonstrated that better estimates could be achieved and also stakeholders could manage in a better manner for overall success of the project. The studies also showed that quality control tools/ methodologies like ISO, CMMI would be highly useful in delivering good quality projects.

Stakeholders have different set of obligations, constraints, necessities, and expectations as projects in R&D organisations are knowledge intensive requiring more sharing of ideas. These differences have been highlighted in the research papers [23-24]. The studies showed that knowledge sharing areas/ sites for the researchers to share/ propose/ discuss their ideas by using blogs, discussion groups, and chat rooms would be highly useful to enhance creativity and innovations.

Researchers had been able to study various project management methodologies, tools, and techniques etc. empirically and statistically [25-33]. The studies showed that various factors like Management commitment, financial constraints, organisational structure, reward system, education and training of project teams play a crucial role in the successful management of projects. The project managers also tailor their methodology to suit the needs to reduce the cost and time for delivering the projects. The studies also proposed that scope and change management training would be highly useful to project managers for executing, monitoring, and controlling the projects successfully. The studies also showed that in new dynamic environment newer technologies would be more useful than traditional models of development and management of projects. Various techniques like critical path method, PERT, baselining, scope, change, and risk management would be of great help for successful delivery and management of the projects.

They had collected data from various projects to study different areas of project management and analysed it for understanding the effectiveness of various models, techniques or tools. Use of alternative statistical techniques and models such as Fuzzy Logic, NPV, approach-avoidance theory, had been explored in the project management area [33-38]. The studies showed that techniques like NPV, fuzzy logic would be highly useful in estimating and forecasting the project execution with better monitoring and control of the projects. The studies also showed that decision tree analysis, fishbone or cause-effect diagrams could be used effectively to manage quality, risks and deliver successful projects. Project managers need different skills/ tools/ techniques / methodologies to manage various projects efficiently.

Leadership abilities, communication and management abilities along with IQ, EQ, and MQ of the project manager have a huge impact on the projects [39-42]. Teaching of project management is not only about just teaching of GANTT and PERT chart but also must include various software and other tools for successfully managing projects. The studies have shown that project managers should use various communication tools/ methods with different stakeholders. Project managers should have the qualities to be a good team player, manager, having the ability to take decisions with responsibility to deliver successful projects. The studies showed that project manager should be able to manage the expectation of team members, motivating them, understand their training needs and communicate them at different levels so to manage various stakeholders, teams in a better manner and make them more productive.

Distributed environment of projects in the present multinational organisations gives rise to more complexities in all areas of project management [43-66]. Therefore standard project management methodologies have to be enhanced to meet diverse requirements from various stakeholders. The studies showed that distributed work environment has its own challenges and advantages. The challenges could be such as managing different time zones, cultural differences, virtual communication environments and costs associated with them, and many more. The advantages could be in terms of providing good quality projects at lower cost. This requires proper documentations, setup the correct expectations, managing various stakeholders and also managing the cross cultural issues effectively and efficiently. The conflict resolution criterion and transparent communication is the key to success in global scenarios and managing successful projects.

Maturity level is able to give the reliability of organisation in a particular area [67-76]. There are a number of maturity and capability levels for project management such as $(PM)^2$, OPM3, CMMI, IEEE, etc. against which an organisation can be appraised. The maturity models takes into account PM processes, factors and characteristics and shows the organisation's progress from functional to project driven organisation. The maturity models describe how mature the processes are, the success rate of the organisation in past, and also the probability that it would be able to deliver good quality projects. The organisation could use these models to build robust processes and also optimise them with feedback and changing scenarios and environment.

REFRENCES

[1] A Guide to the Project Management Body of Knowledge, 5th edition, PMI, USA, 2013

[2] Managing Successful Projects with PRINCE2, OGC, UK, 2009

[3] P3M3-Portfolio, Programme and Project Management Maturity Model, OGC, UK, 2008

[4] Ur Rehman, A., "Software Project Management Methodologies/ Frameworks Dynamics – A Comparative Approach", International Conference on Information and Emerging Technologies, ICIET 2007, Pakistan, 2007, pp 1-5

[5] Schweyer, B., "Formal Specifications For A Project Management Approach", IEEE International Conference on Systems, Man and Cybernetics", Canada, 1995, vol.2, pp 1170-1175

[6] Cederling, U.; Ekinge, R.; Lennartsson, B.; Taxen, L.; Wedlund, T., "A Project Management Model Based On Shared Understanding", Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Sweden, 2000

[7] Bailetti, A.J.; Callahan, J.R.; DiPietro, P., "A Coordination Structure Approach To The Management Of Projects", IEEE Transactions on Engineering Management, 1994, vol. 41, pp 394-403

[8] Milosevic,D.Z, "Standardizing Unstandardized Project Management", IEEE Technical Applications Conference, Northcon/96, USA, 1996, pp 12 – 17

[9] G. Rodrigues, Dr. Alexandre, "Managing and Modelling Project Risk Dynamics A System Dynamics-based Framework", 4th European PMI Conference, London, 2001, pp 1-7

[10] Silva, C. A.; Ferrão, Paulo, "A Systems Modelling Approach to Project Management: The Green Islands Project example", Second International Symposium on Engineering Systems, Massachusetts, 2009, pp 1-12

[11] Lin, Jyhjong; Yeh, Chunshou, "An Object-Oriented Formal Model For Software Project Management", Sixth Asia Pacific Software Engineering Conference, (APSEC '99), 1999, pp 552-559

[12] Ashman, R., "Project Estimation: A Simple Use-Case-Based Model", IT Professional, vol. 6 , Issue: 4, 2000, pp 40–44

[13] Walt Scacchi, "Process Models in Software Engineering", Encyclopaedia of Software Engineering, 2nd Edition, John Wiley and Sons, 2001

[14] Reddy, N.G., "Designing Software Project Management Models Based on Supply Chain Quality Assurance Practices", WRI World Congress on Computer Science and Information Engineering, USA, 2009, pp 659 – 663

[15] Shlaer, Sally; Grand, Diana; Mellor, Stephen J., "The Project Matrix: A Model for Software Engineering Project Management", 3rd Software Engineering Standards Application Workshop (SESAW III), 1984, USA, pp 1-10

[16] Callegari, D.A.; Bastos, R.M., "Project Management and Software Development Processes: Integrating RUP and PMBOK", International Conference on Systems Engineering and Modeling, ICSEM '07, 2007, Israel, pp 1-8

[17] Hewagamage, Champa; Hewagamage, K. P., "Redesigned Framework and Approach for IT Project Management", International Journal of Software Engineering and Its Applications, vol. 5 No. 3, July, 2011, pp 89-106

[18] Van Wyngaard, C.J.; Pretorius, H.C.; Pretorius, L., "Strategic Management Of The Triple Constraint Trade-Off Dynamics - A Polarity Management Approach", IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2011, pp 824 – 828

[19] Sarantis, D.; Askounis, D.; Smithson, S., "Critical appraisal on project management approaches in e-Government", 7th International Conference on ICT and Knowledge Engineering, 2009, pp 44-49

[20] Aguirregoitia, Amaia; Javier Dolado Cosín, José;, Presedo, Concepción, "Software Project Visualization Using Task Oriented Metaphors", Journal of Software Engineering & Applications, vol 3, 2010, pp 1015-1026

[21] C Nienaber, Rita; Barnard, Andries, "A Generic Agent Framework to Support the Various Software Project Management Processes", Interdisciplinary Journal of Information, Knowledge, and Management, vol. 2, 2007, pp 149-162

[22] Xu , Shuobo; Xu, Dishi, "Project Management Methodologies: Are They Sufficient To Develop Quality Software" 2nd IEEE International Conference on Emergency Management and Management Sciences (ICEMMS), China, 2011, pp 175–178

[23] Gilbert, M.G., "A Systems Management Approach To Improving Performance And Reducing Risks In Research Projects And Programs", IEEE Aerospace Conference Proceedings, vol. 7, 2002, pp 3467-3471

[24] Dingyong, Tang; Yizhen, Tao; Long, Jiang; Zheng, Cheng, "Application Research of Knowledge Management in R&D Enterprise Project Management", International Conference on Information Management, Innovation Management and Industrial Engineering, vol. 4, 2009, pp 447-452

[25] White, Diana; Fortune, Joyce, "Current Practice in Project Management–an Empirical Study", International Journal of Project Management, 2002, pp 1-11

[26] Chen, L.Y.; Cian Hui Kao, "The Effects Of Strategic Implementation Of Project Management And Performance", 2nd IEEE International Conference on Information Management and Engineering (ICIME), 2010, China, pp 194-198

[27] Dekkers, C.; Forselius, P., "Increase ICT Project Success with Concrete Scope Management", 33rd EUROMICRO Conference on Software Engineering and Advanced Applications, Germany, 2007, pp 385–392

[28] Berander, P.; Wohlin, C., "Identification of Key Factors in Software Process Management - A Case Study", International Symposium on Empirical Software Engineering, Italy, 2003, pp 316-325

[29] Othman, Marini; Mohd Zain, Abdullah; Razak Hamdan, Abdul, "A Review On Project Management And Issues Surrounding Dynamic Development Environment Of ICT Project: Formation Of Research Area", International Journal of Digital Content Technology and its Applications, vol. 4, no. 1, 2010, pp 96-105

[30] Attarzadeh, Iman, "Modern Project Management: Essential Skills and Techniques", Communications of the IBIMA, vol. 2, 2008, pp 1-9

[31] Jugdev, K.; Mathur, G.; Fung, Tak;, "Project Management Assets And Project Management Performance: Preliminary Findings", Technology Management in the Energy Smart World (PICMET), USA, 2011, pp 1-7

[32] Ghasemabadi, M.A.; Shamsabadi, P.D., "Application Of Five Processes Of Project Management Based On PMBOK-2008 Standard To Run EPM-2010 Project Management System: A Case Study Of Arya Hamrah

Samaneh Co.", 2nd IEEE International Conference on Emergency Management and Management Sciences (ICEMMS), 2011, China, pp 792–795

[33] Pan, G.; Pan, S.L.; Newman, M., "Managing Information Technology Project Escalation and De-Escalation: An Approach-Avoidance Perspective", IEEE Transactions on Engineering Management, vol. 56 , Issue: 1, 2009 , pp 76-94

[34] Deutsch, M.S., "An Exploratory Analysis Relating The Software Project Management Process To Project Success", IEEE Transactions on Engineering Management, vol. 38 , Issue: 4, pp 365-375

[35] Feng, Li; Junyin, Wei, "A Fuzzy Approach for the Project Management", International Conference on Wireless Communications, Networking and Mobile Computing, WiCom-2007, China, 2007, pp 5180-5183

[36] Zwikael, O.; Smyrk, J., "An Engineering Approach For Project Scoping", IEEE 18th International Conference on Industrial Engineering and Engineering Management (IE&EM), 2011, China, pp 2135-2137

[37] Wetekamp, W., "Net Present Value (NPV) As A Tool Supporting Effective Project Management", IEEE 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), Czech Republic, 2011, pp 898–900

[38] Hoon Kwak, Young; T. Anbari, Frank, "Analyzing Project Management Research: Perspectives From Top Management Journals", International Journal of Project Management, 2009, pp 435–446

[39] Hossein Fazel Bakhsheshi, Amir; Rashidi Nejad, Safoora, "Impact of Project Managers' Personalities on Project Success in Four Types of Project", 2nd International Conference on Construction and Project Management, IPEDR-2011, Singapore, vol.15, 2011, pp 181-186

[40] Liu, Shuangqin; Liu, Cheng, "Management Innovation of IT Project Managers", International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII), China, 2010, vol. 3, pp 62-65

[41] Muller, R.; Geraldi, J.; Turner, J.R., "Relationships Between Leadership and Success in Different Types of Project Complexities", IEEE Transactions on Engineering Management, vol. 59 Issue:1, 2012, pp 77 – 90

[42] Tatnall, A.; Shackleton, P, "IT Project Management: Developing On-Going Skills In The Management Of Software Development Projects", International Conference on Software Engineering: Education and Practice(SE:EP'96), USA, 1996, pp 400-405

[43] Armstrong, D.; Cole, P., "Managing Distances And Differences In Geographically Distributed Work Groups" in P. Hinds & S. Kiesler (ed.) Distributed work, MIT Press, 2002, pp. 167-186

[44] Salger, F.; Englert, J.; Engels, G., "Towards Specification Patterns for Global Software Development Projects - Experiences from the Industry", 7th International Conference on the Quality of Information and Communications Technology (QUATIC), Portugal, 2010 , pp 73-78

[45] Salger, F.; Sauer, S.; Engels, G.; Baumann, A., "Knowledge Transfer in Global Software Development - Leveraging Ontologies, Tools and Assessments", 5th IEEE International Conference Global Software Engineering (ICGSE), USA, 2010, pp 336-341

[46] Narayanan, Sidharth; Mazumder, Sumonta; R., Raju, "Success of Offshore Relationships: Engineering Team Structures", International Conference on Global Software Engineering, ICGSE'06, USA, 2006, pp 73- 82

[47] Persson, J.S.; Mathiassen, L.; Boeg, J.; Madsen, T.S.; Steinson, F., "Managing Risks in Distributed Software Projects: An Integrative Framework", IEEE Transactions on Engineering Management, vol. 56, Issue: 3, 2009 ,pp 508–532

[48] Khan, H.H.; Malik, N.; Usman, M.; Ikram, N., "Impact Of Changing Communication Media On Conflict Resolution In Distributed Software Development Projects", 5th Malaysian Conference in Software Engineering (MySEC), Malaysia, 2011, pp 189-194

[49] Lane, M.T.; Agerfalk, P.J., "Experiences in Global Software Development - A Framework-Based Analysis of Distributed Product Development Projects", 4th IEEE International Conference on Global Software Engineering, ICGSE, Ireland, 2009, pp 244 – 248

[50] Niinimaki, T., "Face-to-Face, Email and Instant Messaging in Distributed Agile Software Development Project", 6th IEEE International Conference on Global Software Engineering Workshop (ICGSEW), Finland, 2011, pp 78 - 84

[51] Czekster, R.M.; Fernandes, P.; Sales, A.; Webber, T., "Analytical Modeling of Software Development Teams in Globally Distributed Projects", 5th IEEE International Conference on Global Software Engineering (ICGSE), Ireland, 2010, pp 287–296

[52] Bartholomew, R., "Globally Distributed Software Development Using An Immersive Virtual Environment", IEEE International Conference on Electro/Information Technology, EIT, USA, 2008, pp 355-360

[53] Hashmi, J.; Ehsan, N.; Mirza, E.; Ishaque, A.; Akhtar, A., "Comparative Analysis Of Teams' Growth In Offshore And Onshore Software Development Projects", IEEE International Conference on Management of Innovation and Technology (ICMIT), Singapore, 2010, pp 1163–1167

[54] Hinds, P.J.; Bailey, D.E., "Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams", Organization Science, 2003, vol. 14 (6), pp 615-632

[55] Swan, Bret; Belanger, France; Beth Watson-Manheim, Mary, "Theoretical Foundations for Distributed Work: Multilevel, Incentive Theories to Address Current Dilemmas", IEEE 37th Hawaii International Conference on System Sciences, Hawaii, vol. 1/04, 2004, pp 1-10.

