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

Volume 5 Issue 5

May 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

THE SCIENCE AND INFORMATION ORGANIZATION

www.thesai.org | info@thesai.org

OAIster getCITED Google Scholar BETA BASE Bielefeld Academic Search Engine ULRICHSWEB GLOBAL SERIALS DIRECTORY arXiv.org

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS IET InspecDirect INDEX COPERNICUS INTERNATIONAL WorldCat Window to the world's libraries Microsoft Academic Search EBSCO HOST Research Databases

Editorial Preface

From the Desk of Managing Editor…

It is our pleasure to present to you the May 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 5 May 2014
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

- **Abbas Karimi**
  Islamic Azad University Arak Branch
- **Abdel-Hameed Badawy**
  Arkansas Tech University
- **Abdelghni Lakehal**
  Fsdm Sidi Mohammed Ben Abdellah University
- **Abeer Elkorny**
  Faculty of computers and information, Cairo University
- **ADEMOLA ADESINA**
  University of the Western Cape, South Africa
- **Ahmed Boutejdar**
- **Dr. Ahmed Nabih Zaki Rashed**
  Menoufia University
- **Aderemi A. Atayero**
  Covenant University
- **Akbar Hossin**
- **Akram Belghith**
  University Of California, San Diego
- **Albert S**
  Kongu Engineering College
- **Alcinia Zita Sampaio**
  Technical University of Lisbon
- **Ali Ismail Awad**
  Luleå University of Technology
- **Alexandre Bouënard**
- **Amitava Biswas**
  Cisco Systems
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
  Maranatha Christian University, INDONESIA
- **Anirban Sarkar**
  National Institute of Technology, Durgapur, India
- **Anuranjan misra**
  Bhagwant Institute of Technology, Ghaziabad, India
- **Andrews Samraj**
  Mahendra Engineering College
- **Arash Habibi Lashakri**
  University Technology Malaysia (UTM)
- **Aris Skander**
  Constantine University
- **Ashraf Mohammed Iqbal**
  Dalhousie University and Capital Health
- **Ashok Matani**
- **Ashraf Owis**
  Cairo University
- **Asoke Nath**
  St. Xaviers College
- **Ayad Ismaeel**
  Department of Information Systems Engineering-Technical Engineering College-Erbil / Hawler Polytechnic University, Erbil-Kurdistan Region- IRAQ
- **Babatunde Opeoluwa Akinkunmi**
  University of Ibadan
- **Badre Bossoufi**
  University of Liege
- **Basil Hamed**
  Islamic University of Gaza
- **Bharti Waman Gawali**
  Department of Computer Science & information
- **Bhanu Prasad Pinnamaneni**
  Rajalakshmi Engineering College; Matrix Vision GmbH
- **Bilian Song**
  LinkedIn
- **Brahim Raouyane**
  FSAC
- **Brij Gupta**
  University of New Brunswick
- **Bright Keswani**
  Suresh Gyan Vihar University, Jaipur (Rajasthan) INDIA
- **Constantin Filote**
  Stefan cel Mare University of Suceava
- **Constantin Popescu**
  Department of Mathematics and Computer Science, University of Oradea
- **Chandrashekhar Meshram**
  Chhattisgarh Swami Vivekananda Technical University
- **Chao Wang**

CONTENTS

Integrating Android Devices into Network Management Systems based on SNMP

Fernando Hidalgo
Escuela de Computación
Central University of Venezuela
Caracas, Venezuela

Eric Gamess
Laboratorio de Comunicación y Redes
Central University of Venezuela
Caracas, Venezuela

***Abstract*—Mobile devices are becoming essential for today life. In developed countries, about half of the people have a smartphone, resulting in millions of these electronic devices. Android is the most popular operating system for smartphones and other electronic devices such as tablets. Hence, for network administrators, it is essential to start managing all the Android based devices. SNMP is the de facto standard for network administration, where agents that are running in managed devices are polled by management stations. Some primitive tools have already been developed to transform an Android device as a basic management station. However, so far, there is no SNMP agent for this operating system. In this paper, we develop the first SNMP agent for Android. We also propose an SNMP benchmark to study the SNMP traffic that can be supported by our SNMP agent over some real and actual Android devices. The results obtained show that it is realistic to integrate mobile Android devices in network management systems since they can handle a high number of SNMP requests in a reasonable period of time.**

Keywords—Network Management Systems; SNMP; Android; Performance Evaluation; Benchmarks

I. Introduction

All over the world, millions of cell phones have been sold. In developed countries, smartphone penetration among cell phone users are around 45% in US and 50% in Europe, resulting in a huge number of new devices with a high processing power.

Even if the majority of smart phones and tablets are owned and operated by people, for an organization it can be very useful to integrate these devices in its network administration system. That is, the organization could manage the smart mobile devices that are used for its operation, and the smart mobile devices of its employees with their consent.

The industry standard for network administration is called Simple Network Management Protocol (SNMP) [1][2], and has been around for more than two decades now. SNMP is based on the client/server model, where clients are also referred as network management stations and servers are agents running in the administrated devices. In SNMP, management stations regularly poll agents for information. Even though SNMP is widely spread, it is still rare in these new mobile devices. Most of the SNMP applications proposed for mobile devices are small management tools that allow a limited monitoring of classical network devices such as servers, switches, or routers from the mobile devices.

There are several Operating Systems (OSs) proposed for mobile devices, such as Symbian [3] from Nokia, BlackBerry OS [4] from BlackBerry Limited, iOS [5] from Apple, Windows Phone [6] from Microsoft, and Android [7][8] from Google. Recently, Android has taken the most important part of the market with this OS installed in more than 70% of the smartphones shipped. To the best of our knowledge, there is no SNMP agent for Android, making it impossible for network administrators to incorporate these devices to their monitoring systems. In this work, we develop an SNMP agent for Andriod. It implements many of the SNMP objects and can be used to integrate Android mobile devices to network management systems. To study the SNMP traffic that can be handled by our SNMP agent in real Android based devices, we also develop an SNMP benchmarking tool to get two important metrics (Response Time and Reply Request Ratio). The results obtained show that it is realistic to integrated mobile Android devices in network administration systems since they can handle a high number of SNMP requests in a reasonable period of time.

The rest of this paper is structured as follows. Section II presents the related work. In Section III, we introduce SNMP, the de facto network management protocol. Android is shortly described in Section IV. The development of our SNMP agent is discussed in Section V. Section VI is focused on evaluating the SNMP traffic that can be handled by our agent in real life Android devices. Finally, the paper concludes in Section I.

II. Related Work

Despite of its popularity among developers and users, just a few applications have been released for Android in the field of network administration based on SNMP. SNMP MIB Browser, created by ZOHO Corporation, can be downloaded for free from Google Play, formerly known as the Android Market. Google Play is a distribution platform for applications for the Android operating system operated by Google. SNMP MIB Browser enables users to browse/view the MIB data of SNMP enabled network devices such as servers, switches, routers, etc., from an Android based device. Users can load any standard MIB and fetch values from the managed devices to show the MIB data in an intuitive manner, just by clicking an object, a group of objects, or a table. It supports all the versions of SNMP (v1, v2c, v3). In the case of SNMPv3 [9], it implements MD5 and SHA algorithms for authentication, and DES, 3DES, AES-128, AES-192 and AES-256 algorithms for encryption.

SNMP Manager is similar to SNMP MIB Browser. That is, it is another basic application for Android that allows users to browse/view MIBs. Its support for SNMPv3 is actually in development. However, it has additional features such as trap reception and transmission. It can be freely downloaded from Google Play.

ezNetScan[1] is a free application developed by VR Software Systems Pvt Ltd from India, for network administration for mobile devices. It has been ported to Android and can be downloaded from Google Play. ezNetScan can display basic information of WiFi networks, scan the WiFi networks the Android device is connected to, and display information about the other connected devices, reporting their IP addresses, hostnames, MAC addresses, and more. Through SNMPv1 or SNMPv2c, ezNetScan can do basic operations and collect information about monitored devices such as operating system, file system capacity, installed applications, and running processes. Other features of ezNetScan include TCP port scanning, ping, and traceroute.

SNMP Traffic Grapher is a basic application that can be downloaded from Google Play. It allows real-time graphing of two SNMP Object Identifiers (OIDs) at the same time. The idea is to monitor the download and upload streams that transit the interface of a classical network device such as a server, switch, or router. The first OID will be graphed in green (download) and the second in blue (upload). The application retrieves the SNMP data with SNMPv2c.

SNMP Trap Agent[2] is a commercial product developed by Maildover LLC that provides a limited monitoring of Android phone's performance (percentage of busy CPU), usage (battery charge level, percentage of free memory), and location (latitude and longitude). It uses SNMP traps to send the selected information to the management stations. Even if it allows some flexibility for administration, it is not an SNMP agent since it is restricted to SNMP traps and cannot respond to SNMP get- and set-requests (GetRequest, GetNextRequest, and SetRequest).

As reported in this section, most of the development done so far is limited to SNMP managers, i.e., applications to monitor network devices. Unlike this previous work, our work bring the first implementation of an SNMP agent to Android.

III. Simple Network Managment Protocol

Simple Network Management Protocol [10][11] (SNMP) is a protocol for network management defined by the Internet Engineering Task Force (IETF) that is widely used since it is simple and easy to implement. SNMP is an application layer protocol that facilitates the exchange of management information between agents (managed devices) and Network Management Systems (NMSs). NMSs are also called managers. It is part of the TCP/IP protocol suite and uses User Datagram Protocol (UDP) as a transport protocol. Agents listen to queries on UDP port 161, while NMSs received traps on UDP port 162.

SNMPv1 specifies five core Protocol Data Units (PDUs): GetRequest, GetNextRequest, SetRequest, GetResponse and Trap. GetRequest is sent by a manager to retrieve the value of some objects managed by an agent. GetNextRequest is used iteratively by a manager to get tables or subtrees from administrated systems such as the Address Resolution Protocol (ARP) cache, or the routing table. SetRequest is used by a manager to modify an object in a managed device. GetResponse is sent by agents to respond with data to get-requests (GetRequest, GetNextRequest) and set-requests (SetRequest). Trap is used by agents to report an alert or other asynchronous events to managers. SNMPv1 does not allow manager-to-manager interactions as SNMPv2c and SNMPv3 do.

SNMPv2c is a revised version of SNMPv1 and includes improvements in the areas of performance, manager-to-manager communications, and error-handling. Three new PDUs were added in SNMPv2c: GetBulkRequest, InformRequest, and Report. The purpose of GetBulkRequest is to request the transfer of a potentially large amount of data including, but not limited to, the efficient and rapid retrieval of large tables. Compared to GetNextRequest, GetBulkRequest minimizes the number of requests and responses necessary to complete the transfer. InformRequest is sent by a manager to provide management information to a remote manager. Usage and precise semantics of Report are not specified in [2]; therefore, any SNMP administrative framework making use of this PDU must define it. The SNMPv2c improved error-handling includes expanded error codes that distinguish different kinds of error conditions; these conditions are reported through a single error code in SNMPv1.

A Management Information Base [11] (MIB) is a formal description of a set of network objects that can be managed using SNMP. Standard minimal MIBs have been defined (MIB-I, MIB-II, Host Resources MIB, etc), and vendors often have private enterprise MIBs. MIB-I [12] was defined to manage TCP/IP-based internets. MIB-II, defined in [13], is basically an update of MIB-I. Another fundamental concept of SNMP is the notion of Object Identifiers (OIDs). An OID is a tag that allows a management entity to refer unambiguously to a particular object. OIDs are allocated in a tree fashion and described in the MIB. The value of the OID is a sequence of integers that refers to a particular traversal of the object tree.

IV. Android Operating System

Android Inc. was founded in Palo Alto, California in October 2003 to initially develop an advanced operating system for digital cameras. However, with the goal of reaching a bigger market, Android Inc. diverted its efforts to produce a smartphone OS to rival those of Nokia, BlackBerry Limited, Apple, and Microsoft. The first version of Android was unveiled in November 2007. It is based on the Linux kernel, and designed primarily for touchscreen mobile devices such as smartphones and tablets. It is also used in televisions, games consoles, digital cameras, and many other electronic devices.

On August 2005, Google acquired Android Inc. Part of the success of Android is due to its license. Android source code is released by Google under the Apache License, which allows the software to be freely modified and distributed by device manufacturers, wireless carriers, and the community.

[1] http://www.eznetscan.net
[2] http://www.maildover.com/eurotrap.html

The user interface of Android is based on direct manipulation, using touch inputs that loosely correspond to real-world actions, like swiping, tapping, pinching and reverse pinching to manipulate on-screen objects. Internal hardware such as accelerometers, gyroscopes and proximity sensors are used by some applications to respond to additional user actions, for example adjusting the screen from portrait to landscape depending on how the device is oriented.

As of May 2011, Android has become the leader of the market for mobile OSs, having the largest installed base in almost all the countries of the world. More than one million applications for Android are available to be downloaded from Google Play. Also, it is by far the most popular platform for mobile OSs developers. For all the previous reasons, we decided to develop an SNMP agent for this well accepted architecture.

V. SNMP Agent For Android

Our SNMP agent runs in background and was developed by extending class `Services` from the Android Application Programming Interface (API). It is configured through a GUI (Activity) and some of the parameters that can be set include: the version of SNMP, the read-only community, the read-write community, and the UDP port where the agent is listening (see Figure 1). Users can start and stop the service through a simple switch in the GUI.

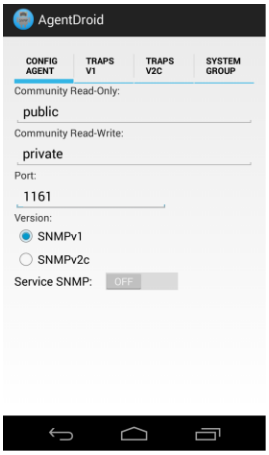


Fig. 1. Interface to Configure the Android SNMP Agent

For the implementation of SNMP, we used a Java package called `SNMP Package`[3]. Our agent supports both, SNMPv1 and SNMPv2c. We implemented all the SNMP messages: GetRequest, GetNextRequest, GetBulkRequest, SetRequest, GetResponse, Trapv1, and Trapv2c. We also added a module for the configuration of some important OIDs of the System Group, such as `sysContact`, `sysLocation`, and `sysName` (see Figure 2).

We implemented all the OIDs of MIB-II in our agent. It also supports part of Host Resources MIB [14], such as the Installed Software Group (hrSWInstalledTable), allowing the

Fig. 2. Interface to Configure Important Objects of the System Group

monitoring of the actual version of the installed applications or the installation of new software. The storage of read-write OIDs is done in a database kept in the system.

We made exhaustive tests to validate our SNMP agent for Android using SNMP JManager [15], a famous Java based manager that we developed a few years ago, and freely available from SourceForge[4]. We also captured the traffic with Wireshark [16] to validate the PDUs generated by our agent.

VI. Performance Evaluation

To evaluate the performance of our SNMP agent for Android, we decided to test its capacity over different Android based devices. Two metrics are important when evaluating the performance of an SNMP agent: response time and Reply Request Ratio (RRR). Response time is the time elapsed between sending an SNMP request (GetRequest, GetNextRequest, or SetRequest) and receiving the corresponding reply (GetResponse). RRR is the ratio between the number of replies received and the number of requests sent. This last metric is very useful when evaluating the number of requests that can be handled in one second or to study the behavior of the agent in stressed conditions. To the best of our knowledge, there is no tool developed so far to evaluate the performance of an SNMP agent. Hence, we developed our own benchmarking tool.

For the response time, it is important to report a meaningful time, i.e., a time that is not based on the sending of a unique SNMP request and the reception of its associated reply. A response time based on a unique SNMP request will have an important error, due to (1) clock precision and (2) the possibility of heavy OS processes being executed during the test, such as disk swapping, which will alter significantly the final results. Hence, our benchmark to measure the response time is based on the client/server model. Basically, an SNMP request (GetRequest, GetNextRequest, or SetRequest) with a fixed number of OIDs is sent from the benchmarking tool (client) to the agent (server) a number of times.

[3] http://jsevy.com/snmp

[4] http://sourceforge.net/projects/snmpjmanager

As soon as the reply arrives to the benchmarking tool, the next request is sent, so there is no idle time introduced by the benchmarking tool between two consecutive requests. We take a timestamp before and after the interchange. The difference of the timestamps is divided by the number of time the SNMP request is done, to get the average response time. For the response time test, our benchmarking tool has two parameters: (1) the list of OIDs to be fetched or modified, and (2) the number of time the request must be done before obtained the total response time.

In the case of the RRR test, our benchmarking tool has three parameters: (1) the list of OIDs to be fetched or modified, (2) the frequency of the requests or the number of requests that must be done in a second, and (3) the duration of the experiment.

For a better explanation, let say that the frequency is 5 and the duration is 30 seconds. Therefore, the benchmarking tool will be sending 5 requests every second, during 30 seconds, i.e., a total of 150 requests will be sent during the experiment. The benchmarking tool will also count the number of received replies, which must be less than or equal to 150. At the end of the experiment, the RRR is computed by dividing the number of replies received by the number of requests sent.

We tested the performance of our agent with two different Android devices: (1) a Galaxy Nexus i9250 phone and (2) a Galaxy Tab 8.9 tablet. We chose these devices since they are widely spread all over the world. We also used some PCs to run the benchmarking tools. The specifications of the devices are in Table I. We developed our benchmarking tool with Java, so we can run it in all the platforms that have a Java Virtual Machine (JVM).

TABLE I. SPECIFICATIONS OF DEVICES USED IN EXPERIMENTS

	Device		
	Phone	***Tablet***	***PC***
Brand	Samsung	Samsung	Hewlett Packard
Model	Nexus i9250	Tab 8.9	HP xw4600
Processor	Dual-Core 1.2 GHz	Dual-Core 1.0 GHz	Core 2 Duo 2.6 GHz
RAM	1 GB	1 GB	4 GB
OS	Android v4.3	Android v3.0	Windows 7 Pro

A. *Results for the Response Time*

Fig. 3 depicts the response time obtained with the Nexus i9250 phone. We have three curves representing the response time for GetRequest, GetNextRequest, and SetRequest messages.

We varied the number of OIDs in the requests from 1 to 10. As expected, the response time of the GetRequest is the smallest, since we only fetch the value of the OIDs in the Android device. GetNextRequest messages first search for the next OID and then fetch its value. SetRequest messages modify the value of the OID, and modifying is usually longer than fetching.

The SetRequest experiments have a better performance than the GetNextRequest when the SNMP messages have a small number of OIDs, and the behavior is inverted for higher number of OIDs. Also note that all the three curves are increasing with the number of OIDs as expected, with SetRequest increasing with the fastest rate.

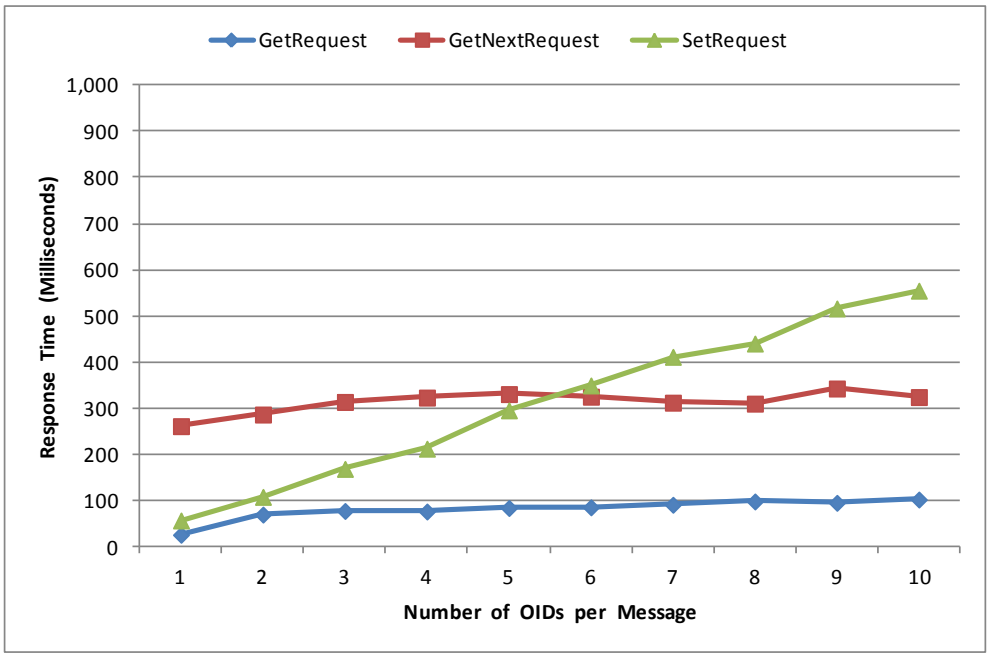


Fig. 3. Response Time for Different SNMP Messages in the Nexus i9250 Phone

Fig. 4 shows the response time obtained with the Tab 8.9 tablet. We have three curves representing the response time for GetRequest, GetNextRequest, and SetRequest messages. We varied the number of OIDs in the requests from 1 to 10. The results obtained are similar to the one of the Nexus i9250 phone (see Fig. 3). However, the response time is a little

smallest (better in this case) for the phone since it has a better processor than the tablet.

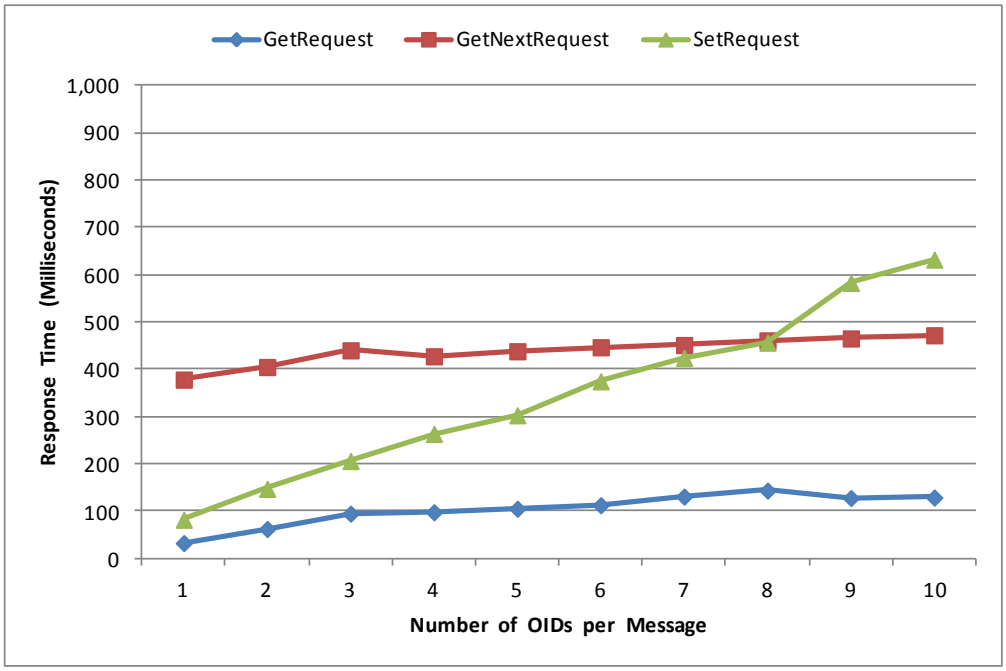


Fig. 4. Response Time for Different SNMP Messages in the Tab 8.9 Tablet

Results for the Reply Request Ratio

Fig. 5, Fig. 6, and Fig. 7 show the results obtained for the RRR, for GetRequest messages with 1, 3, and 6 OIDs, respectively. We varied the number of GetRequest messages sent by second according to the following values: 1, 2, 4, 6, 8, 10, 12, 14, and 16.

For each value of the number of GetRequest messages sent by second, we have six bars. The first two bars represent the RRR for the Nexus i9250 phone and the Tab 8.9 tablet, respectively, with one PC sending GetRequest messages. The following two bars (third and fourth) represent the RRR for the Nexus i9250 phone and the Tab 8.9 tablet, respectively, with two PCs sending GetRequest messages at the same time. In other words, the Android device receives twice the number of requests in one second, from two different sources. The last two bars (fifth and sixth) represent the RRR for the Nexus i9250 phone and the Tab 8.9 tablet, respectively, with three PCs sending GetRequest messages at the same time. In other words, the Android device receives three times the number of requests in one second, from three different sources. We can observe from these figures than the Nexus i9250 phone has a better RRR than the Tab 8.9 tablet, as expected, since it has a better processor. The idea of these experiments is to stress the Android devices and to see how much SNMP GetRequest traffic can be handled by our agent before getting saturated.