[56] Bailey, D. E.; Kurland, N. B., "A Review Of Telework Research: Findings, New Directions, And Lessons For The Study Of Modern Work", Journal of Organizational Behavior, vol.23, 2002, pp 383-400

[57] Pinsonneault, A.; Boisvert, M., "The Impacts of Telecommuting on Organizations and Individuals: A Review of the Literature", in "Telecommuting and Virtual Offices: Issues and Opportunities", Johnson, N.J., London: Idea Group Publishing, 2001, pp 163-185

[58] Pearlson, K.E.; Saunders, C.S., "There's No Place Like Home: Managing Telecommuting Paradoxes", Academy of Management Executive, vol. 15, 2001, pp 117-128

[59] Belanger, France; Beth Watson-Manheim, Mary; Jordan, D.H., "Aligning IS Research and Practice: A Research Agenda for Virtual Work," Information Resources Management Journal, vol. 15, 2002, pp. 48-70

[60] Igbaria, M., "The Driving Forces in the Virtual Society," Communications of the ACM, vol. 42, 1999, pp 64-70

[61] Alveson, M, "Knowledge Work and Knowledge-Intensive Firms". Oxford University Press, New York, 2004

[62] Hornett, A., "The Impact of External Factors on Virtual Teams: Comparative Cases", in Pauleen, J. (ed.), "Virtual Teams: Projects, Protocols and Processes", Idea Group Publishing, UK, 2004

[63] Turkington, D., "Remote Resourcing", The Beca Infrastructure Board, Auckland 2004

[64] Bélanger, F.; Collins, R.W., "Distributed Work Arrangements: A Research Framework", The Information Society, vol 14, 1998, pp 137-152

[65] Cramton, C.D., "Attribution in Distributed Work Groups", in Hinds, P.J.; Kiesler, S. (ed.) "Distributed Work", MIT Press. London, England, 2002, pp 191-212

[66] Mohammad Jafari, M.; Ahmed, S.; Dawal, S.Z.M.; Zayandehroodi, H., "The Importance Of E-Collaboration In SMES By Project Management Approach A Review", 2nd International Congress on Engineering Education (ICEED), Malaysia, 2010, pp 100–105

[67] Grant, K.P.; Pennypacker, J.S., "Project Management Maturity: An Assessment Of Project Management Capabilities Among And Between Selected Industries", IEEE Transactions on Engineering Management, vol. 53 , 2006, pp 59-68

[68] Brookes, Naomi; Clark , Robin, "Using Maturity Models to Improve Project Management Practice", POMS 20th annual Conference, USA, 2009

[69] Hoon Kwak, Young; Ibbs, C. William , "Project Management Process Maturity $(PM)^2$ Model", Journal of Management In Engineering, 2002, pp 150- 155

[70] Sukhoo, Aneerav; Barnard, Andries; M.Eloff, Mariki; A. Van der Poll, John, "An Evolutionary Software Project Management Maturity Model for Mauritius", Interdisciplinary Journal of Information, Knowledge, and Management, vol. 2, 2007, pp 99-118

[71] EPA Guidance for Quality Assurance Project Plans for Modelling, Office of Environmental Information, EPA QA/G-5M, USA Environment Protection Agency, 2002

[72] Organisational Project Management Maturity Model (OPM3), 2nd edition, Project Management Institute, USA, 2008

[73] IEEE standard for Software Project Management Plans, IEEE std 1058-1998, IEEE, USA

[74] CMMI for Development, Version v1.3, Software Engineering Institute Process Management Program, 2010, Carnegie Mellon University, USA, 2010

[75] CMMI for Acquisition v1.3, Software Engineering Institute Process Management Program, 2010, Carnegie Mellon University, USA, 2010

[76] CMMI for Services v1.3, Software Engineering Institute Process Management Program, 2010, Carnegie Mellon University, USA, 2010

E-mail use by the faculty members, students and staff of Saudi Arabian and Gulf states Universities

Fahad Alturise Paul R. Calder Brett Wilkinson
School of Computer Science, Engineering and Mathematics
Flinders University
Adelaide, Australia

***Abstract*— Electronic mail systems (Email) constitute one of the most important communication and business tools that people employ. Email in the workplace can help a business improve its productivity. Many organisations now rely on email to manage internal communications as well as other communication and business processes and procedures. This paper compares email use by university stakeholders (i.e. faculty members, staff and students) between Saudi Arabia on one hand, and the Gulf States - Qatar, Oman, United Arab Emirates (UAE) and Bahrain – on the other. A questionnaire that was expert-reviewed and pilot-tested, was used to collect data from ten universities in Saudi Arabia and five universities in the Gulf States. Slight differences emerged in the Saudi Arabia and Gulf States universities' stakeholders' use of email in terms of having email, frequency of checking email, and skills in using email. The Saudi Arabian universities must improve their IT infrastructure, including the provision of suitable connection networks and formal training of staff in utilising IT resources. This study's findings aim to advise the Saudi Arabian and Gulf States' universities on their plans and programmes for e-learning and the consolidation of required resources.**

Keywords*— *Email; Saudi Arabia; Gulf States

I. Introduction

Email systems constitute the main communication method of electronic learning and are emerging as a strategic business tool. However, electronic learning applications place extra security risks on organisations and businesses due to the problems associated with direct electronic interaction with other entities (Duane and Finnegan 2004). Email systems have traditionally been initiated by IT departments or sections without being part of a business-led strategy. This is despite email having evolved over time to become more of a corporate-wide service (Jackson, Dawson et al. 2001). Through the evolution of email systems, their strategic importance has increased but the benefits of email do not accrue automatically (Stevens and McElhill 2000). Along with increases in electronic business activity and the use of email systems, an increase in employee abuse of email technology has been documented. Virus infection arising from email use and deliberate abuse of email facilities are the leading causes of security breaches, suggesting poor controls and overarching policies that govern email use are to blame (Jackson, Dawson et al. 2003).

Problems emanating from email systems have become critical as technological advances are made towards inter-organisational networking. As organisations struggle to derive value from information technologies, particularly in periods of reduced IT budgets (Jackson, Dawson et al. 2002), organisations spend and often waste money buying technology that does not suit the human infrastructure, policies and procedures so that abuses can be prevented or curbed (Burgess, Jackson et al. 2005). For an organisation to shift its focus from operating as a traditional business to an electronic one it must define its practices, procedures and processes so they can be monitored, analysed, and regulated (Jackson, Burgess et al. 2006). Such issues, consequently, inhibit significantly the growth of electronic business. It is imperative that organisations formulate coordinated and comprehensive responses to email systems management. Specifically, businesses and institutions should anticipate the harmful effects of email system abuses or cyber-hacking to prevent them from occurring (Jackson, Dawson et al. 2003).

While the nature and scope of information systems threats have been well documented in the past, there are no real practical measures to stopping or controlling such dangers. Organisations lack analytical tools to examine their practices or to ensure email systems are used only for corporate reasons and not people's preferences (Duane and Finnegan 2004). This paper presents the findings a study that investigated stakeholders in universities in the Gulf States and Saudi Arabia and how they used email systems. The next section examines the frequency of using email generally and in university email systems particularly. This is followed by a discussion of the research method, and a presentation of the research findings. The paper concludes by identifying key factors regarding issues that may influence email usage in Saudi Arabia's universities.

II. Literature review

Even though email usage has been examined for more than twenty years, the differences of email usage between the Saudi Arabian and Gulf States' universities' stakeholders have not been investigated. Current practices in email use re often studied to identify design implications for improving email, but to date studies have not accounted for potential differences among Gulf States users. It can be assumed that the users of an email application are the major source of problems, as they create and receive the emails that periodically create problems of security, privacy/confidentiality and professionalism. It is essential to identify the major problems users face with email and then administer training schemes on how to become a more effective and responsible email communicator (Jackson, Burgess et al. 2006).

Research conducted over the past 10 years has focused on the overwhelming nature of electronic mail communication, where stress in the workplace, negative social behaviours, and diminished productivity among knowledge workers have been documented (Jackson, Dawson et al. 2002; Burgess, Jackson et al. 2004; Ducheneaut and Watts 2005; Neustaedter, Brush et al. 2005). Many studies have suggested the positive and negative effects of using email and how this tool helps or impacts on productivity. Duane and Finnegan (2004) reported the negative effects of email use, for example security infractions, productivity drain, non-business communication use, increased cost of usage, profanities, bad news, and illicit use. They reported some solutions to overcome these problems such as proper policy and its practical implementation in the workplace.

Smith (2008) reported that the average employee spends between 90 minutes and two hours per day reading email messages. As the email inbox becomes cluttered with retained emails, incoming messages, irrelevant chain mail, and spam, workers may become victims of email overload or, at a minimum, face rising frustration attempting to manage electronic communication in a disciplined way (Jackson, Dawson et al. 2003, Jackson, Dawson et al. 2003). Many studies on the subject of email overload have emerged in the U.K. (Jackson, Dawson et al. 2003; Jackson, Dawson et al. 2003; Burgess, Jackson et al. 2005), and this subject has now generated more interest in more recent times. It is reported that the average corporate email user sends and receives approximately 156 messages per day, “and this number is expected to grow to about 233 messages a day by 2012” (Radicati Group, Inc., 2008, p. 4). One outcome of the explosive growth in email volume is that businesses and institutions increasingly face key decisions on how to address the email monster (Dudman, 2003). At Loughborough University in England, Jackson conducted a series of research projects (Jackson et al., 2002; 2003a, 2003b; Jackson, Burgess, & Edwards, 2006) examining email tolerance levels, cost of email to organizations, and reduction of email defects through training of workers.

Issues identified in the studies included the following: poorly written email, email as a distraction, email used improperly (i.e., when face-to-face was warranted), and email carbon copy abuse. Jackson et al. (2003a) found that 65% of emails sent to recipients failed to provide enough information for the receiver to respond appropriately. Similarly, email messages may not provide the reader with enough information to accurately determine the context or tone of what was being sent (Whittaker, Bellotti et al. 2006). Consequently, the recipient faces additional pressure and frustration attempting to resolve the communication (Burgess, Jackson, & Edwards, 2004). Other studies show the importance of email in education and how people can benefit by it such as increased productivity, social interaction and well-being (Chase and Clegg 2011). Some analyses reported the challenges affecting the utilisation of email in education institutions.

III. Objectives and methodology of the study

The overall objective of this study was to investigate the problems inherent in ICT systems in Saudi Arabian and Gulf States’ universities and try to suggest solutions. One of the most important factors is using communication services such as email among faculty members, administrative staff and students at various universities in these institutions. This study employed a survey to elicit answers to key research questions. A perusal of earlier studies on ICT infrastructure indicates that the questionnaire-based survey has been the most popular method used. Some studies have included open-ended questions (Husain 2001) and selective interviews (Fusayil 2000) to obtain additional data. Most questionnaires have been paper-based enterprises with a few that were web-based (Chu 2002; ur Rehman & Ramzy 2004; Al-Ansari 2006). For this study it was decided to use a paper-based questionnaire because many respondents of this study will not be able to manage a web-based instrument.

Several studies and a few available questionnaires served as the basis for developing a questionnaire with closed-ended questions. It was divided into five parts; two of them are relevant to this paper. Part I contained questions concerning the demographic characteristics of respondents and Part II consisted of questions concerning university computers and internet, and frequency of using several IT services and their effectiveness. The questionnaire, prepared by the researcher, was reviewed by the supervisor and co-supervisor. Also, it reviewed by a statistical consultant and an expert in English and Arabic languages who can ensure clarity, proper language structure, and elimination of language ambiguities. It was pilot-tested on 18 students resulting in minor modifications.

An email was sent to 10 universities in Saudi Arabia and 5 universities in the Gulf States, i.e. Oman, Bahrain, Qatar, and UAE. The number of participants was 142 faculty members, 121 staff/administration members and 511 students. A package consisting of information sheets for participants and a questionnaire in the Arabic language was devised for this study.

IV. Findings and discussion

A. *Non-users of the Email*

One interesting finding is that, given the prevalence of IT in education and when the university is planning to enter e-learning in a big way, 86 (11.11%) of the 774 respondents neither used nor had an email account. This situation is predominant in the Saudi Arabian universities (94.18%). Of this total, (12.34%) were faculty members and (20.98%) were administrative staff. The rest were students and of these only (5.81%) were from the Gulf States universities. They did not give any reasons for not using email. It seems that they were educated either at a time or in an environment where IT was non-existent and consequently had little or need for any IT in their working or everyday lives. The following sections present the data for this study's questionnaires.

B. Demographic Characteristics of the Respondents

Out of the 142 faculty members, 107 (75.35%) worked in Saudi universities and 35 (24.64%) were from the Gulf States universities. 106 (74.6%) were male and 36 (25.4%) were female. Most of them have PhD (50.7%) and were 45 years old and over (25.4%) or between 35 and 44 (46.5%).

Out of the 121 respondents of administrative staff, 95 (78.51%) were from Saudi universities and 26 (21.48%) were from the Gulf States universities. 72 (59.5%) were male and 49 (40.5%) were female. Most of them have Bachelor degrees (54.5%) and (19.8%) have high school or less (54.5%). Age-wise most were between 25 to 34 years old (54.5%) and (9.1%) were 45 years old and over. More than half (57.9%) have working experience at university between 1 to 5 years although a few (4.1%) have more than 20 years.

Out of the 511 students, 388 (75.92%) were from Saudi universities and 123 (24.07%) were from Gulf States universities. In terms of gender 347 (64.7%) were male and 189 (35.3%) were female. Most of them were in their first year of study (34.1%) and (28.7%) were in the second year, with 19.6% in the third year. (80%) of them live with their family and most of them (89%) were aged between 18 and 24.

C. Email Habits of the Respondents

TABLE I. UNIVERSITY STAKEHOLDERS WITH ACCESS TO EMAIL

University Region	**Email**			
	No		***Yes***	
	Count	***Percent***	***Count***	***Percent***
	Faculty members			
Saudi Universities	10	9.30%	97	90.70%
Gulf States Universities	0	0.00%	35	100.00%
	Administrative staff			
Saudi Universities	17	17.90%	78	82.10%
Gulf States Universities	0	0.00%	26	100.00%
	Students			
Saudi Universities	54	13.90%	334	86.10%
Gulf States Universities	5	4.10%	118	95.90%

Table I shows the percentage of university stakeholders who have email accounts either at home or university. About (9%) of faculty members in Saudi universities do not have an email address while all employees in the Gulf States universities have email addresses. Approximately (18%) of administrative staff in Saudi universities do not have email address while all staff employees in the Gulf States universities have email addresses. Finally, about (14%) of students in Saudi universities do not have email addresses compared to about (4%) in the Gulf States universities not having an email address. The table below shows us that email accounts are more common in the Gulf States universities compared to Saudi Arabia, and this is due to a variety of factors such as familiarity with technology, easy access to IT services and effective role for IT in their day life.

D. Frequency of Access Email of the Respondents

Table II shows the frequency of use of emails by stakeholders for different reasons. About (8%) of Saudi Arabian faculty members never use email or only their check emails once a month or less. About (10%) of Saudi Arabia faculty members check email once a week compared to about (3%) of Gulf States faculty members. About (31%) of Saudi Arabia and Gulf States faculty members check their email once a day. Finally, (52%) of Saudi Arabian university faculty members use and check email several times per day compared to (66%) for the Gulf States members. The data indicates that faculty members in the Gulf States universities use email more frequently for their daily activities. This use helps them to understand the benefits of email and how it can be used for learning.

The table also shows that, about (17%) of Saudi Arabia universities' administrative staff never use email or only check emails once a month or less. About (32%) of Saudi Arabia administrative staff check emails once a day compared (8%) of Gulf States universities' administrative staff. Finally, (30%) of Saudi Arabia universities' administrative staff use and check email several times per day compared to (92%) for the Gulf States faculty members. This table show how administrative staff in the Gulf States institutions use email more frequently in their daily duties. They understand the benefits of email and how it assists them in their university work.

Finally, about (46%) of Saudi Arabian university students never use email or only check their emails once a month or less compared to (26%) in the Gulf States universities. About (19%) of Saudi Arabian students check their emails once a week compared to (14%) of the Gulf States students. Finally, about (23%) of Saudi Arabian students use and check email several times per day compared to (46%) of the Gulf States students. The data suggests that students in the Gulf States universities use email more frequently than Saudi students. This use also helps them to understand the benefits of email and its applicability to their learning. This use of email could be due to faculty encouragement or the university system which compels to understand emails in everyday transactions.

E. Having and Frequency of Access University Email of the Respondents

Table III shows the number of stakeholders who have access to university email system and their frequency of use. Five percent of Saudi Arabia university faculty members state that the email system is not available and if it is available (9%) of them never use it. Furthermore (17%) of Saudi Arabian faculty members use email university compared to (3%) of faculty members in the Gulf States. In Saudi Arabia (24%) of faculty members use email system in university occasionally to (6%) of faculty members in the Gulf States universities. Finally, (46%) of Saudi Arabian faculty members use university email system frequently compared it to (92%) of faculty members in the Gulf States.

TABLE II. UNIVERSITY STAKEHOLDERS' FREQUENCY OF CHECK EMAIL

University Region	Email											
	Never		*Less than once a month*		*About once a month*		*About once a week*		*About once a day*		*Several times per day*	
	Count	*Percent*	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*
	Faculty members											
Saudi Universities	2	1.90%	3	2.80%	3	2.80%	10	9.30%	33	30.80%	56	52.30%
Gulf States Universities	0	0.00%	0	0.00%	0	0.00%	1	2.90%	11	31.40%	23	65.70%
	Administrative staff											
Saudi Universities	5	5.30%	4	4.20%	7	7.40%	20	21.10%	30	31.60%	29	30.50%
Gulf States Universities	0	0.00%	0	0.00%	0	0.00%	0	0.00%	2	7.70%	24	92.30%
	Students											
Saudi Universities	79	20.40%	40	10.30%	59	15.20%	72	18.60%	50	12.90%	88	22.70%
Gulf States Universities	12	9.80%	5	4.10%	15	12.20%	17	13.80%	18	14.60%	56	45.50%

These statistics indicate that faculty members in the Gulf States universities use the email system more frequently and this may be due to more proactive university IT policies, support from decision-makers and people simply being used to working with email systems. The Saudi Arabian universities' problems stem from the lack of IT services, lack of infrastructure, lack of motivation, lack of skills and others.

According to (10%) of Saudi Arabian universities' administrative staff the email system is not available. Another (5%) have a university email account but never use the email system. Moreover, (23%) of Saudi Arabian universities' administrative staff use email university rarely while (25%) use email university occasionally. Furthermore (37%) of Saudi Arabian universities' administrative staff use university email system frequently compared to all (100%) of Gulf States universities' staff using email frequently. This complete participation could be due to better policing of the university email system and the fact that Gulf States university staff are comfortable in contacting stakeholders by email.

Finally, (11%) of Saudi Arabian universities' students point out the university email system is not available and if it is available (14%) of them have never used it. This contrasts with less than (1%) in Gulf state countries universities do not have this service and about (3%) who never use if it is available. Moreover, (25%) of students in Saudi Arabian universities' use email university rarely compared to (7%) of Gulf States universities' students. Twenty percent of students in Saudi Arabian universities use the email system occasionally compared to (12%) in the Gulf States institutions. Furthermore (30%) of students in Saudi Arabian universities' use the email system frequently compared to (77%) for the Gulf state universities' students.

TABLE III. UNIVERSITY STAKEHOLDERS' USING UNIVERSITY EMAIL

University Region	University email system											
	N/A		*Never*		*Very Rare*		*Rare*		*Occasional*		*Frequent*	
	Count	*Percent*	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*	*Count*	*Percent*
	Faculty members											
Saudi Universities	5	4.70%	9	8.40%	2	1.90%	16	15.00%	26	24.30%	49	45.80%
Gulf States Universities	0	0.00%	0	0.00%	1	2.90%	0	0.00%	2	5.70%	32	91.40%
	Administrative staff											
Saudi Universities	9	9.50%	5	5.30%	9	9.50%	13	13.70%	24	25.30%	35	36.80%
Gulf States Universities	0	0.00%	0	0.00%	0	0.00%	0	0.00%	0	0.00%	26	100.00%
	Students											
Saudi Universities	40	10.30%	54	13.90%	33	8.50%	68	17.50%	76	19.60%	117	30.20%
Gulf States Universities	1	0.80%	4	3.30%	4	3.30%	5	4.10%	15	12.20%	94	76.40%

TABLE IV. UNIVERSITY STAKEHOLDERS' EMAIL SYSTEM SKILLS

University Region	**Email**									
	Non User		***Beginner***		***Moderate***		***Competent***		***Expert***	
	Count	***Percent***	***Count***	***Percent***	***Count***	***Percent***	***Count***	***Percent***	***Count***	***Percent***
	Faculty members									
Saudi Universities	0	0.00%	1	0.90%	9	8.40%	30	28.00%	67	62.60%
Gulf States Universities	1	2.90%	0	0.00%	0	0.00%	9	25.70%	25	71.40%
	Administrative staff									
Saudi Universities	1	1.10%	4	4.20%	11	11.60%	36	37.90%	43	45.30%
Gulf States Universities	0	0.00%	1	3.80%	0	0.00%	6	23.10%	19	73.10%
	Students									
Saudi Universities	27	7.00%	23	5.90%	67	17.30%	116	29.90%	155	39.90%
Gulf States Universities	2	1.60%	3	2.40%	13	10.60%	36	29.30%	69	56.10%

F. Skills of Using Email of the Respondents

Table IV summarises the skills of using email from the university stakeholders' point of view. In Saudi Arabia (1%) of faculty members are at the beginner or learning stage compared to (3%) in the Gulf States universities. It is evident that (8%) of Saudi Arabian faculty members have a moderate level of skills of using email.

Also, (26%) of both Saudi Arabian and Gulf States faculty members have a competent level of skills of using email. Finally, (63%) of Saudi Arabia faculty members are expert in using email compare to (72%) of faculty members in the Gulf States. It is evident that faculty members in Saudi Arabia and Gulf States universities have the same or similar level of skills in using email. However, this still means that they need training courses to motivate them and educate them in improving email skills. On the other hand, (5%) of Saudi Arabian administrative staff point out that their skills are at the beginner stage or less compared to (4%) in the Gulf States universities. For the Saudi Arabian administrative staff, (12%) have moderate level of skills of using email, and (38%) of Saudi Arabia faculty members have a competent skill level compared to (23%) of administrative staff in the Gulf States. Finally, (45%) of Saudi Arabian administrative staff are expert in using email compare to (73%) of administrative staff in the Gulf States.

These statistics reveal that administrative staff in Saudi Arabian institutions have less skills or expertise than administrative staff in the Gulf States. This means that Saudi Arabian university administrative staff require good quality training courses and support to overcome difficulties and to improve how they employ email for designated tasks. Finally, (13%) of Saudi Arabia students point out that their skills in using email are at the beginner level or less compared to (4%) in the Gulf States universities. Apparently, (17%) of Saudi Arabian students have a moderate level of skills of using email compared to (11%) of students in the Gulf States. In both Saudi Arabia and the Gulf Staters, (29%) of students have a competent level of skills when using email. It emerges that (40%) of Saudi Arabian students are expert in using email compared to (56%) of students in the Gulf States. This indicates that Saudi Arabian students have poorer skills compared to those in the Gulf States universities. They need more encouragement from faculty members to communicate via email and it is apparent that Saudi universities must overcome this dilemma. Faculty members have to help students facilitate the email system at the beginning of each semester.

Saudi Arabian university stakeholders participate less in email compared to their counterparts in the Gulf States, and are less skilled as well. In particular the administrative staff's use of email is lower in Saudi Arabia because it has not been actively promoted or implemented. It is paramount that the decision-makers and senior executives in Saudi Arabia's universities lead the way in changing how staff work with modern electronic technologies and Internet systems that incorporate email services.

V. CONCLUSIONS AND RECOMMENDATIONS

The analysis presented above shows that there is a difference in email usage among faculty members, administrative staff and students of Saudi Arabia and Gulf States universities. However, if the results of this study are viewed in light of the universities' recent interest in promoting e-learning, then the level of interest, nature of use and ability of these respondents seem to fall short of the requirements of developing e-learning via using email services.

The respondents' dependence on learning by themselves how to use email indicates a deficiency in formal training opportunities. A greater use of email by these respondents for personal objectives, be they communication, research or writing, and little interest shown in using it for teaching and classroom work points to another critical gap that contradicts or undermines the university's plans to emphasise IT's place in teaching and learning activities.

An encouraging sign in the findings is the awareness shown by participants of the usefulness of the Internet and its resources for academic work and for identifying concrete problems they face while using these resources. Participants are mindful of the need to upgrade their Internet expertise and experience in online searching.

The observations made above have serious implications for the universities' future academic learning and development plans. These demand immediate and serious attention in the following areas:

- Identification and preparation of plans for upgrading, as soon as possible, the IT infrastructure including libraries, to bring them to a level compatible with the requirements of intensive IT-based teaching, learning and research.

- Conducting a "training needs analysis" which will identify gaps in IT use skills among the faculty, staff, and students.

- Developing formal and differential training packages based on the results of the training needs analysis to improve IT competencies.

- University management must improve their IT applications, staff IT competence and qualification, and remotely available information resources, and focused formal training programmes essential for these resources and services.

The results of this survey point to some issues on which further research is required. There is a need to measure email use skills of faculty, staff and students in a more concrete manner so that differentiated training packages can be prepared for various groups. In other words, a training needs analysis of various segments of the academic population should be done. Prior to that, students' use of the Internet needs to be analysed. The training programs offered by the universities should be analysed in terms of their effectiveness. There is also a need to investigate the level of information evaluation skills of email users and create an awareness of the importance of this activity in selecting and using internet resources for teaching, learning and research. Also, university decision-makers have to encourage and motivate faculty members and staff to increase their use of IT services and email systems. They can do this by establishing training courses, reducing the teaching load, increase salaries and other methods.

ACKNOWLEDGMENT

This research is supported by Saudi Arabia Cultural Mission and Flinders University.

REFERENCES

[1] Al-Ansari, H. (2006). "Internet use by the faculty members of Kuwait University." Electronic Library, The **24**(6): 791-803.

[2] Burgess, A., et al. (2004). "The effectiveness of training in reducing email defects."

[3] Burgess, A., et al. (2005). "Email training significantly reduces email defects." International Journal of Information Management **25**(1): 71-83.

[4] Chase, N. M. and B. Clegg (2011). "Effects of Email Utilization on Higher Education Professionals." International Journal of Technology and Human Interaction (IJTHI) **7**(4): 31-45.

[5] Chu, Y.-h. (2002). Factors related to adoption of internet resources in instruction by faculty at the Pennsylvania State University, Pennsylvania State University.

[6] Duane, A. and P. Finnegan (2004). Managing email usage: A cross case analysis of experiences with electronic monitoring and control. Proceedings of the 6th international conference on Electronic commerce, ACM.

[7] Ducheneaut, N. and L. A. Watts (2005). "In search of coherence: a review of e-mail research." Human–Computer Interaction **20**(1-2): 11-48.

[8] Fusayil, A. (2000). The adoption of the internet by faculty members at Ohio University, Ohio University, June.

[9] Husain, S. P. R. (2001). "Adoption of the internet as a teaching and learning tool: patterns of use, motivators and barriers among outstanding faculty in community colleges."

[10] Jackson, T., et al. (2001). "The cost of email interruption." Journal of Systems and Information Technology **5**(1): 81-92.

[11] Jackson, T., et al. (2003). "Reducing the effect of email interruptions on employees." International Journal of Information Management **23**(1): 55-65.