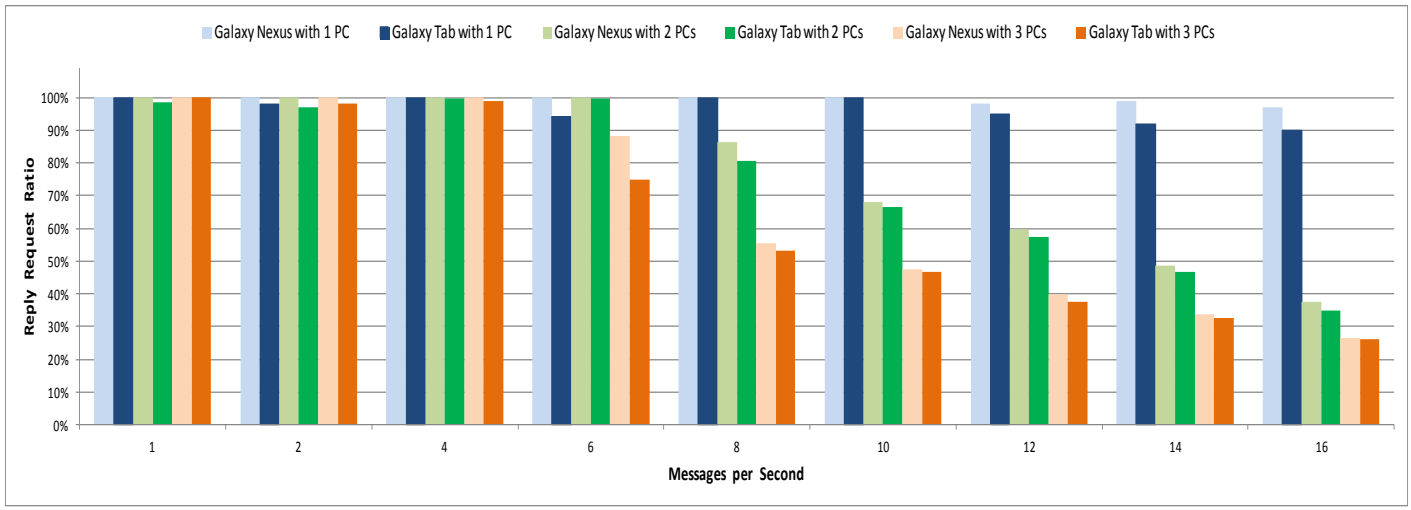


Fig. 5. RRR for GetRequest Messages with 1 OID

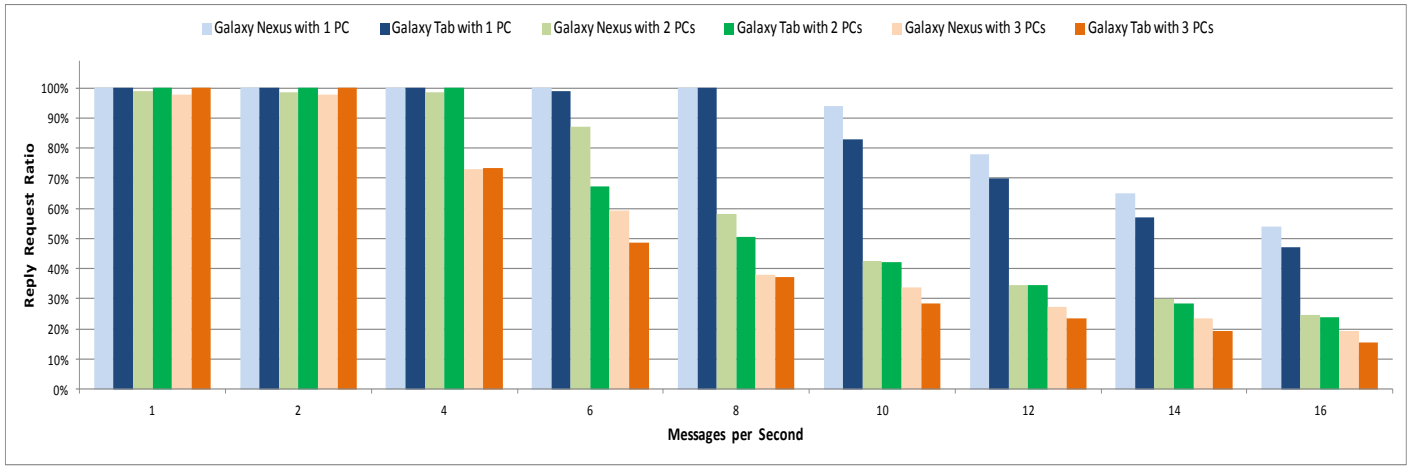


Fig. 6. RRR for GetRequest Messages with 3 OIDs

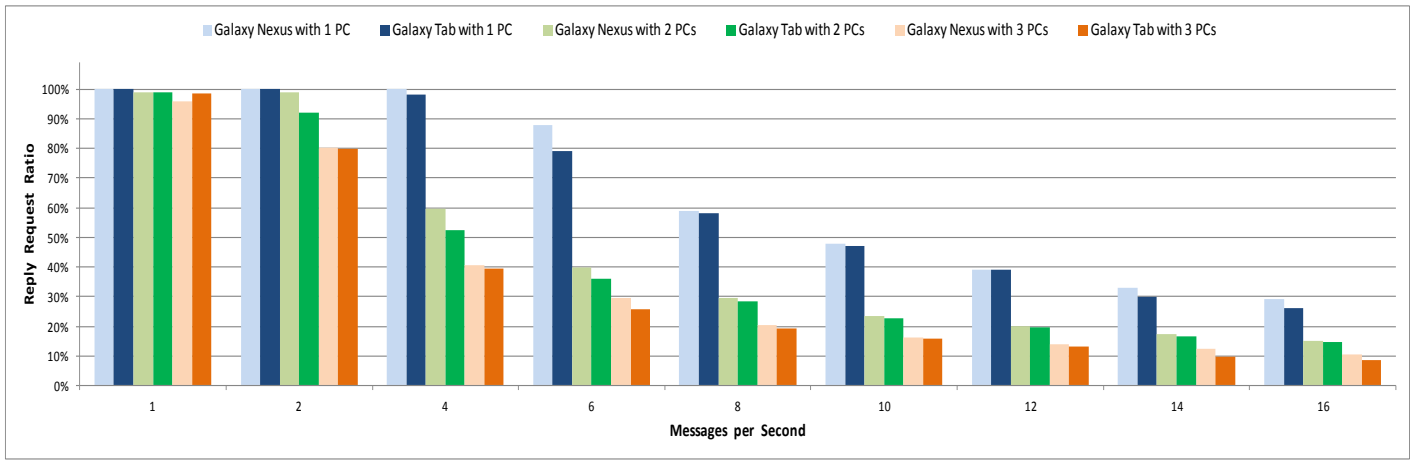


Fig. 7. RRR for GetRequest Messages with 6 OIDs

Fig. 8, Fig. 9, and Fig. 10 show the results obtained for the RRR, for GetNextRequest messages with 1, 3, and 6 OIDs, respectively. We varied the number of GetNextRequest messages sent by second according to the following values: 1, 2, 4, 6, 8, 10, 12, 14, and 16. Similarly to the GetRequest experiments, we can observe from these figures than the Nexus i9250 phone has a better RRR than the Tab 8.9 tablet, as expected, since it has a better processor. That is, the Nexus i9250 phone can handle a bigger number of SNMP requests before been saturated.

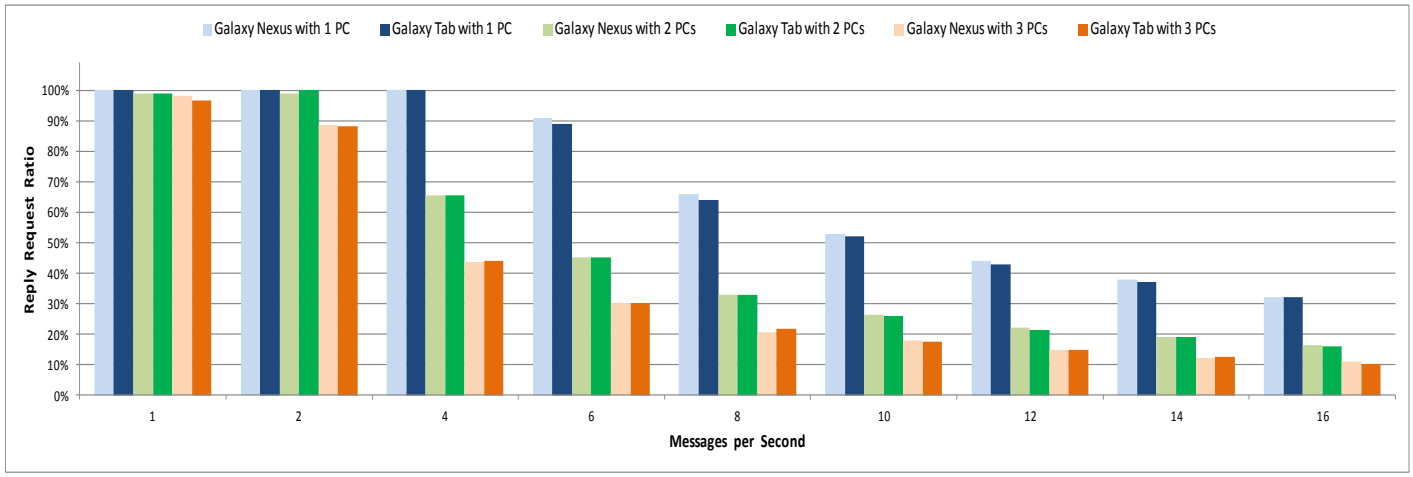


Fig. 8. RRR for GetNextRequest Messages with 1 OID

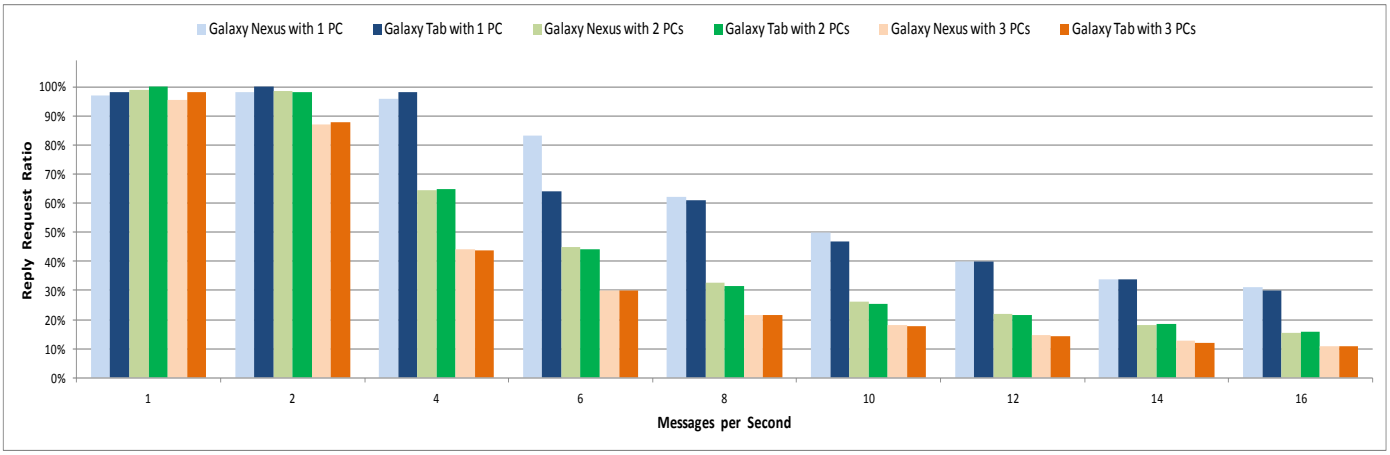


Fig. 9. RRR for GetNextRequest Messages with 3 OIDs

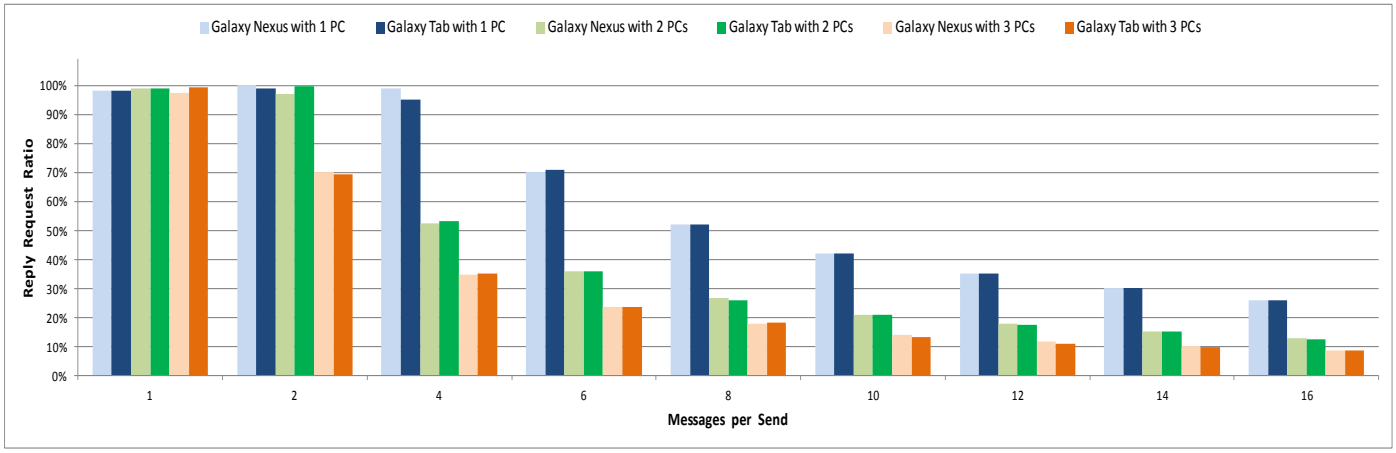


Fig. 10. RRR for GetNextRequest Messages with 6 OIDs

Fig. 11 shows the result of the RRR when varying the number of SNMP requests sent by second to the Nexus i9250 phone from 3 PCs, by using the following values: 1, 2, 4. 6, 8, 10, 12, 14, and 16 requests per seconds. We have 3 curves for GetRequest, GetNextRequest, and SetRequest messages. As expected, GetRequest has the best performance of the requests. GetNextRequest and SetRequest are showing similar results.

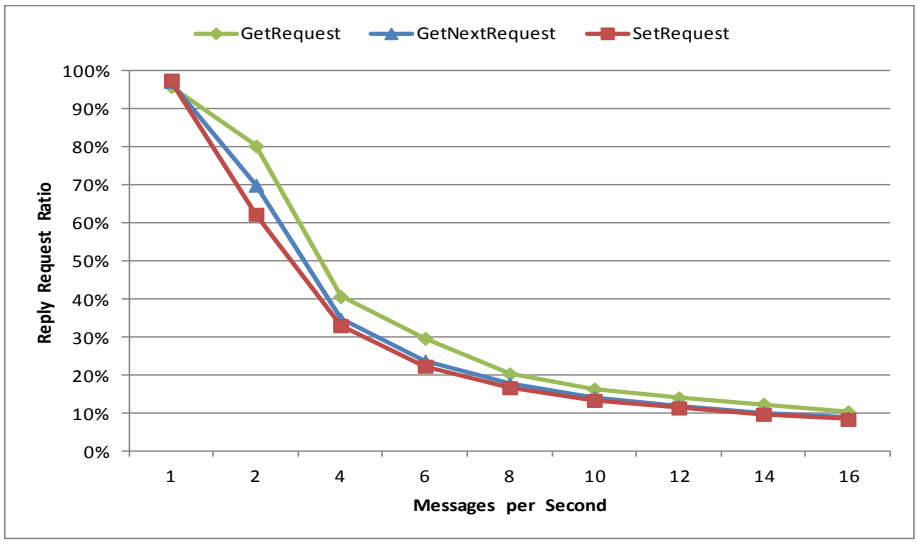


Fig. 11. RRR for GetRequest, GetNextRequest, and SetRequest Messages Sent to the Nexus i9250 Phone with 3PCs and 6 OIDs

Fig. 12 shows the results of the RRR when varying the number of SNMP requests sent by second to the Tab 8.9 tablet from 3 PCs, by using the following values: 1, 2, 4. 6, 8, 10, 12, 14, and 16 requests per seconds. We have 3 curves for GetRequest, GetNextRequest, and SetRequest messages. The results are similar to the ones obtained for the Nexus i9250 phone (see Fig. 11). However, the RRR is better in the case of the Nexus i9250 phone, due to the better processor. These results also show that common Android devices can manage a high volume of SNMP request in a short period of time.

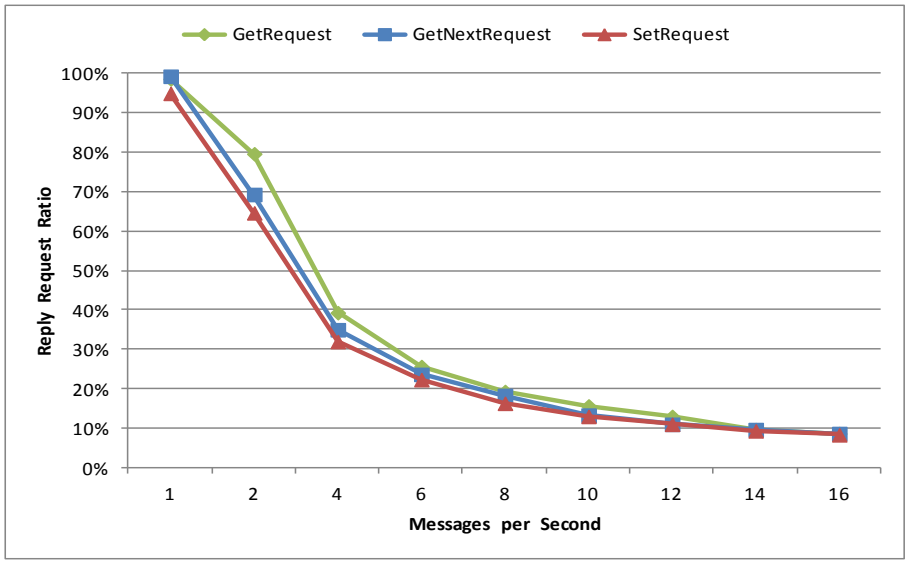


Fig. 12. RRR for GetRequest, GetNextRequest, and SetRequest Messages Sent to the Tab 8.9 Tablet with 3PCs and 6 OIDs

VII. Conclusions and Future Work

Million of Android devices have been sold all over the world, making Android the favorite mobile OS. Due to the growing interest in these devices, changes in network administration to integrate them in NMSs are required. Since SNMP is wildly spread in NMSs, an SNMP agent for Android is the first step for the inclusion of Android devices in monitoring systems. In this paper, we presented the first implementation of an SNMP agent for Android. Our agent has support for SNMPv1 and SNMPv2c. With our agent, users can retrieve almost all the OIDs of MIB-I, MIB-II, and Host Resources MIB.

To validate the agent and to show that the actual Android mobile devices have enough power to be integrated in a management system, we also proposed a performance benchmarking tool for SNMP. Our tests showed that common Android devices (Nexus i9250 phone and Tab 8.9 tablet) can be integrated in NMSs, since they can handle a high number of SNMP request in one second.

As future work, we plan to extend our agent to support SNMPv3 [9]. Today, authentication and privacy are important due to the numerous security threads in networks, especially in wireless networks where the radio waves spread in all directions.

Acknowledgment

We want to thank the CDCH-UCV (Consejo de Desarrollo Científico y Humanístico) which partially supported this research under grant number: PG 03-8066-2011/1.