[12] Jackson, T. W., et al. (2006). "A simple approach to improving email communication." Communications of the ACM **49**(6): 107-109.

[13] Jackson, T. W., et al. (2002). "The cost of email within organizations." Strategies for eCommerce Success: 307.

[14] Jackson, T. W., et al. (2003). "Understanding email interaction increases organizational productivity." Communications of the ACM **46**(8): 80-84.

[15] Neustaedter, C., et al. (2005). The Social Network and Relationship Finder: Social Sorting for Email Triage. CEAS.

[16] Stevens, G. R. and J. McElhill (2000). "A qualitative study and model of the use of e-mail in organisations." Internet Research **10**(4): 271-283.

[17] ur Rehman, S. and V. Ramzy (2004). "Internet use by health professionals at the Health Sciences Centre of Kuwait University." Online Information Review **28**(1): 53-60.

[18] Whittaker, S., et al. (2006). "Email in personal information management." Communications of the ACM **49**(1): 68-73.

Privacy Preserving Data Publishing: A Classification Perspective

A N K Zaman
School of Computer Science
University of Guelph
Guelph, ON, CANADA

Charlie Obimbo
School of Computer Science
University of Guelph
Guelph, ON, CANADA

***Abstract*—The concept of privacy is expressed as release of information in a controlled way. Privacy could also be defined as privacy decides what type of personal information should be released and which group or person can access and use it. Privacy Preserving Data Publishing (PPDP) is a way to allow one to share anonymous data to ensure protection against identity disclosure of an individual. Data anonymization is a technique for PPDP, which makes sure the published data, is practically useful for processing (mining) while preserving individuals sensitive information. Most works reported in literature on privacy preserving data publishing for classification task handle numerical data. However, most real life data contains both numerical and non-numerical data. Another shortcoming is that use of distributed model called Secure Multiparty Computation (SMC). For this research, a centralized model is used for independent data publication by a single data owner. The key challenge for PPDP is to ensure privacy as well as to keep the data usable for research. Differential privacy is a technique that ensures the highest level of privacy for a record owner while providing actual information of the data set. The aim of this research is to develop a framework that satisfies differential privacy standards and to ensure maximum data usability for a classification tasks such as patient data classification in terms of blood pressure.**

Keywords*—*privacy preserving data publishing; differential privacy

I. Introduction

Increase in large data repositories in the recent past by Corporations and Governments have given credence to developing information-based decision-making systems through data-mining. For example, all California based, licensed hospitals have to submit person-specific data (Zip, Date of Birth, admission and release dates, principal language spoken etc.) of all discharged patients to the California Health Facilities Commission to make that data available for interested parties (e.g., insurance, researchers) to promote Equitable Healthcare Accessibility for California [1]. In 2004, the Information Technology Advisory Committee of the President of the United States published a report with the title Revolutionizing Health Care through Information Technology [2], to impose emphasis to implement a nationwide electronic medical record system to promote and to encourage medical knowledge sharing throughout the computer-assisted clinical decision support system. Publishing data is beneficial in many other areas too. For example, in 2006 Netflix (online DVD Rental Company) published 500,000 movie ratings data set from subscribers to encourage research to improve the movie recommendation accuracy on the basis of personal movie preferences [3]. From Oct 2012 Canada and US governments started a pilot project called ”Entry/Exit pilot project” [4] to share travellers (citizens and permanent residents of both countries) biographic data of people who cross the US/Canada border among The Canada Border Services Agency (CBSA) and the Department of Homeland Security (DHS). This is an example of data sharing between two governments.

Table I presents a raw data about patients, where, every row belongs to a single patient. After applying, generalization, anonymized data is published in Table II.

TABLE I: RAW DATA ABOUT PATIENT

Record ID	Sex	Age	Disease Code	Class
1	Female	33	15, 16, 31, 32	N
2	Female	60	15, 31	Y
3	Female	37	16	Y
4	Female	35	15, 16	N
5	Male	16	15	N
6	Male	36	16, 31	Y
7	Female	46	15, 16, 31, 32	N
8	Male	27	16, 31, 32	Y

TABLE II: ANONYMIZED DATA TABLE FOR PUBLICATION

Age	Sex	Disease Code	Class	Count
[15-60)	Female	1*	Y	2
[15-60)	Female	1*	N	3
[15-60)	Male	1*	Y	2
[15-60)	Male	1*	N	1
[15-60)	Female	1*, 3*	Y	1
[15-60)	Female	1*, 3*	N	2
[15-60)	Male	1*, 3*	Y	2
[15-60)	Male	1*, 3*	N	0

The taxonomy tree used for generalization Table I is given in Figure 1. The taxonomy tree is presenting two attributes age and disease code. The attribute age could be divided into two different calsses as 15 to 30 and 30 to 60. In a similar way, four different disease codes are generalized as 1* and 3*. In Table I, although there is no identifiable information (e.g. name or phone number) about any patient, still the privacy of patient is vulnerable due to background knowledge of a malicious user of the data set. For example, if a malicious user knows that the disease code 32 belongs to a Male patient, then it is easy to identify the record #8, as it is the only Male

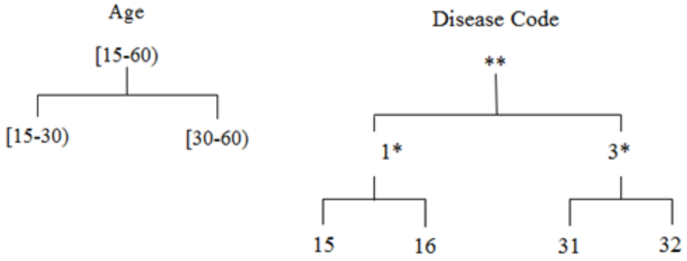


Fig. 1: Taxonomy tree for Attributes Age and Disease Code

has disease 32. On the other hand, after location that patient, the malicious user can also know that the targeted Male has diseases 16 and 31.

The rest of the paper is organized as follows: Section II surveys the related recent published work. Section III sates the problem statement. Section IV discusses the proposed system and experimental setup. Section V mentions the contributions of this research. Section VI presents the pseudocode of the proposed algorithm. Section VII concludes this paper.

II. Related Works

Researchers have proposed many algorithms for Privacy Preserving Data Mining (PPDM) and PPDP, however, not much is found in literature that addresses the privacy preservation to achieve the goal of classification [5]. The following section will discuss recent works on data anonymization for classification.

Iyengar [6] first wrote his paper on privacy of data and classification. He proposed usage based metrics (general loss metric, LM and Classification metric, CM) and showed by applying generalization and/or suppression, the anonymized data is still usable for classification tasks.

A bottom-up anonymization method was proposed by Wang et al. [7], that is only able to handle categorical data for the purpose of the data classification task. Later, the same authors introduced another method called TDS (top-down specialization method) [8] and then another improvement of TDS called TDR [9] (Top-Down Refinement) which is capable to handle both categorical and numerical values for data anonymization.

Lefevre et al. [10] proposed an algorithm called Mondrian and its improved version named as InfoGain Mondrian [11] to address various anonymized data processing including classification. InfoGain Mondrian showed better performance than TDS algorithm, and it is considered as one of the benchmark algorithms for anonymized data classification task. k-anonymous decision trees [12] based algorithms was proposed by Friedman et al. in 2008. Sharkey et al. [13] also proposed a method that generated pseudo data by following the decision tree model.

Kisilevich et al. [14] presented a multi-dimensional hybrid approach called compensation which achieved privacy by utilizing suppression and swapping techniques. The authors investigated data anonymization for data classification by satisfying k-anonymization. They claimed that their work resulted in better classification accuracy on anonymized data. If suppression techniques are applied, then one of the major drawbacks is that sparse data results in high information loss [15].

Li et al. [16] proposed and demonstrated the k-anonymity based algorithm. They utilized global attribute generalization and local value suppression techniques to produce anonymized data for classification. Their algorithms showed better classification performance compared to InfoGain Mondrian [11].

Table III presents some recent works published on data anonymization and classification. Still most published works are using k-anonymity based algorithms.

TABLE III: CLASSIFICATION MODEL USED BY DIFFERENT PRIVACY PRESERVED ALGORITHMS

	K-Anonymity	$\in$-differential privacy	
[17]	Y		Hierarchical Conditional Entropy-based Top-Down Refinement (HCE-TDR)
[18]	Y		SVM Classifier
[14]	Y		Decision tree
[19]		Y	Decision tree

Fung et al. [5] presented different existing anonymization based algorithms in their paper. It is seen that most of the algorithms can handle only two attack models. So, more efficient algorithms are needed to ensure the privacy of a dataset donor and/or owner.

III. Problem Statement

The key challenge for PPDP is to ensure individuals privacy as well as to keep the data usable for research. The aim of this research is to develop a framework that satisfies differential privacy standards and to ensure maximum data usability to deal with the classification task for knowledge miners. The core benefit of this work is to ensure the ease of availability of high quality data to promote collaborative scientific research to achieve new findings.

IV. Proposed system and experimental design

The objective of this research work is to develop a stable PPDP system by addressing specific research issues for publishing anonymized data. One of the primary goals is to publish useful data set to satisfy certain research needs, e.g., classification. The following sections will discuss research questions and the proposed system to be developed:

A. *Privacy Constraint*

One of the main objectives of the proposed system is to preserve individual's privacy. k-anonymization based algorithms are susceptible to attacks that may use individual "contributor's" background and link them to the repository, to expose which tuples belong to the given individual. They are, therefore, vulnerable to record-linkage and attribute-linkage attacks. In literature, it is also found that many privacy preserving models also suffer from table linkage and probabilistic attacks. In the proposed system, differential privacy ($\in$-differential privacy) privacy will be used that is capable to protect date published from the above mentioned attacks.

Differential privacy is a new algorithm that provides a strong privacy guarantee. Partition-based [20] [21] privacy models ensure privacy by imposing syntactic constraints on the output. For example, the output is required to be indistinguishable among k records, or the sensitive value to be well represented in every equivalence group. Instead, differential privacy makes sure that a malicious user will not be able to get any information about a targeted person, despite of whether a data set contains that persons record or not. Informally, a differentially private output is insensitive to any particular record. Therefore, while preserving the privacy of an individual, the output of the differential privacy method is computed as if from a data set that does not contain her record.

Differential privacy also prevents linking a victims sensitive information from an adversary has capturing may be interested in quasi-identifiers.

1) Definition: $\in$-differential privacy: Let us consider two data sets $DB1$ and $DB2$ that differ only in one element. For both data sets $DB1$ and $DB2$, a certain query response Rs should be the same as well as satisfy the following probability distribution Pr:

$$\frac{Pr(An(DB1) = Rs)}{Pr(An(DB2) = Rs)} \leq e^{\epsilon} \tag{1}$$

where, An presents an anonymization algorithm. The parameter $\epsilon > 0$ is chosen by the data publisher. Stronger privacy guarantee could be achieved by choosing a lower value of ϵ. The values could be 0.01, 0.1, or may be $\ln 2$ or $\ln 3$ [22]. If it is a very small ϵ then

$$e^{\epsilon} \approx 1 + \epsilon \tag{2}$$

To process numeric and non-numeric data with differential privacy model, following techniques will be needed.

B. Laplace Mechanism

Dwork et al. [23] proposed the Laplace mechanism to add noise for numerical values and ensure differential privacy. The Laplace mechanism takes a database DB as input and consists of a function f and the privacy parameter λ. The privacy parameter λ specifies how much noise should be added to produce the privacy preserved output. The mechanism first computes the true output $f(DB)$, and then perturbs the noisy output. A Laplace distribution having probability density function

$$pdf(\frac{x}{\lambda}) = \frac{1}{2\lambda}e^{-}|x|/\lambda \tag{3}$$

generates noise, where, x is a random variable; its variance is $2\lambda^2$ and mean is 0. The sensitivity of the noise is defined by the following formula:

$$\hat{f}(DB) = f(DB) + lap(\lambda) \tag{4}$$

where, $lap(\lambda)$ is sampled from Laplace distribution. The mechanism

$$\hat{f}(DB) = f(DB) + lap(\frac{1}{\epsilon}) \tag{5}$$

ensures ϵ-differential privacy.

C. Exponential Mechanism

McSherry and Talwar [24] proposed an exponential mechanism to handle non-numeric data. This mechanism takes as input, a data set DB that encompass an output range, τ, privacy parameter, ϵ and a utility function

$$u : (DB \times \tau) \rightarrow R$$

to produce an output, $t \in \tau$, having real value score where a better utility is proportional to higher score. A probability distribution is introduced by this mechanism which samples an output over the range τ. The sensitivity of the function is defined by

$$\Delta u = max_{\Delta_{(t,DB,\hat{DB})}} |u(DB,t) - (\hat{DB},t)| \tag{6}$$

The probability associated with every output is proportional to

$$e^{\frac{\epsilon u(DB,t)}{2\Delta u}} \tag{7}$$

$i.e.$ the higher score should be chosen exponentially with an output.

D. Anonymization

Ideas of interactive and non-interactive [19] anonymization techniques are as follows. The non-interactive approach is chosen for this research work. In literature, differential privacy method is widely used in an interactive framework [25] [23] [26] [27]. In case of a non-interactive framework, anonymized data set is published by the owner for public use after changing the raw data to an anonymous form. In this research the non-interactive framework is adopted. This is due to the fact that this approach has a number of advantages over its counterpart (interactive approach). In an interactive framework, noise is added to every query response to ensure privacy. To ensure privacy, a database owner has a privacy constraint to answer queries with a certain limit before he/she has to increase the noise level to a point that the answer is no longer useful. Thus, the data set can only support a fixed number of queries for a given privacy budget. This is a critical problem when there are a large number of data miners that want to access the data set, because each user (data miner) can only allow to ask a small number of queries. Even for a small number of users, it is not possible to explore the data for testing various hypotheses.