References

[1] J. Case, M. Fedor, M. Schoffstall, and J. Davin, A Simple Network Management Protocol (SNMP), RFC 1157, May 1990.

[2] J. Case, K. McCloghrie, M. Rose, and S. Waldbusser, Introduction to Community-based SNMPv2, RFC 1901, January 1996.

[3] M. Wilcox, Porting to the Symbian Platform: Open Mobile Development in C/C++, 1st Edition, Wiley, November 2009.

[4] A. Ludin, Learn BlackBerry 10 App Development: A Cascades-Driven Approach, Apress, 1st Edition, March 2014.

[5] J. Conway, A. Hillegass, C. Keur, iOS Programming: The Big Nerd Ranch Guide, 4th Edition, Big Nerd Ranch Guides, February 2014.

[6] A. Whitechapel and S. McKenna, Windows Phone 8 Development Internals, Microsoft Press, July 2013.

[7] R. Meier, Professional Android 4 Application Development, Wrox, 3rd Edition, May 2012.

[8] P. Deitel, H. Deitel, and A. Deitel, Android for Programmers: An App-Driven Approach, Prentice Hall, 2nd Edition, January 2014

[9] W. Stallings, SNMP, SNMPv2, SNMPv3, and RMON 1 and 2, Addison-Wesley Professional, 3rd Edition, October 2013.

[10] J. Gateau, Extending Simple Network Management Protocol (SNMP) Beyond Network Management: A MIB Architecture for Network-Centric Service, Naval Postgraduate School, Monterey, California, March 2007.

[11] L. Walsh, SNMP MIB Handbook, Wyndham Press, March 2008.

[12] K. McCloghrie and M. Rose, Management Information Base for Network Management of TCP/IP-Based Internets, RFC 1156, May 1990.

[13] K. McCloghrie and M. Rose, Management Information Base for Network Management of TCP/IP-based Internets: MIB-II, RFC 1213, March 1991.

[14] S. Waldbusser and P. Grillo, Host Resources MIB, RFC 2790, March 2000.

[15] G. Ayala, P. Poskal, and E. Gamess, SNMP JManager: An Open Source Didactic Application for Teaching and Learning SNMP v1/2c/3 with Support for IPv4 and IPv6, In proceedings of the seventh Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2009), San Cristóbal, Venezuela, June 2009.

[16] L. Chappell and G. Combs, Wireshark Network Analysis: The Official Wireshark Certified Network Analyst Study Guide, 2nd Edition, Laura Chappell University, March 2012.

Acceptance Factors and Current Level of Use of Web 2.0 Technologies for Learning in Higher Education: a Case Study of Two Countries

Razep Echeng
University of the West of Scotland
Paisley, United Kingdom

Abel Usoro
University of the West of Scotland
Paisley, United Kingdom

***Abstract*—The use of Web 2.0 technology tools or social media in educational context is being emphasized in recent times in different parts of the world and this has brought about a significant increase in the number of educational institutions who are aware of their usefulness when either implementing them as a separate system or incorporating them into their learning management systems. However, there is little research on the acceptance and how much these tools are currently being used for learning hence the need for more empirical studies to investigate factors that would influence acceptance and increase the use of these technologies. The study developed hypotheses and a research model which was operationalized into a questionnaire administered to academics and students in Scotland and Nigeria. 317 responses were received from Nigeria and 279 from Scotland. Analysed data was used to validate the research model that is aimed at explaining acceptance and present level of use of Web 2.0 technology tools in learning environments.**

Keywords*—Web 2.0 technologies; acceptance factors; adoption for learning; collaboration; participation; Nigerian higher education; Scotland

I. INTRODUCTION

Acceptance of technology has long been a challenging issue in information systems research [1]. Understanding the reason why people accept or reject technology is very crucial because it serves as a guide to investors, manufacturers, institutions and their managers. Much research has used technology acceptance models (TAM) to measure acceptance of technology [2][3][4][5]. Whereas some research exist in developed countries on acceptance of Web 2.0 tools in learning, not much of such empirical studies has been done in developing economies. Neither has there been a comparative study of these economies. Hence, this study investigates Nigeria and Scotland as well as compares these two countries in terms of the factors of the model such as perceived usefulness, performance expectancy, motivation, ease of use, attitude, behaviour and actual use. The investigation also endeavours to ascertain how the model can be used to explain the acceptance and the use of Web 2.0 social network technology tools in teaching and learning in higher institutions of developed and developing communities. The rest of this paper presents literature review, method, data analyses, discussion, conclusions, implications and areas for future study.

II. LITERATURE REVIEW

A. Nigeria and Scotland

Few empirical studies exist in Nigeria that examined the use of Web 2.0 in learning [6][7]. These researches were interested in the use of Web 2.0 by librarians, academics and students in Nigeria. They found the use of these tools for academic purpose lacking and identified five major problems: personality characteristics, motivation, lack of facilities and lack of computer expertise. They also suggested more research into how these technologies can be adopted for teaching and learning.

Little research has been done on the acceptance of Web 2.0 tools in higher institutions in developed countries [8] and their research inferred that subjective norm of students affect their acceptance, and little has been researched on users' acceptance of Web 2.0 technology tools in learning in Scotland as found in some other developed communities [9][10], hence this research is a comparative study that seeks to bring together factors that influence acceptance and use of Web 2.0 technology tools in learning in order to understand the key factors that could be associated with adoption in these two educational communities.

B. Challenges of Adoption of Web 2.0 in Education

Literature has documented the challenge of getting students and educators to adopt Web 2.0 tools for educational purposes [11]. Some research explained that the limited adoption is due to lack of understanding of the behaviour of users thereby shifting focus from what users want to what is technologically achievable [12]. Though innovative educators appreciate and use Web 2.0 technologies, others are afraid that these technologies would disrupt young people's engagement with "traditional" education [13[12]. These challenges and debates on them have been noticed in higher education of developed economies [11]. However, these technologies are potentially useful in learning activities.

C. Potential of Web 2.0 technology tools

Over the past five to six years, there has been a significant increase in research on educational usefulness and potentials of Web 2.0 [14][15][16]. Most of them have shown that Web 2.0 social network tools can enhance participation, collaboration and interaction in learning. They enable social

networking site users who are mostly young people to create profiles and build personal networks that connect them to each other for a variety of professional and personal reasons. However, there is still the problem of acceptance and use for academic purposes by both students and educators [17][11] and this calls for investigation on constructs that influence such acceptance and use.

D. User Acceptance

User acceptance is exhibited in the willingness of a user or group of users to employ information technology tools for the tasks that they are designed to support. The acceptance of technology has been a challenging issue in information systems research for a long time and researchers have studied a range of issues related to technology acceptance, from individual user characteristics such as cognitive style to internal beliefs and their impact on user's behaviour. It is crucial to understand the reason why people accept or reject technology, because it can serve as a guide to investors, manufacturers, and institutions and for managerial intervention.

Existing research on user acceptance has produced a variety of explanatory and predictive models [18][2][4][8][20][11]. However, these models suggest different and, sometimes, conflicting sets of predictor variables. Ajjan's [8] study used the theory of planned behaviour and his findings inferred that subjective norm of students is a key factor that affects their acceptance whereas Armitage and Cornor [19] studied 185 researches that used the theory of planned behaviour (TPB) from 1980 until 1997 and found that subjective norm was a weak variable for predicting behavioural intention. Hence, the need to review other popularly used technology acceptance models in order to understand the major constructs that could contribute to acceptance and use of Web 2.0 technology in learning.

E. Technology Acceptance Models

Literature review revealed three widely used models of acceptance of technology and these are: theory of reasoned action [18]; technology acceptance model [2]; and unified theory of acceptance and use of technology [4]. Acceptance of technology has been studied in different contexts, but there is limited research on acceptances of Web 2.0 tools in teaching and learning in higher institutions [8] [20] hence, this research developed hypotheses to test acceptance of Web 2.0 technology tools in learning and to answer to the research question which is: What are the factors that would relate to acceptance and use of Web 2.0 tools for learning?

III. Hypotheses Development

A selection of constructs was made from three technology acceptance models that support learning: theory of reasoned action [18]; technology acceptance model [2]; and unified theory of acceptance and use of technology [4]. These constructs are: motivation to use, social factor, facilitating conditions, performance expectancy, ease of use and perceived usefulness. This selection was guided by the preliminary interviews with five ICT directors, five lecturers and 16 students in five Nigerian Universities and one university in Scotland.

Fifteen semi structured questions were used to investigate the situation on learning with Web 2.0 technology tools and the possible motivating factors that could be used to enhance the use of these tools in learning. These interviews were analyzed using NVIVO tag cloud in order to retrieve text that were mostly used by the respondents and the clustering co-occurrence or non-occurrence to determine important constructs to be included [21]. This was done to support the inclusion of constructs from the three models and also this analysis suggested the addition of a new construct (prior knowledge) in the hypothesis development. The hypotheses are presented in the rest of this section.

A. Perceived usefulness (PU)

Perceived usefulness is the belief that the use of technology will improve and progress the work or learning activity of an individual or an organization. Research by Davis et. al.[2] and Venkatesh et. al [4] found that perceived usefulness affects technology acceptance. This research is to examine the effect of perceived usefulness with regards to Web 2.0 technologies for learning with the hypothesis:

H_1: There is a positive relationship between perceived usefulness and behavioural intention to adopt Web 2.0 technologies in learning.

B. Social Factor (SF)

Social factor in this context comes from the impact of social presence on individual behaviour. This could be communication and interaction with students and lecturers which may result in interpersonal agreements that affect behaviour of individuals in a group [23][24][25][8]. This factor was included in Davis et. al.'s [2] Model as an external factor, which they argued may influence technology acceptance. This variable is also included in UTAUT. This research seeks to validate this argument when considering Web 2.0 technologies acceptance for learning. Therefore:

H_2: Social factor has a positive relationship with behavioural intention to accept Web 2.0 technologies for learning.

C. Prior Knowledge (PK)

Prior knowledge can be described as knowledge of a set of circumstances gained in the past sufficient to make actions based on those circumstances. It is often helpful and very useful in learning environments [27][28]. This knowledge or experience could positively relate to acceptance of Web 2.0 technologies for learning, hence the following hypotheses:

H_3: Prior knowledge has a positive relationship with behavioural intention to adopt Web 2.0 technologies for learning.

D. Facilitating conditions (FC)

The access to internet facilities, the availability of good internet signals and the cost of broadband can be regarded as facilitating conditions for the use of Web 2.0 technologies for learning. Therefore, they may be related to the use of Web 2.0 technologies in higher education. Thus, it can be hypothesized that:

H_4: There is a positive relationship between facilitating conditions and behavioural intention to use Web 2.0 technologies in learning.

E. Perceived Ease of use (PEOU)

Perceived ease of use is the degree to which an individual believes that the use of technology will be without much effort, but will help to achieve much in a short time [2][26]. This has been used to predict acceptance of technology [2], and this research suggests that perceived ease of use should explain acceptance of Web 2.0 technology tools for higher education, hence the hypothesis:

H_5: There is a positive relationship between perceived ease of use and behavioural intention to adopt Web 2.0 technology tools in learning.

F. Performance Expectancy (PE)

Performance expectancy is the degree to which an individual or group of people expect to be proficient in their work or education when they are using technology. Ajjan and Harshone's [8] research found this variable as promoting technology acceptance. To investigate this finding in the case of Web 2.0 in learning in higher education, we used the hypothesis:

H_6: There is a positive relationship between performance expectancy and behavioural intention to use Web 2.0 technologies in learning.

G. Motivation to use (MTU)

Motivation in this context involves emotional support, internal or external support that stirs up a learner or gives the desire to act. Motivation can facilitate or hinder change in a learner [18][28][29]. Intrinsic and extrinsic motivation develops personal behaviour which can in turn affect evaluation of choice, goals and achievements. Thus, motivation to use Web 2.0 technology tools for learning is likely to relate to the attitude of the learners, and it should also be related to behavioural intention.

H_7: There is a positive relationship between motivation and intention to use Web 2.0 technologies for learning.

H. Behavioural intention (BI)

Ajzen and Fishbein [18] argued that a person's exhibition of a specific behaviour is determined by their behavioural intention. Behavioural intention to use Web 2.0 technology can relate with actual use. Thus the hypothesis:

H_8: Behavioural intention has a positive relationship with actual use of Web 2.0 technologies for learning.

A conceptual model was developed from the hypotheses that have been presented in this section (see Figure 1)

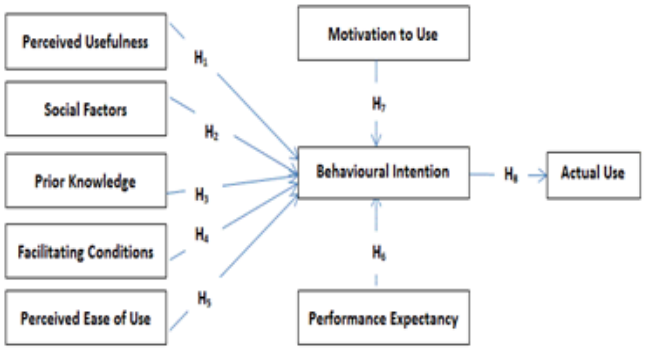


Fig. 1. Conceptual model

IV. Method

The nature of the research question and focus which are on the acceptance of Web 2.0 technologies guided the method adopted. The literature also revealed that most researchers approach similar studies using quantitative research [2][8]. This research operationalized the constructs (see Table 1) into a questionnaire to collect data that would measure the eight constructs in the model.

The questionnaire was divided into three parts: the first part measured students' level of satisfaction in learning and facilities available for teaching and learning; the second part measured the eight constructs in the research model (attitude to use, actual use, perceived usefulness, perceived ease of use, social factor, acceptance and performance. Then the third part investigated demographics (age, gender, educational level, faculty, having personal computer, and having internet access). Items were measured using 5 and 7 point Likert scales with 19 questions. All items in the questionnaire were adapted from earlier and similar research to suit this study [2][8].

A. Content Validation

The questions had strong theoretical basis and besides they were validated by allowing prospective participants to answer them in order to check whether the questions were understandable and answerable. The questionnaire was amended based on comments from these respondents [30].

B. Participants

Questionnaire was sent online to students and lecturers in one university in Scotland and participation was voluntary. 279 (78 lecturers and 201 students) responded. The Nigeria questionnaire was administered by lecturers and the researcher who visited five Institutions and collected 317 usable data from participants. Five universities were selected due to differences in Nigerian educational system.