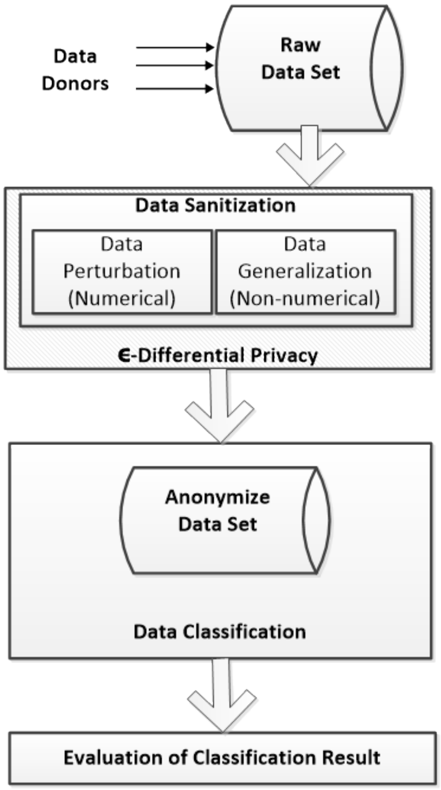

Data Donors
Raw Data Set
Data Sanitization
Data Perturbation (Numerical)
Data Generalization (Non-numerical)
ϵ-Differential Privacy
Anonymize Data Set
Data Classification
Evaluation of Classification Result

Fig. 2: Data Flow Diagram of the Proposed System

E. Generalization

Definition: Let

$$DB = r_1, r_2, ..., r_n$$

be a set of records, where every record r_i represent the information of an individual with attributes

$$A = A_1, A_2, ..., A_d$$

It is assumed that each attribute A_i has a finite domain, denoted by $\Omega(A_i)$. The domain of DB is defined as

$$\Omega(DB) = \Omega(A_1) \times \Omega(A_2) \times ... \times \Omega(A_d)$$

To anonymize a data set DB, the process of generalization takes place by substituting an original value of an attribute with a more general form of a value. The exact general value is chosen according to the attribute partition.

Figure 2 represents the data flow diagram of the proposed system. In the first step, the raw data is collected from the data donors', for example, in case of a medical data, patients of a hospital would be the data donors. After collecting the raw data, a sensitization algorithm is applied on the data to preserve individual's privacy. Finally, the sanitized data is released for the research community for further processing for knowledge mining.

V. CONTRIBUTIONS

The following sections will discuss the important contributions of this research.

A. Securing Data Donors Privacy

By surveying the literature it is found that k-anonymy and various extension are susceptible to different attacks such as attribute linkage attack, background knowledge attack, table linkage attack and probabilistic attack. Differential privacy provides the strongest privacy guarantee and a differentially private output is insensitive to any particular record. Differential privacy model is able to protect all above mentioned attacks. In this research, differential privacy will be used along with generalization.

B. Handling High Dimensionality of Data Set

Measuring and Collecting information about an individual is becoming easier due to the improved technology. As a result, the number of attributes is rising and the size of the domain is growing exponentially. To handle that high dimensional data set, this research is going to implement the idea of multiple releases of anonymized data instead of publishing the whole data in a single time. A data set with different attributes could be utilized with different interest groups for their own research needs. Suppose there is a Table T contains data donors personal data, for example, (Employment Status, Gender, Age, Race, Disease, income). An interested group (for example a health insurance company) for the mentioned Table T, interested to classify data and wants to model the relation between disease and gender, age, income. Another interested group (for example a non-government organization (NGO) that works for different social services) may be interested to cluster data with attributes employment status, age, race. One solution is to publish the attributes in a single release Employment status, Gender, age, race, income for both interested groups; however, the problem with such release is that none of the group needs all released attributes. On the other hand, publishing more attributes together makes data donors vulnerable to malicious users. If the required information for different analysis is separate then publishing data for both cases at once may not necessary for those cases. Alternatively, publishing anonymized data based on the data recipients need is a better way to address the specific need of an analysis. Publishing multiple views of data, may be a more efficient way to handle high-dimensional data sets.

C. Re-usability of Data

Another goal of this research is to increase data re-usability through multiple publications of anonymous data. By the course of time, every day, data is collected and stored. So, multiple publishing of anonymized data gives an opportunity for data re-usability. For example, say the data owner has two sets of health care data for the years 1995-2004 and 2005-2014. One can publish the entire data set in an anonymous form in a single time. However, if any researcher wants data from the range 2004-2009, then the data owner could publish the anonymous data for the desired range instead of giving two different data sets with lots of redundant information.

D. Minimizing Redundancy in Published Anonymized Data

In literature, all the existing non-interactive privacy preserving models publish data once and made the data available for the interested parties. One of the major drawbacks of this paradigm is data redundancy. For this research, purpose-based (e.g. based on time span or based on specific attributes etc.) releases of anonymized data are aimed to address the classification task to avoid data redundancy.

VI. Pseudocode for the proposed algorithm

The following section presents the pseudocode for the proposed system:

```
1) START
2) Read the raw data set DB
3) Create purpose-based tailored data set TDB
     $a$). Based on certain time span [reflects section V(C)] (if NO go to b)
     $b$). Based on selection of attributes [reflects section V(B)]
4) Follow taxonomy tree for TBD
5) Apply generalization and ensure differential privacy:
     $a$). Apply Exponential Mechanism [case of non-numeric data]
     $b$). Apply Laplace Mechanism [case of numeric data]
6) Generate generalized privacy preserve data set GTDB.
7) Apply classification technique
8) Evaluation of classification accuracy.
9) END.
```

VII. Conclusion

In this paper the idea of privacy preserving data publishing is discussed for data classification purpose. The goal of this work is to implement a practical privacy preserving framework to keep privacy of an individual while keeping the anonymized data useful for the researcher. The core benefit of this work is to promote data sharing for knowledge mining. Differential privacy along with generalization creates a strong privacy guarantee and data utility for data miners.

References

[1] R. S. B. David and C. Schoenfelder, "California inpatient data reporting manual, medical information reporting for california," Office of Statewide Health Planning and Development, Tech. Rep., 09 2013. [Online]. Available: http://www.oshpd.ca.gov/hid/mircal/

[2] P. I. T. A. Committee, "Revolutionizing Health Care Through Information Technology," www.nitrd.gov/pitac/meetings/2004/, June 2004.

[3] I. Netflix, "Netflix Prize," http://www.netflixprize.com//index, February 2013.

[4] C. B. S. A. (CBSA), "Entry/Exit pilot project," http://www.cbsa-asfc.gc.ca/media/release-communique/2012/2012-09-28b-eng.html, July 2014.

[5] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010. [Online]. Available: http://doi.acm.org/10.1145/1749603.1749605

[6] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. New York, NY, USA: ACM, 2002, pp. 279–288. [Online]. Available: http://doi.acm.org/10.1145/775047.775089

[7] K. Wang, P. S. Yu, and S. Chakraborty, "Bottom-up generalization: A data mining solution to privacy protection," in *Proceedings of the Fourth IEEE International Conference on Data Mining*, ser. ICDM '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 249–256. [Online]. Available: http://dl.acm.org/citation.cfm?id=1032649.1033461

[8] B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *Proceedings of the 21st International Conference on Data Engineering*, ser. ICDE '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 205–216. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2005.143

[9] C. M. Fung Benjamin, K. Wang, and P. S. Yu, "Anonymizing classification data for privacy preservation," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 5, pp. 711–725, May 2007. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2007.1015

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proceedings of the 22Nd International Conference on Data Engineering*, ser. ICDE '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 25–. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2006.101

[11] L. Kristen, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization techniques for large-scale datasets," *ACM Trans. Database Syst.*, vol. 33, no. 3, pp. 17:1–17:47, Sep. 2008. [Online]. Available: http://doi.acm.org/10.1145/1386118.1386123

[12] A. Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," *The VLDB Journal*, vol. 17, no. 4, pp. 789–804, Jul. 2008. [Online]. Available: http://dx.doi.org/10.1007/s00778-006-0039-5

[13] P. Sharkey, H. Tian, W. Zhang, and S. Xu, "Privacy-preserving data mining through knowledge model sharing," in *Proceedings of the 1st ACM SIGKDD International Conference on Privacy, Security, and Trust in KDD*, ser. PinKDD'07. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 97–115. [Online]. Available: http://dl.acm.org/citation.cfm?id=1793474.1793482

[14] S. Kisilevich, Y. Elovici, B. Shapira, and L. Rokach, "Protecting persons while protecting the people," C. S. Gal, P. B. Kantor, and M. E. Lesk, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, ch. kACTUS 2: Privacy Preserving in Classification Tasks Using k-Anonymity, pp. 63–81. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-10233-2-7

[15] J. Liu and K. Wang, "Anonymizing transaction data by integrating suppression and generalization," in *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*, ser. PAKDD'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 171–180. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13657-3-20

[16] J. Li, J. Liu, M. M. Baig, and R. C. Wong, "Information based data anonymization for classification utility," *Data Knowl. Eng.*, vol. 70, no. 12, pp. 1030–1045, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2011.07.001

[17] M. Ye, X. Wu, X. Hu, and D. Hu, "Anonymizing classification data using rough set theory," *Knowl.-Based Syst.*, vol. 43, pp. 82–94, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.knosys.2013.01.007

[18] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29 2009 - April 2 2009, Shanghai, China*, 2009, pp. 429–440. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2009.19

[19] N. Mohammed, R. Chen, B. C. Fung, and P. S. Yu, "Differentially private data release for data mining," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 493–501. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020487

[20] R. C. wing Wong, J. Li, A. W. chee Fu, and K. Wang, "(, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *In ACM SIGKDD*, 2006, pp. 754–759.

[21] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*, 2007, pp. 106–115. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2007.367856

[22] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, 2011. [Online]. Available: http://doi.acm.org/10.1145/1866739.1866758

[23] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878-14

[24] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, 2007, pp. 94–103. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/FOCS.2007.41

[25] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '03. New York, NY, USA: ACM, 2003, pp. 202–210. [Online]. Available: http://doi.acm.org/10.1145/773153.773173

[26] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," in *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, ser. STOC '10. New York, NY, USA: ACM, 2010, pp. 765–774. [Online]. Available: http://doi.acm.org/10.1145/1806689.1806794

[27] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 493–502. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835868

Reducing Shared Cache Misses via dynamic Grouping and Scheduling on Multicores

Wael Amr Hossam El Din
Computer Department ,Faculty of Engineering,
Cairo University,
Giza,Egypt
wael.amrhossam@gmail.com

Hany Mohamed ElSayed
Communication Department ,Faculty of Engineering,
Cairo University,
Giza,Egypt
helsayed@ieee.org

Ihab ElSayed Talkhan
Computer Department,
Faculty of Engineering,
Cairo University,
Giza,Egypt
italkhan@aucegypt.edu

***Abstract*—Multicore technology enables the system to perform more tasks with higher overall system performance. However, this performance can't be exploited well due to the high miss rate in the second level shared cache among the cores which represents one of the multicore's challenges.**

This paper addresses the dynamic co-scheduling of tasks in multicore real-time systems. The focus is on the basic idea of the megatask technique for grouping the tasks that may affect the shared cache miss rate ,and the Pfair scheduling that is then used for reducing the concurrency within the grouped tasks while ensuring the real time constrains. Consequently the shared cache miss rate is reduced.The dynamic co-scheduling is proposed through the combination of the symbiotic technique with the megatask technique for co-scheduling the tasks based on the collected information using two schemes. The first scheme is measuring the temporal working set size of each running task at run time, while the second scheme is collecting the shared cache miss rate of each running task at run time.

Experiments show that the proposed dynamic co-scheduling can decrease the shared cache miss rate compared to the static one by 52%.This indicates that the dynamic co-scheduling is important to achieve high performance with shared cache memory for running high workloads like multimedia applications that require real-time response and continuous-media data types.

Keywords*—*Shared Cache Miss Rate; Dynamic Scheduling; Multicore; Symbiosis;

I. Introduction

Processor industry has moved towards the multicore technology since the delivered performance of single cores can not meet the needed requirements for running different applications like web servers, multimedia programs and databases. Multicore technology is introduced to increase the required performance and power efficiency. However, there are challenges for this technology, one of which is that the cores share a second level L2 cache. Therefore, with increasing the workload managing the shared cache space becomes essential to avoid higher miss rate which degrades the system performance. The cost of memory access has reached roughly 300 processor cycles in 2006 and has been increasing at the rate of 50% per year [1]. Decreasing the shared cache miss rate is the focus of the research in this paper.

Chip Multiprocessor (can also be named a Multicore processor) refers to a single chip that integrates two or more processors in an area that would have originally been filled with a single large uniprocessor. This solves the power consumption problem when adding more transistors to the uniprocessor and switching it at higher and higher frequencies.

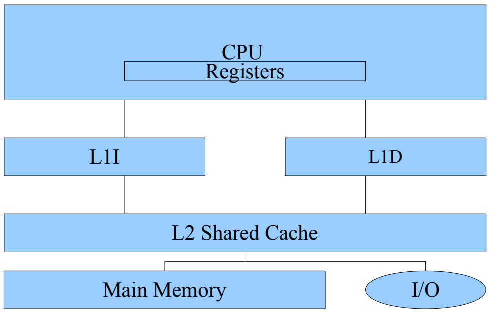


Figure 1 : Conventional Microprocessor

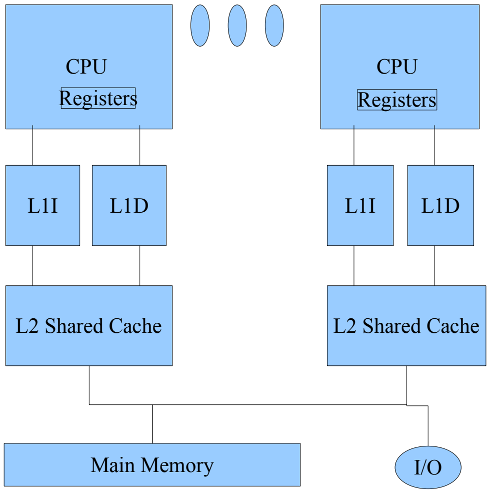


Figure 2 : Simple Chip Multiprocessor

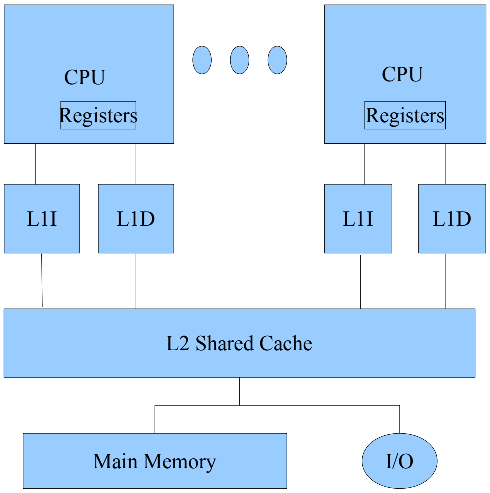


Figure 3 : Shared Cache Chip Multiprocessor

Figures 1-3 show the difference between the uniprocessor and the multicore systems. Figure 1 shows the conventional microprocessor architectures on which the other architectures are based. In figure 2 there are N cores that share only main memory and I/O, but in figure 3 there are N cores that have separate L1 cache memories and shared L2 cache. There are benefits for the shared cache architecture. For example, it can provide high bandwidth, low latency connection for the cores to communicate the shared data with each other[2]. Throughout this paper we will consider the shared cache chip multiprocessor architecture shown in figure 3.