Table I. QUESTIONNAIRE AND CONSTRUCTS

Constructs		Questions	Question No.
Perceived ease of use		How easy do you find using these Web 2.0 tools (listed in question 6) to obtain the resources you need for your studies?	7
Perceived usefulness		To what extent do you agree that Web 2.0 tools would speed up acquisition of knowledge?	12
		To what extent do you agree that Web 2.0 tools will encourage active participation in learning?	13
Actual use		How often do you use Web 2.0 tools for academic purposes per week?	8
Social		To what extent do you agree that the social part of e-learning platforms (e.g. Module and Blackboard) motivates learner to achieve learning objectives?	10b
Motivation		E-learning platforms enable you to send mails, download course materials upload assignments, read announcements, access the library material and discuss with other students, professionals and your lecturers. To what extent do you think such systems would motivate you to achieve your learning objectives?	10a
Facilitating condition		Regarding facilities available for learning and teaching in the university, how satisfied are you? Add any comments regarding conditions necessary to facilitate Web 2.0 in in learning.	4
Performance Expectancy		To what extent do you agree that the use of Web 2.0 technologies for learning will help to improve performance?	14
Prior knowledge		How often do you use Web 2.0 tools (e.g. blogs, Wikis, twitter) for social purposes per week?	6
Behaviour intention		To what extent do you agree that social computing should be adopted in higher education and training for sharing of knowledge and information?	11
Demographics	Gender	What is your gender?	16
	Status	Are you a student or lecturer?	1
	Field	What is your field?	19
	Age bracket	What is your age bracket?	17

V. DATA ANALYSIS

This study adapted the quantitative data analysis. Descriptive analysis with frequency tables and histograms was carried out to describe the general responses of each question. The model was tested for general validity.

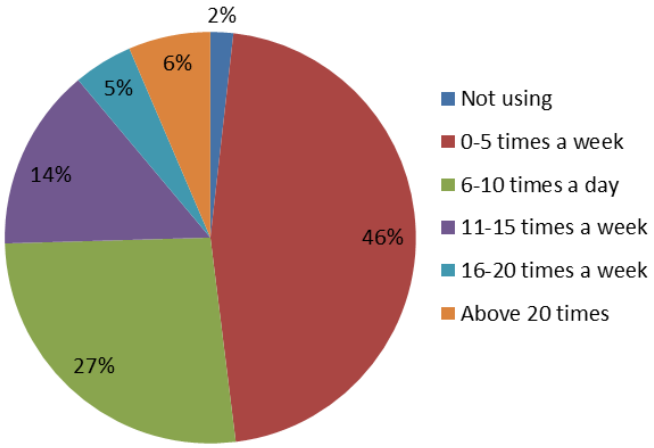


Fig. 2. Level of use in academic activities

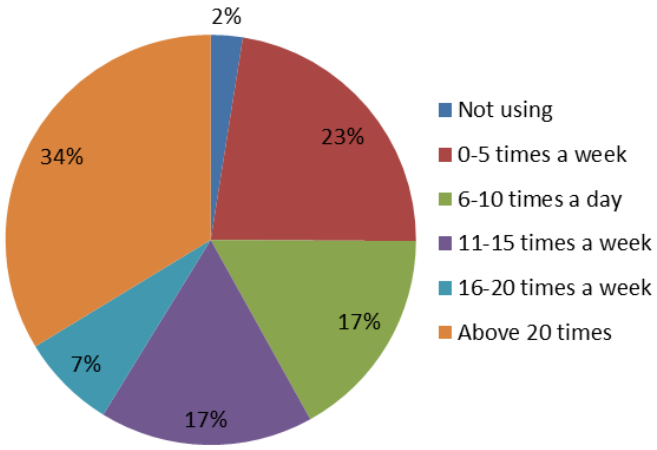


Fig. 3. Level of use for social activities

Inferential statistical analysis in the form of correlation analysis was used to evaluate the relationships between variables, therefore testing the hypotheses of this study (see the conceptual model at Fig. 1). The correlation formula is given as:

$$\rho X, Y = corr(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

where x is one variable, e.g. *motivation to use* and y another, e.g. *behavioural intention;* and $\rho X,Y$ is the correlation coefficient.

Kendall tau rank correlation coefficients were used because we do not have absolute values [30]. Tables 3 is a summary of relationships between variables and links the relationships to hypotheses presented previously in the model. Correlations marked with a single asterisk are significant at level 0.05 and those with double asterisks are significant at level 0.01. The absence of an asterisk indicates no correlation and this is the case in motivation to use and behavioural intention. The rest of this section will discuss each pair of variables before a general summary of the findings and implications are presented.

VI. Discussion

Table 3 is a correlation table from the data collected from Scotland and Nigeria, it shows that all the relationships examined in the model were significant except *motivation to use* in the Nigerian data.

This research validates the constructs: perceived usefulness, performance expectancy, perceived ease of use, prior knowledge, motivation, facilitating conditions, and social factor from TAM, UTAUT and TRA [2][8][18] and also validate the new construct prior knowledge. The correlation between *behavioural intention* (BI) and *perceived usefulness* (PU) in Scotland data and Nigeria is highly significant and reaches the value of .616 and .549, respectively. That means that there is a relationship between acceptance and usefulness in the case of Web 2.0 technologies. The rest of this section will discuss the relationships between BI and other variables.

Table II. General Correlation Between Behaviour Intention and Other Constructs

Dependent Variable	Independent Variable	Correlations Coefficients		Significance		Hypothesis
		Scotland	Nigeria	Scotland	Nigeria	
BI	TAM (PU)	.616**	.549**	Yes 0.01	Yes 0.01	H_1
BI	TAM, UTAUT (SF)	.674**	.520**	Yes 0.01	Yes 0.01	H_2
BI	Mine (PK)	.625**	.153*	Yes 0.01	Yes 0.05	H_3
BI	UTAUT (FC)	.130*	.115*	Yes 0.05	Yes 0.05	H_4
BI	TAM (PEoU)	.221**	.134*	Yes 0.01	Yes 0.05	H_5
BI	UTAUT (PE)	.620**	.431**	Yes 0.01	Yes 0.01	H_6
BI	MtU (TRA)	.290**	.932	Yes 0.01	No	H_7
AU	TAM, UTAUT (BI)	.155*	.189*	Yes 0.01	Yes 0.01	H_8

Figures 2 and 3 show the percentage of academic and social uses, respectively. We observe that the percentage of academic use per week is very low (0-5) for majority of users and social purpose use is as high as 20 times and above for majority of users.

Table 3 also shows the relationship between *Behavioural Intention* and *Performance Expectancy* as highly significant for Scotland and Nigeria (.620**, .431**) at 0.01 level of significance. This is in agreement with existing research [8] that there is a relationship between acceptance and performance expectancy. The correlation between *Behaviour Intention* and *Social Factor* is highly significant (.674** and .520** at 0.01 level of significance. This also agrees with previous research [8][2] meaning there is a relationship between social factor and intention to use Web 2.0 technologies for learning.

The correlation between *behavioural intention and actual use* is significant at 0.05 level in the two economies. This is in line with the technology acceptance model [2][8][33[34]. There is correlation between *behavioural intention* and *motivation* (MtU) in Scotland at 0.05 level of significance (see Table 3) and this agrees with the research by Ajzen and Fisnbein [18]. However, MtU is not significant in Nigeria and the reason was that learning management systems (LMS) are rarely available in Nigerian universities, hence the need for more research in this area when LMS are more available in Nigeria.

The relationship between *behavioural intention* and *perceived ease of use* is highly significant in Scotland at 0.01 level, but just significant at 0.05 level in Nigeria. However this agrees with other research [2][35] that ease of use influences behavioural intention, meaning that there is a relationship between perceived ease of use and behavioural intention. The table also shows that there is a significant correlation between *behavioural intention* and *facilitating conditions,* meaning that there is a relationship between these two variables and this agrees with other research [8][11][20].

On the whole, this research has validated the conceptual model (Figure 1), it agrees with other research [2][3][4][8][18][11][20] and also unveils a new construct, *prior knowledge,* which has a positive relationship with behavioural intention which should explain acceptance of Web 2.0 tools for learning.

VII. Conclusions, Implications and Areas For Further Study

This research examined user acceptance and adoption of Web 2.0 technology tools for learning among populations in Nigeria and Scotland. It aimed to give insight into the very low use of these tools and to proffer key related factors that should be borne in mind by policy makers and system developers who aim to encourage increased use of these tools in teaching and learning. The research has validated its conceptual model. It

has also revealed low motivation in Nigeria to the use these tools for academic purposes. A casual observation suggests that this is caused by inadequate provision of learning management systems (LMS) with social media tools. Thus, future research should investigate in Nigeria or a similar developing country motivation to use Web 2.0 tools when learning management systems (LMS) with social media are popular in such environments. Besides, a qualitative approach should be used to complement the quantitative findings of this research. It is also the intent of the researcher to set up an experimental use of LMS with social media in some of these institutions in Nigeria to further validate this work as well as discover implementation issues.

This work contributes to the body of knowledge on factors that affect acceptance and use of Web 2.0 social networking technology tools in teaching and learning. This will aid management decisions towards enhancing and improving educational experience as they consider the key variables validated in this research.

REFERENCES

[1] E. Swanson. (1994). Information Systems Innovation among Organizations, *Management Science,* Vol. 40, No. 9, 1069-1092.

[2] F. Davis, R. Bagozzi and P. Warshaw (1989). User Acceptance of Computer Technology: a comparison of two theoretical models. *Management Science*, *Vol*. 35, pp. 982 - 1003.

[3] V. Venkatesh and F. Davis, (2000*).* A Theoretical Extension of the Technology Acceptance Model: four Longitudinal Field Studies. *Management Science,* Vol. 46, No. 2, pp. 186-204.

[4] V. Venkatesh, M. Morris, G. Davis, and F. Davis, (2003). User acceptance of information technology: Toward a unified view, *MIS Quarterly*, Vol. 27(3) pp. 425–478.

[5] L. Oshiyanki, P. Cairns, and H. Thimbleby, (2007). Validating Unified theory of acceptance and use of Technology (UTAUT) tool cross-culturally. *British computer society*, Vol. 2 Proceedings of the 21st century BCS group conference.

[6] C. Anunobi and A. Ogbonna (2012). Web 2.0 Use by Librarians in a state in Nigeria, Developing Country Studies, Vol. 2, No. 5, pp. 7-66, 2012.

[7] R. Echeng (2011). The use of Web 2.0 in teaching and learning in Nigerian higher institution, An MSc Thesis submitted to the School of Computing, University of the West of Scotland.

[8] H. Ajjan and R Hartshorne (2008). Investigating faculty decisions to adopt Web 2.0 technologies: Theory and empirical tests, *The Internet and Higher Education*, Vol. 11, No 2, pp. 71–80, 2008.

[9] G. Baxter, M. Stansfield, and T. Connolly (2011). Introducing Web 2.0 in education: A structured approach adopting a Web 2.0 implementation framework. Next Generation Web Services Practices (NWeSP), 2011 7th International Conference on 19-21 Oct. 2011, 499 - 504.

[10] R. Echeng, A. Usoro and G. Majewski, (2013). Acceptance of Web 2.0 in Learning in higher education: an empirical study of a Scottish university, WBC July Conference Proceedings on E-learning.

[11] P. Jucevičienė, G. Valinevičienė,(2010). A Conceptual Model of Social Networking in Higher Education, Electronics and Electrical Engineering – Kaunas: Technology, vol. 6, no. 102, pp. 55–58.

[12] C. Ennew, A. Fernandez-Young (2006). Weapons of mass instruction? The rhetoric and reality of online learning, *Marketing Intelligence & Planning*, Vol. 24 Issue: 2, pp.148 - 157.

[13] J. Njenga and, L Fourie (2010). The myths about e-learning in higher education *British Journal of Educational Technology Vol*. 41 No. 2, pp. 199 – 212.

[14] C. Redecker, Review of Learning 2.0 practices, Study on the Impact of Web 2.0 Innovations on Education and Training in Europe, 2009 Available online https://europa.eu/Publications/pub.cfm?id=2059 [Accessed: 25-11-2011].

[15] T. Franklin and M. Van Harmelen (2007). Web 2.0 for content for learning and teaching in higher education, London: Joint Information Systems Committee, 2007. JISC www. jisc. ac. uk/media/documents/programmes/digitalrepositories/web2

[16] B. Alexander, Web 2.0: A new wave of innovation for teaching and learning? *EDUCAUSE Review*, Vol. 41, no. 2, pp. 32-44, 2006.

[17] T. Valtonen, S. Hacklin, S. Kontkanen, ,A. Hartikainen-Ahia, S. Kärkkäinen and Kukkonen, J. (2013). Pre-service teachers' experiences of using social software applications for collaborative inquiry, *Computers & Education* 69 (2013) 85 - 95.

[18] I. Ajzen and M. Fishbein, *Understanding attitudes and predicting social behavior*, Englewood Cliff, NJ: Prentice Hall, 1980.

[19] J. Armitage, and M. Connor (2010). Efficacy of theory of planned behaviour. British Journal of Social Psychology (2001), 40, 471–499.

[20] S. Mazman, and Y. Usluel, (2010). Modeling educational usage of Facebook. *Computers & Education*, Vol.55, pp. 444-453.

[21] Echeng, R., Usoro, A., Majewski, G. (2013). Acceptance of Web 2.0 in Learning in higher education: an empirical study of a Scottish University, WBC 2013 July conference proceedings on e-learning, pp. 30-38.

[22] N. Fielding and R. Lee (1998). *Computer Analysis and Qualitative Research*. SAGE publications.

[23] B. Guerin (1993). *Social Facilitation*. Cambridge, England: Cambridge University press.

[24] S. Taylor, & Todd, P. A. (1995) Understanding information technology usage: A test of competing models. *Information systems research*, *6*(2), 144-176.

[25] J. Aiello and E. Douthitt (2001) Social Facilitation from Triplett to Electronic Performance Monitoring Group Dynamics. *Theory, Research and Practice*, Vol. 5, No 3, pp. 163-180.

[26] S. Kujawa and L. Huske (1995). *The Strategic Teaching and Reading Project guidebook* Oak Brook, IL: North Central Regional Educational Laboratory.

[27] T. Mitchell, S. Chen, and R. Macredie (2005). Hypermedia learning and prior knowledge: domain expertise vs. system expertise *Journal of Computer Assisted Learning* Vol. 21(1). pp. 53–64.

[28] Eccles, J. and Wigfield, A. (2002). *Annual Review in Psychology*, *Vol.* 53, pp.109–132.

[29] M. Fetscherin, and C. Lattemann (2008). User Acceptance of Virtual Worlds.*Journal of electronic commerce research*, Vol. 9, No. 3, pp. 231-241.

[30] G. Zikmund *Business Research Methods*, Mason, OH: Thomson/South Western, 2003.

[31] O'Cathain, A., Murphy E, Nicholl J (2010). Three Techniques for integrating data in mixed methods studies. BMJ 2010;341:c4587.

[32] V. Terzis, A. Economides (2011). The Acceptance and Use of Computer Based Assessment. *Computers & Education, Vol.* 56, pp. 1032 - 1044.

[33] T. Alrawashdeh, M. Muhairat, and S. Alqatawnah (2012). Factors Affecting Acceptance of Web-Based Training System: Using Extended UTAUT and Structural Equation Modeling. *International Journal of Computer Science, Engineering and Information Technology* (IJCSEIT), *Vol.* 2, No.2, pp. 45-54.

[34] D. Straub, M. Keil, and W. Brenner (1997) Testing the Technology Acceptance Model Across cultures: A Three Country Study. *Information and Management*, Vol. 33, No.1, pp. 1 - 11.

AUTHORS' PROFILE

Razep Echeng is a PhD research student at the University of the West of Scotland and also an associate lecturer teaching oracle database and design with SQL and software project management tutorials. She has a first degree in Computer Science from the University of Calabar, Nigeria and a master's degree in advanced computer systems development at the University of the West of Scotland. Her research is on e-learning with the use of Web 2.0 social media tools for learning. She has published six conference and journal articles in this area and written one book chapter. She is also interested in analysing and designing e-learning systems.

Dr Abel Usoro lectures and researches in information systems in the School of Computing, University of the West of Scotland. He has published widely in refereed journals, conferences and book chapters. He has served in the scientific committee of a few conferences two of which he has chaired. He is the lead editor of *Leveraging Developing Economies with the Use of Information Technology: Trends and Tools* (http://www.igi-global.com/book/leveraging-developing-economies-use-information/61661). His current research interest are knowledge management and e-learning. He has graduated three PhDs and also acts as director of studies to a few research students.

Experimental Analysis of the Fault Tolerance of the PIM-SM IP Multicast Routing Protocol under GNS3

Gábor Lencse
Department of Telecommunications
Széchenyi István University
Győr, Hungary

István Derka
Department of Telecommunications
Széchenyi István University
Győr, Hungary

***Abstract*—PIM-SM is the most commonly used IP multicast routing protocol in IPTV systems. Its fault tolerance is examined by experimenting on a mesh topology multicast test network built up by Cisco routers under GNS3. Different fault scenarios are played and different parameters of the PIM-SM and of the OSPF protocols are examined if they influence and how they influence the outage time of an IPTV service. The failure of the Rendezvous Point (RP) of the given IP multicast group as well as the complete failure of a router in the media forwarding path of the multicast stream are examined. A method is given how the service outage time caused by the complete failure of a router can be limited by an appropriate choice of the Dead Interval parameter of OSPF.**

Keywords—IP multicast protocols; PIM-SM; OSPF; fault tolerance; GNS3; modelling and simulation

I. Introduction

Protocol Independent Multicast – Sparse Mode [1] (PIM-SM) is the most commonly used multicast routing protocol in IPTV systems. The customers of these systems expect uninterrupted entertainment, which requires fault tolerance from the transport network. PIM-SM allows only one Rendezvous Point (RP, see later) per multicast groups [2] thus the RP can be a critical part of the multicast network. PIM version 2 introduced a fault tolerance mechanism with the help of the bootstrap mechanism [3]. This mechanism makes possible for a multicast based IPTV system to survive the failure of the RP; however the switching over to the new RP is not always invisible for the customers, but may cause service outage for a certain amount of time. In our previous research papers [4] and [5], we investigated the possible length of the service outage time and the parameters it may depend on. Different scenarios were investigated and parameters were tested whether they have an influence on the length of the service outage time, and if so, how they influence it. The measurements were performed on a mesh topology network built up by four times four XORP [6] routers. However, the behaviour of the fault tolerance of the PIM-SM implementation in the tested XORP system might depend on implementation specific details. In our current research, we performed our previous tests again using Cisco images under GNS3 [7] to validate our previous results.

The remainder of this paper is organised as follows. First, an introduction to the operation of PIM-SM and a very brief summary to IPTV systems are given. Second, the test environment is described. Third, the different kinds of experiments are presented and the results are interpreted. Fourth, the directions of our future research are outlined. Finally our conclusions are given.

II. Introduction To PIM-SM and IPTV Systems

The following descriptions are the taken from our previous paper [5] with some shortening. For a more detailed description with illustrative figures, see our original paper or [8].

A. The Operation of PIM-SM

Protocol Independent Multicast builds multicast trees on the basis of routing information obtained from a unicast routing protocol (e.g. RIP, OSPF) – this is why PIM is called "protocol independent". It has four variants, from which we deal with PIM-SM only. PIM-SM [9] does not suppose group members everywhere in the network but it sends multicast traffic into those directions where it has been requested using unidirectional *shared trees* rooted at the *Rendezvous Point* (RP). It may optionally use shortest path trees per source. PIM-SM does not have an own topology discovery method, but uses the Routing Information Base (RIB) of the unicast routing protocol applied in the *Autonomous System* (AS). With the help of this "outer" *Routing Information Base* (RIB), PIM-SM builds its own *Multicast Routing Information Base* (MRIB). Unlike unicast RIB (that specifies the next router towards the destination of the packets) MRIB specifies the reverse path from the subnet to the router.

As PIM-SM is an *Any-Source Multicast* (ASM) protocol, the receivers need to find the source(s). The so-called *Rendezvous Point* (RP) is used for this purpose. The RP can be set statically by the administrator of the AS, or it can be elected from among the RP candidate routers. There can be only one RP per multicast groups in the AS (or multicast domain) at a time.

The operation of PIM-SM has three phases. Now, we briefly describe what happens in these phases.

1) Phase One: RP-Tree

The *Rendezvous Point Tree* (RP-tree) is being built in the following way. The receivers send their IGMP (or MLD) Join messages with the required group address as destination IP address. The *Designated Router* (DR) of the receiver (that was elected from among the local routers before) receives the IGMP Join message and sends a *PIM Join* message to the RP

of the required multicast group. This PIM Join message travels through the routers in the network and the visited routers prepare the appropriate MRIB entries thus the *RP-tree* is being built. The PIM Join messages have the marking: (*, G), where the first element is the IP address of the streaming source and the second one is the IP address of the multicast group.

The star ("*") means that when a receiver joins a group, it will receive the traffic from all the sources that send steam to multicast group G. The PIM Join messages do not need to travel until the RP; it is enough to reach a point where the RP-tree has already been built. (The RP-tree is also called *shared tree* because the multicast traffic from all the sources uses the same tree.) The PIM Join messages are resent periodically while there is at least a single member in the group. When the last receiver of a leaf network leaves the group then DR sends a (*, G) *PIM Prune* message towards the RP so as to cut back the tree until the point where there are other active receivers connected.

When an S data source starts sending to a group, the first hop router (DR) of the source encapsulates the data packets of the source into unicast messages called *Register* messages and send them to the RP. The RP router knows from the Register messages that the source is ready to send the stream. The RP decapsulates the Register messages, and forwards the contained streaming data message to the appropriate multicast group (if it has at least a single member) using the RP-tree.

Note that the multicasting is fully functional at end of phase one; the following two phases serve efficiency purposes only.

2) Phase Two: Register-Stop

The RP sends an (S, G) Join message to the source. As this message travels to the source, the routers along its path register the (S, G) pair to their MRIB (if they do not have it yet). When this Join message arrives to the subnet of the source (S) or to a router that already has an (S, G) pair registered in its MRIB, then the streaming data start flowing from the S source to the RP by multicast routing. Now, a *source-specific multicast tree* between the S source and the RP was built. After that, the RP sends a *Register-Stop* message to indicate that the first hop router of the source does not need to send Register messages (encapsulating the multicast data packets into unicast messages).

3) Phase Three: Shortest-Path Tree

The path of the packets from the source to the receivers through the RP may be not optimal. To eliminate this, the DR of the receiver may initiate the building of a *source specific shortest-path tree* (SPT) towards the source (in this way possibly leaving out RP from the path). To do this, DR sends an (S, G) Join message to S. When this message arrives to the subnet of S or to a router that already has an (S, G) pair, then the streaming data start flowing from S to the receiver using this new SPT.

Now, the receiver receives all the streaming data packets twice. To eliminate this, the DR of the receiver sends an (S, G) Prune message towards the RP. (This is also known as an (S, G, rpt) Prune.) This message will prune the unnecessary tree parts and the streaming data will not arrive to the receiver through the RP-Tree any more.

4) The Built-in Fault Tolerance Mechanism of PIM-SM

It is an important element of the fault tolerance of PIM-SM that the RP does not need to be set up manually, it can be automatically elected from among those PIM-SM routers that were configured *Candidate RP* (C-RP).

The election uses the bootstrap mechanism described in [10]. The *BSR router* is elected dynamically from among the PIM-SM routers that were configured *Candidate BSR* (C-BSR). All the C-BSR routers flood the multicast domain with their *Bootstrap messages* (BSM). The one with the higher priority wins. During the BSR election all the routers – including C-RP routers – learn the IP address of the BSR. After that, all the C-RP routers send their *Candidate-RP-Advertisement* (C-RP-Adv) messages to the BSR periodically. (The C-RP-Adv messages are sent in every *C_RP_Adv_Period* seconds, the default value is 60 seconds.) The BSR collects these messages, builds an RP list and advertises it also periodically for all the routers. The list is encapsulated into a BSM and is sent in every *BS_Period* seconds. All the routers – including the BSR and the C-RPs – can decide the *winner RP* by the priority of the C-RPs. If the current RP fails to send its C-RP-Adv message to the BSR within *RP Holdtime* (a value included in the C-RP-Adv message) then BSR decides that it is dead and starts advertising the new RP list leaving out the dead one.

Notes:

*1) Ref. [10] says that the RP candidate routers should set RP Holdtime to a value that is not less than 2.5*max{BS_Period, C_RP_Adv_Period} so that the system is able to tolerate the loss of some Bootstrap messages and/or C-RP-Adv messages.*

2) The C-BSR routers also take care if the elected BSR fails, but that is not addressed in this paper.

3) The Choice of the Underlying Unicast Routing Protocol

As PIM-SM is *protocol independent*, there is a certain freedom in the choice of the underlying unicast routing protocol. The two most widely used protocols are the Routing Information Protocol [11] (RIPv2) and the Open Shortest Path First [12] (OSPFv2) for routing within a single autonomous system. Even though RIP is much simpler and more widely used in LANs than OSPF, it is not scalable and therefore it is not appropriate for the size of networks that are often used for providing IPTV services. This is why OSPF was chosen for our test network.

Note that OSFP also uses a fault tolerance mechanism but it is much simpler than that of PIM-SM. The OSPF routers take care for their neighbours only. All the OSPF routers send *Hello* messages in every *Hello Interval* seconds to their neighbours. If they do not see a Hello message from a neighbour within the so called *Dead Interval* time they consider the given neighbour dead and they calculate new routes leaving out the dead neighbour.

B. IPTV in a Nutshell

Nowadays, several data transmission technologies are available to transmit digital data (that may represent various media types, e.g. video, audio, text, etc. – the standard handles them in a uniform way) over different channels such as

DVB-S/S2 via satellite, DVB-T/T2 via terrestrial, DVB-C/C2 via cable TV links and so on. In the TCP/IP based networks, the commonly used solution for delivering the digital video, audio and auxiliary data is based on the DVB-IPTV [13].

A general property of the above mentioned technologies is that they use the same MPEG2 Transport Stream (MPEG2-TS) format to organize the digital data (video, audio, etc) and additional service information (SI/PSI tables) into a common frame. The MPEG2-TS is divided into 188 bytes long packets (4 bytes header and 184 bytes data). In the IPTV environment, usually seven TS packets are embedded into one IP/UDP or IP/UDP/RTP packet and they are sent through the network. Unlike other DVB technologies, IPTV does not use broadcasting to deliver these packets. Instead, it uses IP multicast for the live or online streaming (e.g. live TV) and unicast for the offline services, for example VoD or Timeshift.

When a subscriber would like to watch the selected IPTV program his/her receiver joins to the TV program's pre-programmed IP multicast group. After the join process (a few seconds) the receiver will get continuously the MPEG2-SPTS packets of the TV program through the IP multicast enabled network. If the subscriber switches over to another IPTV program then the receiver will leave the current one and join to the next IP multicast group.

III. Test Environment

In our previous paper [5], a mesh topology network was used in a virtualized environment with XORP [6] routers. Now the same topology is used and the GNS3 [7] environment was chosen for our experiments.

For our experiments, one Sun Server SunFire X4150 was used with the following configuration: two Quad Core Intel Xeon 2.83GHz CPUs, 8GB DDR2 800MHz RAM, 160GB HDD, Gigabit Ethernet NICs.

The topology of the test network was a mesh network built up by 4 times 4 virtual Cisco 7200 routers interconnected via Gigabit Ethernet point-to-point links as shown in Fig. 1. There were some further devices used: three virtual computers and three virtual hubs. The virtual computers were realized as Virtualbox virtual machines with 1GB RAM and 10GB HDD per virtual machine. The computers called Server and Client were used for media streaming, and the one named Manager was used for both managing the experiments and monitoring the traffic. For this reason, it had direct connections to all the three hubs (but the lines were omitted from the figure for aesthetic and clear-cut considerations). The usage of the hubs was inevitable for monitoring for the traffic and it also helped to keep away the management traffic from the mesh network.

The virtual computers had Ubuntu 12.04 LTS operating system. As for media streaming, the VLC software of VideoLAN was used for both server and client purposes.

A. IP Configuration

Private IP addresses were used from the 192.168.0.0/16 network. The IP addresses of the virtual routers were configured manually as shown in Fig. 1. The network segments between two routers displayed by horizontal and vertical lines got IP addresses from 192.168.{1-12}.0/24 and 192.168.{13-24}.0/24 networks respectively. The last octets of the IP addresses of the interfaces are written next to the interfaces. The IP addresses of the network segments connecting the server and the client virtual computers are displayed in a similar manner. Plus the Manager computer had additional IP addresses from the subnets it was connected to (through hubs).