This paper addresses the reduction of the L2 shared cache miss rate while ensuring the appropriate satisfaction of real time constraints of the tasks. For this purpose, we present two schemes for re-packing tasks in groups.Each task has a utilization and working set size (WSS)[3].The utilization indicates the core share that each task requires.The working set is defined as the collection of information referenced by the task during the task interval time ,while the working set size (WSS) is defined as the number of pages in the working-set.The temporal working set size (TWSS) is defined as the WSS every certain number of clock cycles.The first scheme is based on re-calculating the TWSS of each running task at run time. Then, it re-packs the tasks in groups. Each group has a number of tasks such that their total temporal WSS is less than or equal certain threshold. The second scheme is based on using the on-line counters for misses of each task at run time. Then it re-packs the tasks such that the miss rates are equally distributed on the groups. By this way, we can avoid the situation in which any group has a much higher miss rate than other groups. After that, each group in these two schemes is assigned a certain number of cores such that the maximum combined TWSS of all executing tasks is bounded at a value less than the capacity of the second level L2 shared cache. This will reduce concurrency within each group, eliminate the L2 shared cache thrashing, and reduce its miss rate. Finally the Pfair scheduling algorithm is used for selecting some tasks within each group for running.

The rest of the paper is organized as follows. In section II, we review the related work. In section III, we give details of the grouping and scheduling techniques that the proposed methods are based on. In section IV, we present the simulation methodology. In section V, we show the simulation results. Finally, we present the conclusions in section VI.

II. Literature Review

Different scheduling algorithms for multicore systems have been introduced in previous researches. Fedorova *et al.* introduced using the instruction mix as a heuristic for the scheduling decisions [4]. Then, in [5] they showed that the miss rate for the second level (L2) shared cache can have the greatest negative impact on processor performance, consequently they introduced the balance-set principle for grouping all the runnable threads, such that the combined working set of each group fits in the cache. After that they introduced in [1],[6] the non-work-conserving , the target-miss-rate, and the cache fair algorithms for reducing miss rate of the L2 shared cache. These algorithms are based on using analytical performance models and online performance monitoring. Although they showed different techniques for resolving the L2 shared cache miss rate, they did not consider real time scheduling.

Real time scheduling was introduced by Anderson *et al.* [7] on which we build our proposed method. They introduced the concept of a megatask that simply represents a set of tasks to be treated as a single scheduling entity. They proposed a scheme for incorporating the megatask concept into a Pfair scheduled system. In [8] they proposed heuristics and other methods like the spread-cognizant method [9] to support both encouragement and discouragement of the co-scheduling of groups of tasks simultaneously while ensuring the satisfaction of real-time constraints. On the other hand, Anderson's work was achieved under static co-scheduling, while we consider dynamic co-scheduling.

There are other papers that introduced scheduling algorithms that aim at increasing the system throughput of the multicore platform without considering the L2 shared cache miss rate or considering the real time scheduling. For example,Zhang *et al.* [10] proposed a hot-page coloring approach for the L2 shared cache partitioning. Cong *et al.*[11] proposed algorithms for reconfigurable resource allocation and job scheduling for achieving high performance. Azimi *et al.* [12] proposed a cache partition mechanism for partitioning the L2 shared cache among the applications based on guiding the allocation of its physical pages. Yang et al. [13] proposed a cache-aware scheduling policy which improves cache performance by considering data reuse, memory footprint of co-scheduled tasks, and coherency misses. Wang et al. [14] presented a hybrid approach for partitioning the multicores into clusters that share the L2 cache, then the tasks which access a common region of memory are statically assigned to the same cluster.

III. Grouping & Scheduling Tasks Among Multicore Platform

The main two steps of our proposed approach is grouping and scheduling. For grouping, we are interested in combining between the two task-grouping techniques which are the symbiotic techniques and the megatasking. For scheduling, we consider applying the Pfair scheduling algorithm for ensuring the satisfaction of real time constraints for tasks.

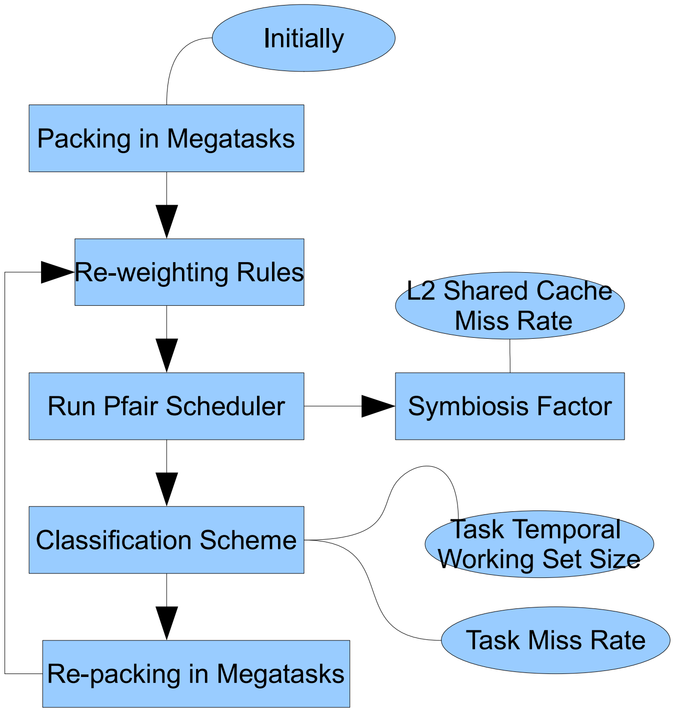


Figure 4 : The Proposed Dynamic Co-scheduling Technique

Figure 4 shows the main sequence for the proposed dynamic co-scheduling technique. It is an iterative method based on packing the tasks in groups, then running the scheduler for certain number of clock cycles. Then, getting some statistical information about the running tasks, re-grouping of tasks based on the obtained statistical information is made .These steps will be explained more in the following subsections.

A) Grouping

Megatasks

At the initialization phase,the tasks are chosen randomly and grouped into megatasks[7]. Each megatask has a utilization that is equal to the total utilizations of its tasks. This utilization can be also termed the cumulative weight that is used to allocate one or more processors time in discrete quanta. Let Ɣ be a megatask. Its cumulative weight can then be expressed as:

$$W_{sum}=\sum_{T\in\gamma} wt(T) \quad (1)$$

where T is the component task within the Ɣ, and wt(T) is the utilization(i.e. weight) of each task.

Packing Strategy

As mentioned before, the tasks are grouped into megatasks such that one megatask is created at a time. The megatask is closed when the ratio of the total TWSS of the packed tasks to the size of the L2 shared cache is equal to or greater than certain integer, then another new megatask is created and so on until all the tasks are grouped.

Re-weighting Rules

After creating the megatasks,each megatask cumulative weight W_{Sum} should be inflated.This inflation is referred as the megatask scheduling weight W_{Sch} .It is computed for each megatask Ɣ using the re-weighting rules[7] in order to guarantee that its grouped-tasks meet their deadlines.This is done by assigning each megatask a number of cores indicated by the calculated W_{Sch} as shown:

$$W_{Sch}=W_{Sum}+\Delta f \quad (2)$$

Where Δf is the inflation value and can be calculated in [7].

Symbiosis Factor

After assigning each megatask with a certain number of cores based on its W_{Sch} ,then the Pfair scheduler starts running and at run time ,the symbiosis factor is calculated. Symbiosis is a co-scheduling technique whose concept is derived from the meditation of the nature in which close and often long-term interaction between two or more different biological species is established so that they can rely on and benefit from each other.

Similarly, symbiosis is applied on the scheduling of tasks, hence symbiosis is a factor that indicates the performance of tasks that are co-scheduled and compete in hardware resources every cycle [15]. This factor may be based on system performance, system utilization, energy delay product, cores energy, cores power, average normalized turnaround time (time between submitting a job to the system and its completion) ,cache sensitivity and cache intensity, or average shared cache miss rate[15-18]. Throughout this paper, we will consider the symbiosis factor as the miss rate for the second level L2 shared cache.

It is found that computing the symbiosis factor is based on two main techniques which are sampling and probabilistic modeling.

Sampling This technique is known as SOS (Sample, Optimize, Symbiosis) [15, 17, 19]. It is based on producing a profile for all the possible combinations for scheduling the tasks, then taking the schedule decision based on the highest symbiosis factor.

Probabilistic Modeling The main drawback from the sampling technique is the large overheads resulted from profiling the different combinations of tasks to have the information necessary for scheduling. Hence the probabilistic job symbiosis modeling [18] is used to eliminate this drawback through predicting the symbiosis factor for the co-scheduled tasks without the need for the “Sampling Phase”. This technique is designed based on the following main steps:

1. The cycle stack [18,20] is calculated for each task.It is consists of three components:

Base cycle count: number of times the processor dispatches instructions for the task.

Miss event cycle count: number of times the processor consumes cycles handling miss events.

Waiting cycle count: number of times the processor dispatches instructions for another task and therefore can not make progress for the given task.

2. The probabilities for base cycle count,miss event cycle count and waiting cycle count for each task are calculated.This is done through normalizing each cycle count (i.e. Base ,miss ,or waiting cycles count) to their overall sum (i.e. Base cycle count + Miss cycle count +Waiting cycle count).

As a consequence of the advantages of using probabilistic symbiotic modeling rather than using sampling, we use its concept to calculate the miss cycle count for each task without including the base and waiting cycles counts. Then we use this miss cycle count in calculating the task miss rate as shown:

$$Task\ Miss\ Rate = \frac{Task\ Miss\ Counter}{L2\ Miss\ Counter + L2\ Hit\ Counter} \quad (3)$$

where “Task Miss Counter” is the total misses in the L2 shared cache for a task,”L2 Miss Counter” is the total misses for all the running tasks ,and ”L2 Hit Counter” is the total hit for all the running tasks.

Classification Scheme

The scheduler takes decision to re-pack the system tasks into new megatasks every T clock cycles. This decision should be taken based on the classification scheme and the obtained information about the tasks. The classification scheme reflects how the threads affect each other when they are competing for shared resources. Consequently, it enables the scheduler to predict the performance effects of co-scheduling any group of threads in a shared cache system.

There are many different classification schemes like animalistic taxonomy, SDC, and pain. The most suitable classification schemes in our case of decreasing the miss rates among the L2 shared cache were proposed in [21,22]. These papers propose the classification schemes based on the collected information at run time. We use one classification scheme based on miss rate and propose another one based on temporal working set size.

Tasks can be classified based on the miss rate which plays a key role in the performance. The performance degradation is exacerbated by the tasks that have high miss rate due to memory controller contention, memory bus contention, and prefetching hardware contention. Hence the miss rate of each task can be obtained online using hardware counters, then the scheduler identifies the high miss rate applications and separates them into different groups, such that no one group will have a much higher total miss rate than any other group. Other metrics rather than miss rate can be also used, such as cache access rate and IPC, but the miss rate has been proved to give the best results. Hence, the miss rate classification scheme is a suitable scheme for our work. Besides that ,we propose another new classification scheme which is based on the TWSS of each task that is calculated every T clock cycles at run time.

Dynamic Grouping

Finally,the re-packing of the tasks in new megatasks should be done based on the classification scheme:

- In case of classification based on TWSS: the criteria of packing is exactly the same at the initialization phase.
- In case of classification based on miss rate (MR), the criteria of packing is based on that proposed in [21], in which the scheduling algorithm Distributed Intensity Online (DIO) takes the decision based on the miss rate classification. DIO uses performance counters at run time to get the miss rates of tasks (according to equation (3)). Hence, DIO observes the miss rates periodically not more frequently than once every one billion cycles in order to account for phase changes of tasks with low overhead resulted from the migrations. Then the scheduler assigns the tasks across the initially created megatasks such that the miss rates are distributed as evenly as possible according to the miss rate (MR) classification scheme.

Then , the W_{Sch} for each megatask is computed using the re-weighting rules.

Condition of Qualified Megatask

The total number of cores ,that are assigned to the megatasks, should not exceed the number of the system cores.

B) Scheduling

Pfair Scheduler

The second main phase is that at run time, the Pfair scheduling algorithm is used to serve the tasks within each megatask under assumption that every core is single-threaded (i.e. can only serve one data address request and one instruction address request). The most efficient Pfair scheduling algorithm is an algorithm called PD^2 .

Pfair scheduling can be used to schedule a periodic task system τ in which the tasks are assigned with the processor time in discrete time units that is represented with the time interval [t, t + 1), where t is a nonnegative integer, as slot t. The sequence of these scheduling decisions over time defines a schedule[23].

Each task T of the task system τ is assigned a rational weight wt(T) ∈ (0, 1] that denotes the processor share it requires. For a periodic task T ,

$wt(T) = \frac{T_e}{T_p}$ (4)where T_e and T_p are the (integral) execution cost and period of T .

Tasks' Division Each task T in τ is divided into an infinite sequence of quantum-length subtasks, $T_1, T_2, \cdots, T_i$ where each subtask T_i has an associated release $r(T_i)$ and deadline $d(T_i)$, defined as follows (for proof see [24]):

$$r(Ti)=\left\lfloor \frac{i-1}{wt(T)} \right\rfloor \quad (5)$$

$$d(Ti)=\left\lceil \frac{i}{wt(T)} \right\rceil \quad (6)$$

Tie-breaking Rules The Pfair scheduler PD^2 has two tie-breaking rules which are used for breaking between the sub-tasks that have the same deadlines.

First tie-break:The successor bit b(T)

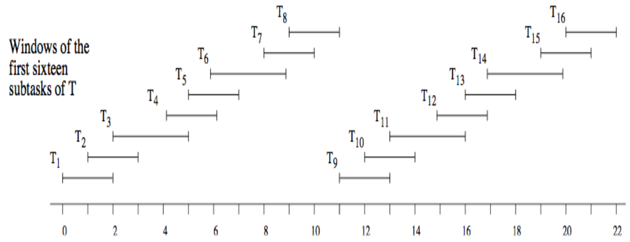


Figure 5 : Windows of the first 16 subtasks of Task T

As shown in the above figure 5,successive windows of a sub-task are either disjoint or overlap by one slot.