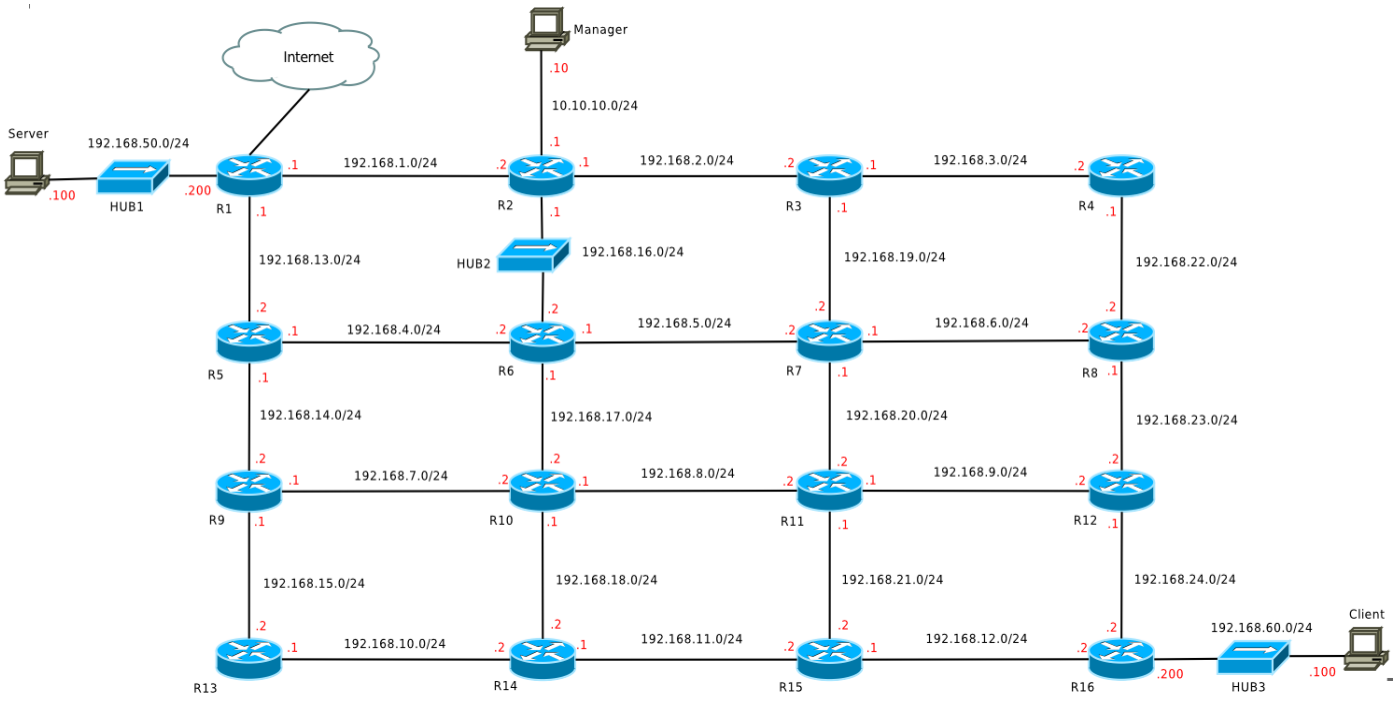


Fig. 1. Topology of the test network

B. OSPF Configuration

Because of the nature of the mesh, the OSPF protocol could be configured by the definition of peer-to-peer connections (it can be done if the neighbouring routers are interconnected by point-to-point links).

The typical general OSPF configuration commands were (they belong to R2):

```
router ospf 20
 router-id 192.168.1.2
 log-adjacency-changes
 network 192.168.1.0 0.0.0.255 area 1
 network 192.168.2.0 0.0.0.255 area 1
 network 192.168.16.0 0.0.0.255 area 1
```

And the typical interface configuration fragment for an interface looks like follows (it belongs to an interface of R2):

```
interface GigabitEthernet1/0
 ip address 192.168.2.1 255.255.255.0
 ip ospf network point-to-point
 ip ospf cost 1
 ip ospf hello-interval 10
 ip ospf dead-interval 40
 negotiation auto
```

Configuring OSPF in this way made the network fully connected: unicast IP packets can be sent from anywhere to anywhere. Note that PIM-SM uses the unicast routing table (RIB) when building its own multicast routing table (MRIB).

C. PIM-SM Configuration

For PIM-SM, those and only those interfaces should be configured where PIM-SM has to handle multicast traffic. A typical configuration for an interface looks like follows:

```
interface GigabitEthernet1/0
 ip pim sparse-mode
```

In order to be able to experiment with the fault tolerance of PIM-SM, the dynamic election of the RP was used. This required us to configure some routers as C-RP and at least one router as C-BSR. Routers R2, R4 and R14 were configured as both C-RP and C-BSR but with different priorities[1].

The R2 router was the highest priority C-RP, the R4 was the second highest priority one; R14 was the highest priority C-BSR. A typical configuration for a router that was set as both C-RP and C-BSR looks like follows:

```
ip pim bsr-candidate GigabitEthernet1/0 1 2
ip pim rp-candidate GigabitEthernet1/0 priority 2
```

Considering the fact that in phase three there is no need for the RP, but a source-specific shortest path tree (SPT) is used for the transmission of the stream (that may not contain RP, or even if it contains RP then RP acts like a simple multicast router only), PIM-SM was configured so that it would never enter phase three:

```
ip pim spt-threshold infinity
```

[1]As it is defined in [9], the lower numeric value means higher priority.

D. Time Synchronization

The important events of the measurements were logged into text files. In order be able to compare the timestamps of the events occurred on different virtual routers or computers, the system times of the other virtual routers and computers were synchronized to R1 using the standard NTP protocol. The R1 router had an Internet connection and it was synchronized to a stratum 1 time server in the Internet. The configuration of R1 had the following commands:

```
ntp master 3
ntp server 130.149.17.21
```

The other routers had the following command:

```
ntp server 192.168.1.1
```

E. Streaming

A single program transport stream (SPTS) – that was demodulated and demultiplexed from a Hungarian DVB-T multiplex – was pre-recorded and used for all the measurements. The VLC server sent the stream to the 239.1.1.1 multicast IP group address using UDP. The VLC client received the stream and the standard `tshark` program was used by the Manager virtual machine to capture (and record for offline analysis) the stream on the receiver side through HUB3.

IV. Experiments and Results

A. Testing the Failure of the RP

Hypothesis 1: Killing the RP on R2 router will not stop the stream (unlike it was stopped when XORP was used [5]) because Cisco PIM-SM fully complies with the PIM-SM standard [9] and thus the RP functionality is no more needed at the end of phase two.

The measurements were controlled by a script executed on the Manager virtual computer. The RP functionality was switched off by the remote execution of the following commands on the R2 router:

```
no ip pim bsr-candidate GigabitEthernet1/0 1 2
no ip pim rp-candidate GigabitEthernet1/0 priority 2
```

The experiments were executed multiple times and we did not experience service outage. We have also checked that R2 stopped functioning as an RP and R4 was elected as the new RP.

B. Testing the Failure of the Complete PIM-SM router

Hypothesis 2: Switching off the operation of the complete R2 router will stop the stream for a while, but the stream will be restored when the underlying unicast routing (OSPF) finds a new route (that does not contain the R2 router) from the DR of the server to the DR of the client. The length of the service outage time is expected to show no correlation with the time elapsed from the last C-RP-Adv message when the RP is killed.

The measurements were controlled by a script executed on the Manager virtual computer. This script did the following: after the establishment of the OSPF routing table on R2 and starting the streaming, it made sure that R2 is the actual RP.

Then it waited until PIM-SM sent a C-RP-Adv message. From that time it waited until a *predefined delay* (it was a parameter, see later). After that it started the scripts and it sent a marker (ICMP echo request) to the client and then it remotely switched off both unicast and multicast routing functionality of R2 by shutting down of its network interfaces.

The following commands were used:

```
access-list 100 deny   ospf any any
interface GigabitEthernet1/0
 ip access-group 100 in
 ip access-group 100 out
 shutdown
interface GigabitEthernet2/0
 ip access-group 100 in
 ip access-group 100 out
 shutdown
interface GigabitEthernet3/0
 ip access-group 100 in
 ip access-group 100 out
 shutdown
```

Note that the filtering out of all the OSPF messages was necessary because otherwise the OSPF running on R2 would have notified its neighbours (using OSPF LSA updates). However, our aim was to simulate not the regular shut down but the failure of R2.

The *predefined delay* was increased from 5 seconds to 55 seconds in 5 seconds steps. (As C-RP-Adv is sent in every 60 seconds by the defaults settings of the Cisco IOS, thus there would be no point in increasing the delay above 55 seconds.) The whole measurement was executed 11 times. The complete script can be found in the Appendix.

The results can be found in Fig 2. They justify hypothesis 2: the stream stopped for a while and it continued after a certain amount of time; the service outage time shows no correlation with the time elapsed from the last C-RP-Adv message. Both prove that *no new RP is necessary for the restoration of the stream.*

Hypothesis 3: The length of service outage time caused by the switching off the operation of the complete R2 router depends on the *time elapsed from the last Hello message* of the OSPF protocol.

The default values of the OSPF *Hello Interval* and *Dead Interval* are 10 seconds and 40 seconds respectively. For testing purposes, the first one was raised to 35 seconds and similar series of the measurements were performed in the way that the *delay* from the last OSPF *Hello message* before the stopping of R2 was increased from 5 seconds to 30 seconds in 5 seconds steps.

The results can be found in Fig. 3. They justify hypothesis 3: the average service outage times are approximately 5 seconds higher than the time that was left from the *Dead Interval* of OSPF at the time of stopping R2. (The stream was restored because OSPF calculated a new route that did not contain the R2 router.)

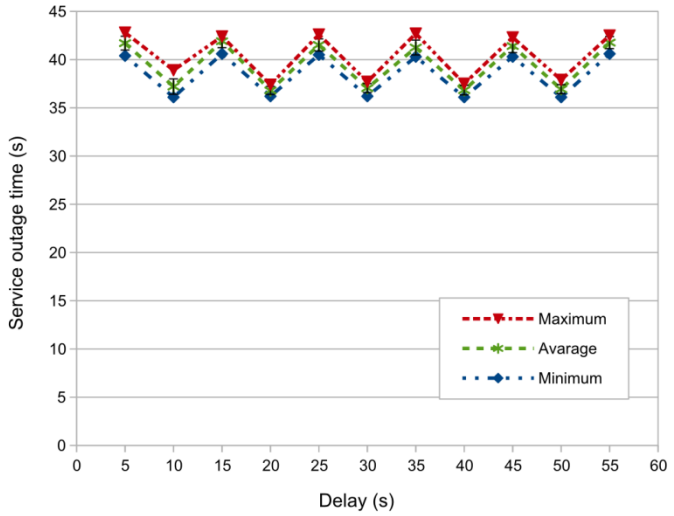


Fig. 2. Service outage times in the function of the delay from the last **C-RP-Adv Message** to the stopping of the R2 router

C. *Limiting the service outage time by parameter tuning*

As we have shown in section B, if the service outage was caused by the complete failure of a multicast router[2] which is an element of the path from the DR of the server to the DR of the client then the service outage time was determined by the parameters of the underlying unicast routing protocol. In our experiments, the service outage time was upper bounded by the *Dead Interval* of OSPF. The actual value of the service outage time depended on the elapsed time from the last OSPF *Hello* message at the time of the failure of R2.

Hypothesis 4: The service outage time caused by the complete failure of a multicast router can be limited by an appropriate setting of the OSPF *Dead Interval* parameter. The measurements were taken in the usual way but using 20 seconds and 15 seconds as OSPF *Dead Interval* and *Hello Interval*, respectively. The used values of the delay from the last OSPF *Hello* message to the failure of R2 were 5 and 10 seconds. The results can be found in Table 1. They justify hypothesis 4.

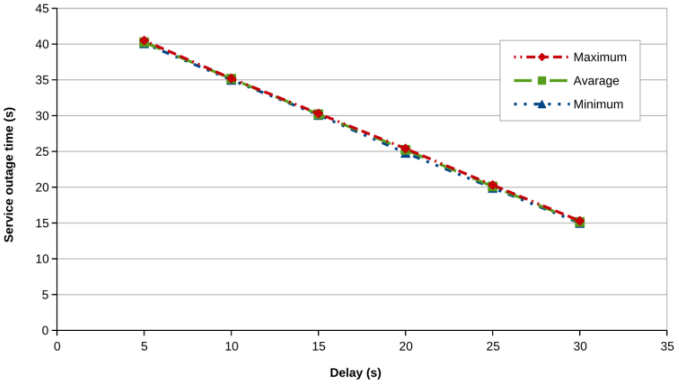


Fig. 3. Service outage times in the function of the delay from the last **OSPF Hello Message** to the stopping of the R2 router

[2]It can be the RP, but it is not necessarily the RP.

TABLE I. SERVICE OUTAGE TIMES IN THE FUNCTION OF THE DELAY FROM THE LAST **OSPF HELLO MESSAGE** TO THE STOPPING OF XORP ON XORP2 ROUTER **USING 20 SECONDS OSPF DEAD INTERVAL**

Delay (s)	Service outage time (s)			
	min	max	average	std. dev.
5	19.7	20.8	20.07	0.26
10	14.9	15.1	15.05	0.07

The significance of the findings of hypotheses 4 is that the time of the service outage caused by the complete failure of a multicast router can be limited by the appropriate choice of the *Dead Interval* parameter of OSPF. Note that the service outage time cannot be arbitrarily decreased in this way for at least two reasons:

1) The choice of the Dead Interval parameter of OSPF has a consequence on the frequency of the OSPF Hello messages. This frequency should not be too high as these messages consume both network and router capacity.

2) The exchange of the topology information and the recreation of the routing tables in OSPF require a certain amount of time. Though this time was negligible in our experiments due to the small size of our test network, the situation can be different in the case of a real life multicast network for IPTV.

V. FUTURE RESEARCH AND RELATED WORK

It is also our aim with the experiments described above to collect both measurement data and experience with PIM-SM implementations to be able to prepare a new or to improve an existing PIM-SM simulation model. Simulation is a powerful tool for the performance and fault tolerance analysis of complex ICT (Information and Communication Technology) systems [14]; and our measurement results may help in building a good simulation model.

Our preferred simulation environment is the OMNeT++ [15] discrete event simulation framework for multiple reasons:

- it is modular, extensible and convenient
- it is open source and free for academic purposes
- we had good experiences with it in our previous research projects [16] and [17].

The first PIM-SM model for OMNeT++ was published in [18]. The implementation was the MSc final project work of Juan Ángel Cachinero in 2009. His teacher, Raquel Perez Leal was kind enough to collect and send us the source code in 2013, however it turned out that the code was developed under a very old version (3.3) of OMNeT++ thus we abandoned it.

PIM-DM has been recently implemented for OMNeT++ at the Faculty of Information Technologies, Brno University of Technology as a part of the ANSA project and the implementation of PIM-SM is on the way according to [19]. We plan to join to this effort and complete the PIM-SM model with the implementation of its fault tolerance mechanism.

Multipath stream transmission [20] is also a novel and challenging idea. The MPT [21] library was developed and it is being actively researched at the Faculty of Informatics, University of Debrecen, see also [22] and [23]. We have already contacted Béla Almási, the first author of these papers and we are looking for a common ground for our research activities.

VI. CONCLUSIONS

We have given an introduction to the PIM-SM multicast routing protocol and the description of our test environment.

We have shown that RP is no more used at the end of phase two of PIM-SM thus its failure does not interrupt the ongoing media streaming. This is an important difference between the behaviour of the XORP and of the Cisco PIM-SM implementations in phase 2.

We have shown that the complete failure of a PIM-SM router that resides in the media forwarding path results in service outage but the restoration of the stream does not requires an active RP rather it is done by the underlying unicast routing protocol (actually OSPF). The length of the service outage depends on the parameters of the underlying OSPF and can be bounded by the appropriate choice of the *Dead Interval* parameter of OSPF.