For example :

- The deadline of T_1 is 2 while the release time of T_2 is 1.
- The deadline of T_2 is 3 while the release time of T_3 is 2.
- The deadline of T_3 is 5 while the release time of T_4 is 4, and so on.

In other words,Formally let a sub-task T_{i+1} and its predecessor T_i ,so the release time of T_{i+1} will be equal to either :

- deadline of T_i ,or
- deadline of T_i-1

From this point, they have defined a bit $b(T_i)$ that distinguishes between these two possibilities:

- $b(T_i)=1$ if release time of T_{i+1} = deadline of T_i -1
- $b(T_i)=0$ if release time of T_{i+1} = deadline of T_i

Second tie break:The group deadline D(T)

Consider a sequence $T_i,\ldots,T_j$ of subtasks such that $b(T_k)=1$ and $|windowLength(T_{K+1})|=2$ for all i ≤ k < j.

For introducing the group deadline,as shown in figure 5,scheduling T_i in its last slot forces the other subtasks in this sequence to be scheduled in their last slots.

For example, scheduling T_3 in slot 4 forces T_4 and T_5 to be scheduled in slots 5 and 6, respectively. So the group deadline of a subtask T_i , denoted $D(T_i)$, is the earliest time by which such a “cascade” must end.

Formally, it is the earliest time t, where $t \geq deadline(T_i)$, such that either:

- $t=deadline(T_k)$ and $b(T_k)=0$,or
- $t+1=deadline(T_k)$ and $|windowLength(T_k)|=3$ for some subtask T_K .

For example, in the above Figure, $D(T_3)=d(T_6)-1=8$ and $D(T_7)=d(T_8)=11$.

Now after defining the successor bit b(T) and the group deadline D,the next step is showing the PD^2 priority rules.

The PD^2 Priority Definition The PD^2 Priority is based on the successor bit b(T),the group deadline D ,and the deadline of each subtask d(T) as will be shown.

Under PD^2 , subtask T_i priority is at least that of subtask U_j , denoted $T_i \preceq U_j$, if one of the following rules is satisfied:

I. $d(T_i)<d(U_j)$.

II. $d(T_i)=d(U_j)$ and $b(T_i)>b(U_j)$.

III. $d(T_i)=d(U_j)$, $b(T_i)=b(U_j)=1$,and $D(T_i)\geq D(U_j)$.

IV. Simulation Methodology

In this section we are going to show the stages of building the simulation environment. The cache simulator is based on trace-driven model and is written in C++. It models the private cache among each core, the shared cache, the main memory and the memory requests. Also It models the Modified Exclusive Shared Invalid (MESI) cache coherence protocol.

Design Phases

First phase: Memory Trace Collection: It's a memory-access trace file based on using the “Pin” dynamic binary instrumentation framework for the IA-32 and x86-64 instruction-set architectures [25]. The “Pin” contains a tool that can be modified for printing the address of every instruction and data that are executed within the running application. The running application is represented by SPECjvm2008 benchmarks [26] that contains 38 workloads intended to represent a diverse set of common types of computation for real-world applications including text/character processing, numerical computations, and bitwise computation. Consequently we run each workload alone with the pin tool to capture all its memory accesses, then these accesses are dumped into two trace files, one for the instructions addresses and the other for the data addresses. These trace files contain entries, where each entry has a cache type (data or instruction), an address and an access type (read or write).

Second phase: Build The Proposed Scheduler:This is the implementation of the Pfair for scheduling the workloads at run time and megatask grouping of the tasks. It includes four configurations that will be run for every test case in section V.

First Configuration: Pfair without grouping:

- There is no grouping for tasks.
- Each task is stored in its own queue.
- Each task is assigned one core at the initialization phase.

Second Configuration: Static Megatask based on Working Set Size (WSS):

- There is only static grouping for tasks such that each task is packed in a megatask at the initialization phase.
- The criteria for closing the megatask and creating a new one is that the ratio of the total TWSS of the packed tasks to the size of the L2 shared cache is equal to or greater than certain threshold.
- Each megatask is represented by a queue.
- Each megatask has its assigned number of cores based on its re-weighting rules.

Third Configuration: Dynamic Megatask based on Working Set Size (WSS):

- The initial steps are exactly like the second configuration.
- Every certain number of clock cycles or certain number of instructions (e.g. once every ten million cycles to avoid overheads due to re-scheduling), all the megatasks are re-created based on the TWSS of each task.

Fourth Configuration: Dynamic Megatask based on Miss Rate:

- The same as the third configuration but the only difference is in the last step as every certain number of clock cycles or certain number of instructions (e.g. once every ten million cycles),all the megatasks are re-created based on the shared cache miss rate (MR) of each task where tasks are distributed on the megatasks such that the total miss rate (MR) of all the tasks is equalized across all the megatasks.

Third Phase: Cache Simulator: writing a cache simulator that models the architecture in figure 3 in which each core has a private cache and there is a shared cache among the cores. Besides that it is responsible for implementing the cache coherence protocol known as “Modified Exclusive Shared Invalid” (MESI).

Operational Scenario

In the proposed simulator, we assume single threaded core, so each core has a separate application. The scheduler serves these cores in a round robin manner. When there is a memory request, the cache simulator checks the cache type and operation type, then it sends it to the private cache L1 Data or L1 Instruction. If there is a hit, then it replies with data after the private cache latency cycles, otherwise it sends the request to the L2 shared cache, then if there is a hit, then it replies with data after the L2 latency cycles, otherwise it sends the request to the main memory, so it returns data after the main memory latency cycles.

Fourth Phase: Test Cases: These are the test cases that are represented by the mixes of different scenarios of real execution. For example we can consider the following types of mixes:

- Total WSS for the workloads that is lower than the L2 shared cache size.
- Total WSS for the workloads that is greater than the L2 shared cache size.
- Total WSS for the workloads that is equal to the L2 shared cache size.

V. Simulation Results

We compared the four configurations mentioned before: Pfair without grouping, Static megatask based on WSS at initialization phase, the newly proposed Dynamic Megatask based on WSS, and the newly proposed Dynamic megatask based on MR.

Table I shows the used configuration values for the main memory and the first level L1 private cache.Each workload in the SPECjvm2008 has a WSS of 2MB.These parameters are used in all the scenarios.

TABLE I. L1 Cache and Main Memory Parameters

Parameters	**Values**
Simulated Hardware Parameters	
Main Memory Latency	200
L1 Data Cache Size	32KB
L1 Data Line Size	64 bytes
L1 Data Associativity	4
L1 Data Latency	2
L1 Instruction Cache Size	32KB
L1 Instruction Line Size	64 bytes
L1 Instruction Associativity	2
L1 Instruction Latency	1
Task Parameters	
Working Set Size (WSS)	2 MB

The simulation results show the resulted Shared Cache L2 Miss Rate for each configuration in which total miss rate for the shared cache that is calculated as

$$\frac{\textit{shared cache L2 Total Misses}}{\textit{shared cache L2 Total Misses} + \textit{shared cache L2 Total Hits}} \quad (15)$$

A) Scenario 1

Table II shows that there are 8 workloads of total WSS 16 MB and the number of cores is 16 with L2 shared cache of size 1MB.

TABLE II. PARAMETERS

Parameters	**Values**
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	8 workloads of total WSS 16 MB
L2 Cache Size	1MB
L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	11
Cores	16

The graph in Figure 6 shows that the dynamic megatask outperforms the static one and the Pfair with no grouping. Also the total miss rates in both the dynamic megatask based on running tasks MR and based on WSS are near to each other.As this scenario represents the workloads of total WSS during a certain number of clock cycles that is equal or slightly greater than the shared cache L2 size.This leads to that the shared cache L2 miss rates of the four configurations are near to each other.The total miss rate tends to decrease with time as the workloads tend to finish and reach its end.

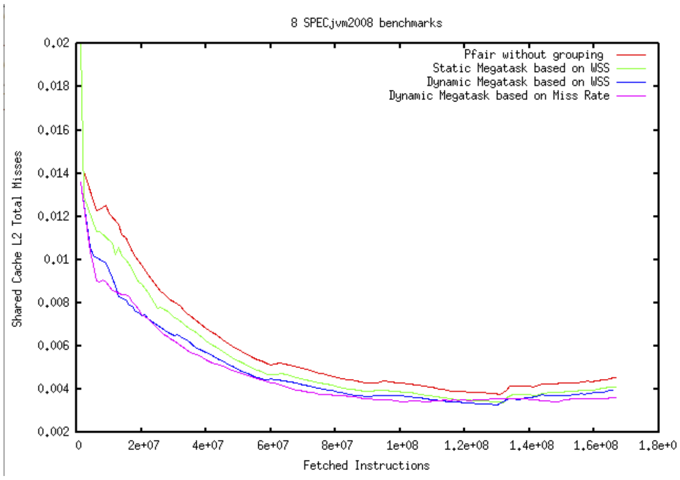


Figure 6 : 8 SPECjvm2008 benchmarks of total size 16 MB that share L2 cache of size 1 MB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

B) Scenario 2

Table III shows that there are 4 workloads of total WSS 8 MB and the number of cores is 4 with L2 shared cache of size 8MB.

TABLE III. PARAMETERS

Parameters	**Values**
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	4 workloads of total WSS 8 MB
L2 Cache Size	8MB
L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	7
Cores	4

Figure 7 shows that when the total WSS for the running workloads during a certain number of clock cycles fits the shared cache L2, the miss rate for the shared cache L2 becomes approximately the same for the static megatask and pfair without grouping while the dynamic megatask based on tasks WSS and based on tasks MR is slightly better than static megatask and pfair without grouping and tends to be the same when the system tasks tend to finish.

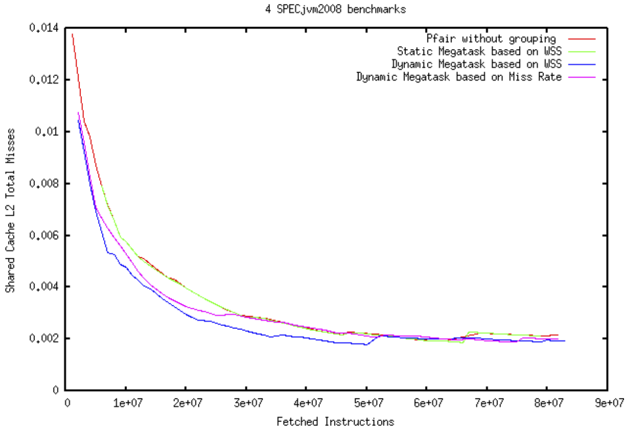


Figure 7 : 4 SPECjvm2008 benchmarks of total size 8 MB that share L2 cache of size 8 MB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

C) Scenario 3

Table IV shows that there are 5 workloads of total WSS 10 MB and the number of cores is 4 with L2 shared cache of size 20MB.

TABLE IV. PARAMETERS

Parameters	**Values**
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	5 workloads of total WSS 10 MB
L2 Cache Size	20MB
L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	20
Cores	4

Figure 8 shows that when the shared cache L2 is large enough, the shared cache L2 miss rate coincides in all the configurations.This case is the ideal case that rarely occurs.

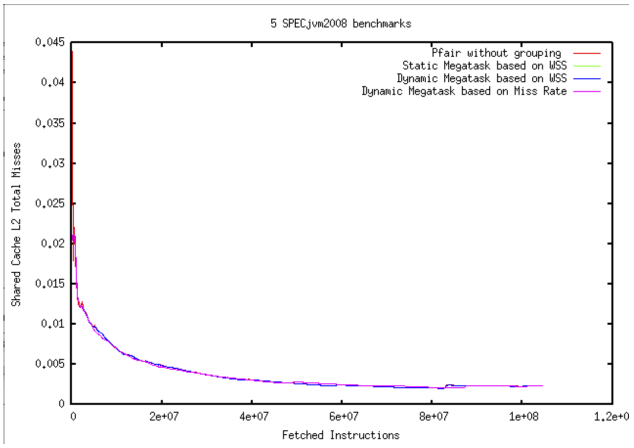


Figure 8 : 5 SPECjvm2008 benchmarks of total size 10 MB that share L2 cache of size 20 MB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

D) Scenario 4

Table V shows that there are 5 workloads of total WSS 10 MB and the number of cores is 4 with L2 shared cache of size 1MB.

TABLE V. PARAMETERS

Parameters	Values
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	5 workloads of total WSS 10 MB
L2 Cache Size	1MB
L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	11
Cores	4

The graph in figure 9 shows a slight difference in the total miss rates as the shared cache L2 still fits the total TWSS for the running workloads.But the dynamic megatask based on the WSS and MR is still the winner.

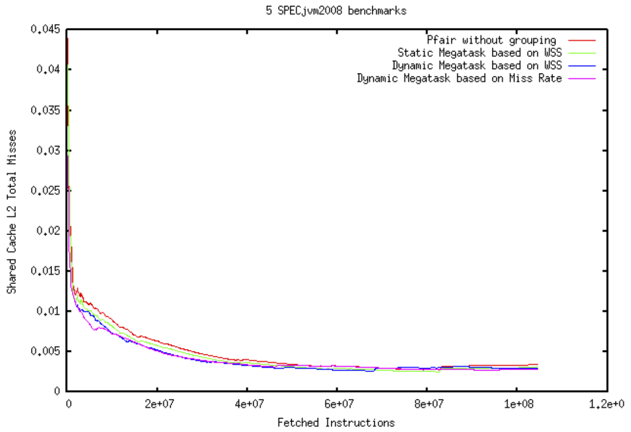


Figure 9 : 5 SPECjvm2008 benchmarks of total size 10 MB that share L2 cache of size 1 MB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

E) Scenario 5

Table VI shows that there are 16 workloads of total WSS 32 MB and the number of cores is 4 with L2 shared cache of size 1MB.

TABLE VI. PARAMETERS

Parameters	Values
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	16 workloads of total WSS 32 MB
L2 Cache Size	1MB
L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	11
Cores	4

In this scenario ,the workloads are increased while the shared cache L2 size is kept the same.This represents the case in which the total TWSS of the running workloads is always greater than shared cache L2,thus the graph in figure 10 shows that the dynamic megatask in general has a dramatic change in the shared cache L2 miss rate rather than that in the static megatask and Pfair with no grouping.The dynamic megatask based on MR slightly outperforms that is based on WSS.Hence the proposed technique is appropriate for the high processing workloads like graphics and audio applications.

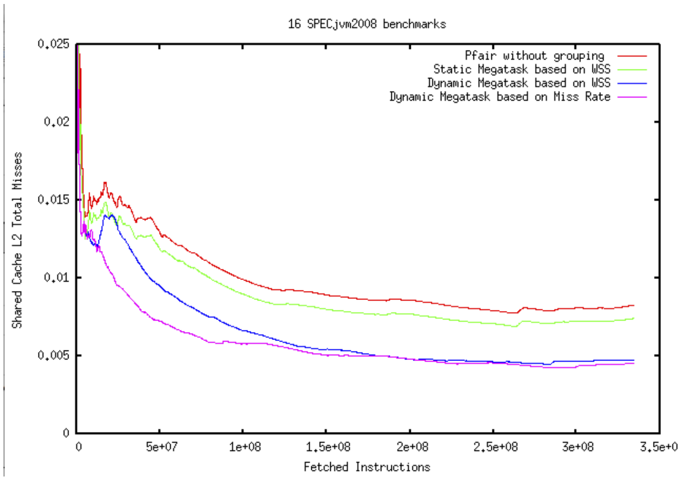


Figure 10 : 16 SPECjvm2008 benchmarks of total size 32 MB that share L2 cache of size 1 MB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

F) Scenario 6

Table VII shows that there are 12 workloads of total WSS 24 MB and the number of cores is 48 with L2 shared cache of size 512MB.

TABLE VII. PARAMETERS

Parameters	Values
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	12 workloads of total WSS 24 MB
L2 Cache Size	512KB

L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	11
Cores	48

In this scenario the shared cache L2 size is very small compared to the total WSS of the running workloads, thus the graph in figure 11 shows that the static megatask outperforms the pfair with no grouping in decreasing the shared cache L2 miss rate ,but the dynamic megatask is still the better than that the static one.This indicates that the static megatask succeeds in reducing the concurrency within the running workloads as in the static megatask ,while the dynamic one succeeds in monitoring the symbiosis factor every certain number of clock cycles ,then re-scheduling based on the two classifier schemes MR and WSS.

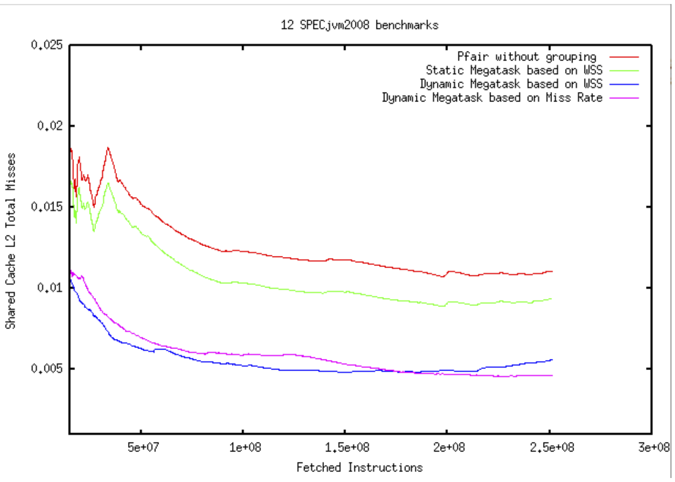


Figure 11 : 12 SPECjvm2008 benchmarks of total size 24 MB that share L2 cache of size 512 KB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

G) Scenario 7

Table VIII shows that there are 10 workloads of total WSS 20 MB and the number of cores is 40 with L2 shared cache of size 512MB.

TABLE VIII. Parameters

Parameters	Values
Simulated Hardware Parameters	
SPECjvm2008 benchmarks	10 workloads of total WSS 20 MB
L2 Cache Size	512KB
L2 Cache Line Size	64 bytes
L2 Associativity	16
L2 Cache Latency	11
Cores	40

This scenario is the same like the previous one except that the total WSS of the workloads is slightly decreased, thus the graph in figure 12 shows that the dynamic megatask is still the better in decreasing the shared cache L2 miss rate. In all graphs the two dynamic megatask configurations are always close to each other in decreasing the L2 shared cache.Hence it is interesting to try other classifier schemes that depends on the workloads requirements and the shared cache L2 characteristics.This can be considered in the future work.

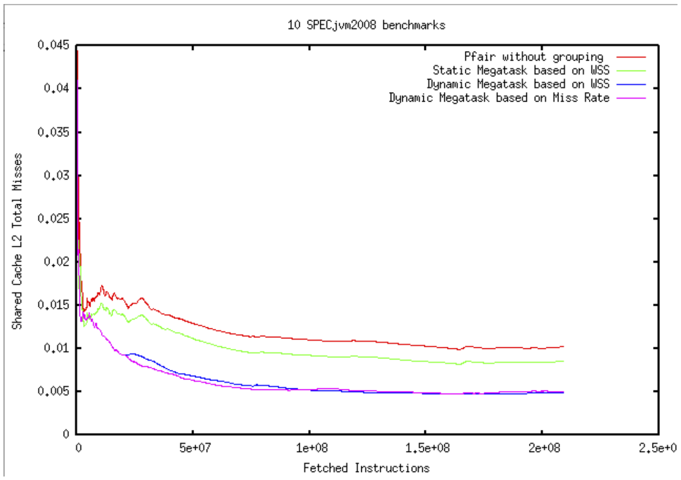


Figure 12: 10 SPECjvm2008 benchmarks of total size 20 MB that share L2 cache of size 512 KB, X-axis represents the fetched instructions and Y-axis represents the shared cache L2 Miss Rate.

Table IX shows the percentage decrease of the miss rates for the shared cache L2 in the above scenarios for the static and dynamic megatask configurations with respect to Pfair without grouping based on the following equation

$$\frac{100*(Average\ Miss\ Rate(Y)-Average\ Miss\ Rate(X))}{Average\ Miss\ Rate(Y)} \quad (16)$$

where Y represents the Pfair without grouping and X represents one of the other three configurations: static megatask , dynamic megatask based on TWSS, or dynamic megatask based on MR.

TABLE IX. Improved Shared Cache L2 Miss Rate

Scenario No.	Static Megatask	Dynamic Megatask based on WSS	Dynamic Megatask based on MR
1	5.47%	16.88%	16.19%
2	19.46%	23.79%	12.49%
3	0.15%	0.15%	0.73%
4	9.05%	13.27%	13.54%
5	10.64%	34.53%	41.98%
6	15.40%	52.70%	51.10%
7	15.17%	47.82%	48.30%

VI. Conclusions

This paper has aimed at increasing the system throughput while ensuring the real-time constraints.It tackles the dynamic grouping technique that is based on mixing between the idea of megatask and symbiosis techniques. The symbiosis techniques is used to predict a factor for each task which is either the temporal working set size or the miss rate. The megatask is used in grouping tasks based on the classification scheme according to the symbiosis factor and calculating the re-weighting rules to ensure that the tasks meet their deadlines.The Pfair scheduling is used at run time for serving bounded number of tasks within each megatask group, hence reducing the concurrency of tasks execution within each megatask which leads to reducing the second level L2 shared cache misses.The simulation results show that the dynamic grouping technique outperforms the Pfair without grouping and the static megatask. This is especially true when the shared cache size is relatively small compared to the tasks requirements such as video coding and multimedia applications.

These results suggest some points for future work. For example, as we assume that each core has single thread, this work can be extended to multi-threaded cores.The challenge key is how to distribute threads across the cores[27].Besides that, timing analysis on multicore platform can be studied. Also our work can be extended to check the overheads for tasks migration and the impact of re-scheduling.This may suggest using another techniques for determining the tasks migration threshold as in [28].

Future work is also required to evaluate these techniques to handle multi-threaded applications.In addition to that, it is interesting to use other classification schemes that is based on cache properties like cache intensity and cache sensitivity.Research is required for proposing heuristics-based co-scheduling by machine learning.

VII. Acknowledgment

The authors thank Islam M. Atta for his help in completing the presented work.

References

1. Alexandra Fedorova.“Operating system scheduling for chip multithreaded processors”,A thesis to the Division Of Engineering And Applied Sciences in partial fulfillment of the requirements for the degree of Doctor of Philosophy , Harvard University Cambridge, Massachusetts September, 2006.

2. Kunle Olukotun, Basem A. Nayfeh, Lance Hammond, Ken Wilson, and Kunyung Chang, “The case for a single chip multi-processor,” Proc. 7th Int’l Conf. Architectural Support for Programming Languages and Operating Systems, ACM Press, New York, 1996, pp. 2-11.

3. Peter J. Denning,”The working set model for program behavior ”,1st ACM Symposium on Operation Systems Principles,1968.

4. A. Fedorova, C. Small, D. Nussbaum and Margo Seltzer.”Chip multithreading systems need a new operating system scheduler”.In Proceedings of 11th ACM SIGOPS European Workshop, Leuven, Belgium, September 2004.

5. A. Fedorova, M. Seltzer, C. Small, and D. Nussbaum.”Performance of multithreaded chip multiprocessors and implications for operating system design”. In Proceedings of the USENIX 2005 Annual Technical Conf., 2005.

6. A. Fedorova,M. Seltzer and M. D. Smith.”A non-work-conserving operating system scheduler for SMT processors”. In Proceedings of the Workshop on the Interaction between Operating Systems and Computer Architecture (WIOSCA), in conjunction with ISCA-33, June 2006.

7. James H.Anderson,John M.Calandrino, and UmaMaheswari C.Devi. “Real-Time scheduling on multicore platforms”.In Proceedings of the 12th IEEE Real-Time and Embedded Technology and Applications Symposium,pp. 179-190,April 2006.

8. James H. Anderson, John M. Calandrino, and UmaMaheswari C.Devi. “Parallel task scheduling on multicore platforms”. In Real-Time Systems Symposium, 2006, RTSS '06, 27th IEEE International, pp. 89-100, December 2006.

9. James H. Anderson, John M. Calandrino, and UmaMaheswari C.Devi. “Cache-Aware real-time scheduling on multicore platforms:heuristics and a case study”. In Proceedings of the 20th Euromicro Conference on Real-Time Systems, 2008, ECRTS '08, pp. 299-308, 2-4 July 2008.

10. R. Azimi, D. Tam, L. Soares, and M. Stumm.”Managing shared L2 Caches on multicore systems in software”.In Workshop on the Interaction between Operating Systems and Computer Architecture, Held in junction with 2007 International Symposium on Computer Architecture (ISCA-34), San Diego, CA, USA, June 2007.

11. J. Cong, K. Gururaj, and G. Han. “Synthesis of reconfigurable high-performance multicore systems”.In proceeding of the ACM/SIGDA international symposium on Field programmable gate arrays, February 2009.

12. R. Azimi, D. Tam, L. Soares, and M. Stumm. “Enhancing operating system support for multicore processors by using hardware performance monitoring”. In SIGOPS Operating Systems Review (OSR), Special Issue on the Interaction Among the OS, Compilers, and Multicore Processors, April 2009.

13. Teng-Feng Yang,Chung-Hsiang Lin,Chia-Lin Yang.”Cache-aware task scheduling on multi-core architecture”.In the proceedings of VLSI Design Automation and Test (VLSI-DAT), 2010 International Symposium on April 2010.

14. Yan Wang,Lida Huang,Renfa Li,Rui Li. “A Shared cache-aware hybrid real-time scheduling on multicore platform with hierarchical cache”.In the proceeding Parallel Architectures, Algorithms and Programming (PAAP), 2011 Fourth International Symposium on December, 2011.

15. A. Snavely,D. Tullsen,and ,and G. Voelker.Symbiotic Jobscheduling with Priorities for a Simultaneous Multithreading Processor. In the Proceedings of the 2002 ACM SIGMETRICS international conference on Measurement and modeling of computer systems,University of California at San Diego.

16. Matthew DeVuyst, Rakesh Kumar, and Dean M. Tullsen. “Exploiting unbalanced thread scheduling for energy and performance on a CMP of SMT processors”.In the Proceedings of the international parallel and distributed processing Symposium 2006,University of California, San Diego.

17. Rohit Jain. “Soft real-time scheduling on a simultaneous multithreaded processor “.Thesis Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science in the Graduate College of the University of Illinois at Urbana-Champaign, 2002.

18. Stijn Eyerman,Lieven Eeckhout. “Probabilistic job symbiosis modeling for SMT processor scheduling”.In the Proceedings of the fifteenth edition of ASPLOS 2010,ELIS Department, Ghent University, Belgium.

19. A. Snavely,N. Mitchell,L. Carter,J. Ferrante. “Exploration in symbiosis on two multithreaded architectures” , Research at IBM in the year 1999.

20. S.Eyerman,L.Eeckhout.“Per-Thread cycle accounting in SMT processors”. In the Proceedings of ASPLOS 2009,ELIS Department,Ghent University.

21. S. Zhuravlev,S. Blagodurov,A. Fedorova.”Addressing shared resource contention in multicore processors via scheduling”.In the Proceedings of ASPLOS 2010,School of Computing Science, Simon Fraser University.

22. Yuejian Xie,Gabriel H. Loh. “Dynamic classification of program memory behaviors in CMPs”.In the Proceedings of 2008 CMP-MSI.

23. J. Anderson and A. Srinivasan. “Mixed Pfair/ERfair scheduling of asynchronous periodic tasks”. Journal of Comp. and Sys. Sciences, 68, 2004.

24. J. Anderson and A. Srinivasan. “A new look at pair priorities”. Technical Report TR00-023, University of North Carolina at Chapel Hill, Sept. 2000.

25. “Pin2.13UserGuide”.

26. “SPECjvm2008 User's Guide”.

27. Stijn Eyerman,Lieven Eeckhout.”The Benefit of SMT in the Multi-Core Era:Flexibility towards Degrees of Thread-Level Parallelism”.In the Proceedings of 19th international conference ASPLOS ,March 2014,NY,USA.

28. Bagher Salami,Mohammadreza Baharani,Hamid Noori ."Proactive Task Migration with a Self-Adjusting Migration Threshold for Dynamic Thermal Management of Multi-Core Processors",The Journal of Supercomputing, 2014 – Springer.