Our results may be used by IPTV service providers for improving the availability and fault tolerance of their IPTV systems.

We plan to use our experience with the different PIM-SM implementations in the development and/or improvement of simulation models for PIM-SM.

ACKNOWLEDGMENT

The text and measurement methods of our previous paper [5] (about our experiments under the XORP platform) were reused in this paper but now our experiments were performed in the GNS3 environment.

REFERENCES

[1] S. Deering, D. Estrin, D. Farinacci, V. Jacobson, C. Liu, L. Wei, “The PIM architecture for wide-area multicast routing”, IEEE/ACM Transactions on Networking, Vol. 4, No. 2, pp. 153-162. April 1996. DOI:10.1.1.39.7251

[2] M. Sola, M. Ohta and T. Maeno: “Scalability of internet multicast protocols”, in: Proc. of INET'98, Geneva, Switzerland, July 1998.

[3] Silvano Da Ros, Content Networking Fundamentals, Cisco Press, 2006, ISBN: 1-58705-240-7

[4] G. Lencse and I. Derka, “Towards the modelling of the fault tolerance mechanism of the PIM-SM multicast routing protocol in an IPTV Environment” Proceedings of the 2012 European Simulation and Modelling Conference (ESM'2012): Modelling and Simulation 2012, (Essen, Germany, 2012. Oct. 22-24.) EUROSIS-ETI, 152-156.

[5] G. Lencse and I. Derka, “Investigation of the fault tolerance of the PIM-SM IP multicast routing protocol for IPTV purposes”, Infocommunications Journal, Vol. V, No. 1. (March, 2013) pp. 21-28.

[6] XORP Inc. and individual contributors, XORP user manual, Version 1.8-CT, 2010.

[7] Graphical Network Simulator – GNS3 http://www.gns3.net/

[8] Beau Williamson, Developing IP multicast networks, Volume 1, Cisco Press, 2000, Indianapolis, IN, USA. ISBN: 1-57870-077-9

[9] B. Fenner, M. Handley, H. Holbrook, and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", IETF, August 2006, RFC 4601

[10] N. Bhaskar, A. Gall, J. Lingard, and S. Venaas, "Bootstrap Router (BSR) Mechanism for Protocol Independent Multicast (PIM)", IETF, January 2008, RFC 5059

[11] G. Malkin, "RIP Version 2", IETF, November 1998, RFC 2453

[12] J. Moy, "OSPF Version 2", IETF, April 1998, RFC 2328

[13] Digital Video Broadcasting (DVB); Transport of MPEG-2 TS Based DVB Services over IP Based Networks, ETSI TS 102 034 V1.4.1 (2009-08)

[14] L. Muka and G. Muka. "Creating and using key network-performance indicators to support the design and change of enterprise infocommunication infrastucture", Proceedings of 2012 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS 2012), (Genoa, Italy, July 8-11, 2012) Volume 44, Books 12, pp. 737-742.

[15] A. Varga and R. Hornig, "An overview of the OMNeT++ simulation environment", Simutools '08: Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops (Marseille, France, March 7, 2008), ICST, http://dl.acm.org/citation.cfm?id=1416290

[16] G. Lencse I. Derka and L. Muka, "Towards the efficient simulation of telecommunication systems in heterogeneous execution environments", Proc. of the 36th International Conference on Telecommunications and Signal Processing (TSP 2013), (Rome, Italy, July 2-4, 2013), pp 304-310. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6613941

[17] G. Lencse and I. Derka, "Testing the speed-up of parallel discrete event simulation in heterogeneous execution environments", *Proceedings of the ISC'2013, 11th Annual Industrial Simulation Conference*, (Ghent, Belgium, May 22-24, 2013) EUROSIS-ETI, 101-107.

[18] R. P. Leal, J. Á. Cachinero, E. P. Martín, "New approach to inter-domain multicast protocol", ETRI Journal, Vol 33. No. 3. June 2011, pp. 355-365. http://dx.doi.org/10.4218/etrij.11.0110.0405

[19] V. Veselý, O. Ryšavý, M Švéda, "IPv6 unicast and IPv4 multicast routing in OMNeT++", Simutools '13: Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques, (Cannes, France, March 5-7, 2013) pp. 346-349. http://dl.acm.org/citation.cfm?id=2512785

[20] B. Almási, Sz. Szilágyi, "Multipath ftp and stream transmission analysis using the MPT software environment", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 11, (November 2013) pp. 4267-4272.

[21] B. Almási, A. Harman, "An overview of the multipath communication technologies", Proceedings of the Conference on Advances in Wireless Sensor Networks 2013 (AWSN 2013), Debrecen University Press, Debrecen, Hungary, ISBN: 978-963-318-356-4, 2013, pp. 7-11.

[22] B. Almási, Sz. Szilágyi, "Throughput Performance Analysis of the Multipath Communication Library MPT", Proceedings of the 36th International Conference on Telecommunications and Signal Processing (TSP 2013), (Rome, Italy, July 2-4, 2013), pp 86-90.

[23] B. Almási, "Multipath Communication – a new basis for the Future Internet Cognitive Infocommunication", Proceedings of the CogInfoCom 2013 Conference, (Budapest, Hungary, December 2-5, 2013), pp. 201-204.

APPENDIX

This script was used for testing our second hypothesis in section 4.B. The scripts used for the other tests are very similar to this one.

```
#!/bin/bash
rtlog="/mnt/data/measurements/runtime.log"
for j in {1..11}
do
  echo -e "$j-th run\t$(date +%T.%N | cut -b1-12)" >> $rtlog
  echo -e "================================================" >> $rtlog
  mkdir /mnt/data/capture/t$j >> $rtlog 2>&1
  for i in {5..55..5}
  do
    echo -e "\nDelay $i seconds\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    echo -e "-----------------" >> $rtlog
    msglog="/mnt/data/measurements/messages_$j-$i.log"
    echo -e "R2 start\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    sshpass -p cisco ssh -t admin@r2 "tclsh disk0:switchon.cfg" >> $msglog 2>&1
    ospf=$(sshpass -p cisco ssh -t admin@r2 "sh ip route ospf" 2>/dev/null | wc -l | tr -cd "[:print:]")
    echo -e "Number of OSPF routes: $ospf\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    while [ $ospf -eq 0 ]
    do
      ospf=$(sshpass -p cisco ssh -t admin@r2 "sh ip route ospf" 2>/dev/null | wc -l | tr -cd "[:print:]")
    done
    echo -e " Number of OSPF routes: $ospf\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    echo -e "R2 started\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    echo -e "Sending the stream\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    screen -S streaming -d -m ssh -t steve@server "/home/steve/streaming.sh"  >> $msglog 2>&1
    echo -e "Receiving the stream \t$(date +%T.%N | cut -b1-12)" >> $rtlog
    screen -S streamplay -d -m ssh -t steve@client "/home/steve/streamplay.sh" >> $msglog 2>&1
    rp=$(sshpass -p cisco ssh -t admin@r2 "show ip pim rp" 2>/dev/null | awk '{ print $4 }' | sed 's/,//' | \
      tr -cd "[:print:]")
    echo -e "Current RP: $rp\t" >> $rtlog
    echo -e "Waiting until the restoration of RP\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    while [ "$rp" != "192.168.2.1" ]
    do
      rp=$(sshpass -p cisco ssh -t admin@r2 "show ip pim rp" 2>/dev/null | awk '{ print $4 }' | sed 's/,//' | \
        tr -cd "[:print:]")
    done
    echo -e "RP was restored\t$(date +%T.%N | cut -b1-12)\tRP: $rp" >> $rtlog
    echo -e "Waiting for Cand-RP-Advertisement\t$(date +%T.%N | cut -b1-12)" >> $rtlog
```

```
    tshark -i eth2 'ip host 192.168.16.1 and pim[0]=0x28' -c 1 -t a >> $msglog 2>&1
    echo -e "Cand-RP-Advertisement was found\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    echo -e "Start Capture\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    screen -S capture -d -m tshark -i eth3 -a duration:180 -s 128 -w /mnt/data/capture/t$j/test-$i.pcap >> \
      $msglog 2>&1
    echo -e "Delaying $i seconds\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    sleep $i >> $msglog 2>&1
    echo -e "Elapsed\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    echo -e "Set Marker\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    ping -c 1 client >> $msglog 2>&1
    echo -e "Stop R2 \t$(date +%T.%N | cut -b1-12)" >> $rtlog
    sshpass -p cisco ssh -t admin@r2 "tclsh disk0:switchoff.cfg" >> $msglog 2>&1
    echo -e "Wait until the end of the capture interval\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    capture=$(ps aux | grep tshark | wc -l)
    while [ $capture -gt 1 ]
    do
      capture=$(ps aux | grep tshark | wc -l)
    done
    echo -e "Stop receiving the Stream\t$(date +%T.%N | cut -b1-12)" >> $rtlog
    screen -S streamplay -X quit >> $msglog 2>&1
    screen -S streaming -X quit >> $msglog 2>&1
    echo -e "End of Capture\t$(date +%T.%N | cut -b1-12)\n" >> $rtlog
  done
  echo -e "================================================" >> $rtlog
done
```

Novel LVCSR Decoder Based on Perfect Hash Automata and Tuple Structures – SPREAD –

Matej Rojc
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia

Kačič Zdravko
Faculty of Electrical Engineering and Computer Science
University of Maribor
Maribor, Slovenia

***Abstract*— The paper presents the novel design of a one-pass large vocabulary continuous-speech recognition decoder engine, named SPREAD. The decoder is based on a time-synchronous beam-search approach, including statically expanded cross-word triphone contexts. An approach using efficient tuple structures is proposed for the construction of the complete search-network. The foremost benefits are the important space savings and higher processing speed, and the compact and reduced size of the tuple structure, especially when exploiting the structure of the key. In this way, the time needed to load the ASR search-network into the memory is also significantly reduced. Further, the paper proposes and presents the complete methodology for compiling general ASR knowledge sources into a tuple structures. Additionally, the beam search is enhanced with the novel implementation of a bigram language model Look-Ahead technique, by using tuple structures and a caching scheme. The SPREAD LVCSR decoder is based on a token-passing algorithm, capable of restricting its search-space by several types of token pruning. By using the presented language model Look-Ahead technique, it is possible to increase the number of tokens that can be pruned without decoding precision loss.**

Keywords—LVCSR decoder; tuple structure; finite automata; perfect hashing; Look-Ahead; language models

I. INTRODUCTION

A LVCSR decoder represents a major component in the development of any continuous speech-recognition system. Since tasks' and systems' complexities are constantly increasing, the decoder becomes an increasingly significant component within the overall development of compact and efficient speech-recognition systems. Therefore, more efficient designs can improve the trade-off between the needed decoding time and the recognition error rate. Furthermore, large knowledge sources are used in the LVCSR decoder, enabling estimation of the most likely word sequence from specific acoustic evidence. In general, these knowledge sources are: acoustic models (HMM - Hidden Markov Models), pronunciation lexicon, and N-gram language models. More and more new application areas require increase in the complexity of acoustic, lexicon and language models used in LVCSR decoders. Consequently, the requirements for time and space efficiency of LVCSR decoders are becoming greater and greater, despite the continuous growth of hardware performance, and GPU-like parallel processing. Therefore, efficient management of all these knowledge-sources, and efficient decoding of the acoustic input, still remains important issues and challenging tasks. Furthermore, in LVCSR decoders a lot of optimisation techniques, specific architectures, and heuristics have to be used and developed in order to achieve lower computational complexity and lower memory requirements. Progress regarding LVCSR decoding algorithms, together with the availability of ever increasing computing power and memory capacity, has also resulted in more accurate and close to real-time LVCSR decoders for tasks such as, e.g. broadcast news transcription, conversational telephone speech recognition systems etc. Technology based on weighted finite-state machines (WFSM) has already shown that it is possible to efficiently encode all those knowledge sources present within a speech-recognition system, such as e.g. language models, pronunciation dictionaries, context decision-trees, etc. By using them, a LVCSR network is usually obtained by a composition of several WFSTs. After using minimisation algoritms, an LVCSR network can be directly used in a Viterbi-based LVCSR decoder. These decoders have already been shown to yield good performance when compared to the classic approaches. This results in the implementations of several Viterbi-based decoders using FSA technology [7, 16, 17, 18, 19, 24]. Nevertheless, the complexity of acoustic and language models used in speech recognition tasks still imposes growing requirements for the efficiency and accuracy of LVCSR decoders, and fosters the development of new approaches and techniques such as, e.g. cross-word acoustic models and long-span language models, already resulted in the development of several solutions for the speech-decoding problem [1, 2, 5, 6, 8, 10, 21, 22].

II. RELATED WORK

Nowadays, a lot of speech-decoding software packages exist that employ a number of different decoding techniques, based on time synchronous Viterbi search and many are also available for research purposes. CMU/Sphinx, released by Carnegie Mellon University (CMU) [26], contains less features and flexibility, but in contrast to HTK [27], focuses more on speed and was one of the first ASR systems to offer support for speaker-independent Large Vocabulary Continuous Speech Recognition (LVCSR) system. Latest versions, although less efficient than previously, are more flexible and enables faster and easier development and maintainance of different applications [17]. Further, the Julius LVCSR decoder is a high-performance, two-pass decoder, focusing on performance, modularity, and availability [9]. The HTK framework is very flexible and comprises a lot of state-of-the-art ASR features,

e.g. vocal tract length normalization (VTLN), heteroscedastic linear discriminant analysis (HLDA) and discriminative training with maximum mutual information (MMI), and minimum phone error (MPE) criteria [15,27]. Some of the decoder implementations have shifted from dynamic search to static graphs in the form of probabilistic weighted finite-state transducers (WFSTs) [4, 7, 18]. Architectures based on the theory of weighted finite-state transducers represent flexible and efficient decoder architectures. The advantages of this implementation can be seen in the simplicity of LVCSR decoders and the seamless composition of lexicon, acoustic, and language models. One of the most efficient solutions for search-network optimization is the WFST framework from [12]. In these architectures, all the knowledge sources are combined together statically. Furthermore, the search network can be optimised for maximal efficiency. In such LVCSR systems the decoding network is usually compiled independently of the LVCSR decoder itself, in this way representing also more flexible solutions for e.g. the incorporation of several application-specific knowledge sources. Nevertheless, Mohri's approach can restrict the complexity of the knowledge sources, and prevent some on-the-fly adaptation [22]. A drawback can also be seen in the memory requirements for the compilation of very large static decoding networks for LVCSR systems, although today this issue has become less crucial because of the availability of 64-bit systems, and a lot of available RAM memory. Another approach from [25] expands the search network dynamically. This approach, on the other hand, can be computationally too expensive for efficient decoding regarding larger LVCSR tasks. Recently, the Juicer WFST decoder has become a popular WFST-based alternative to the tree-based dynamic decoders, as provided with the HTK and Sphinx toolkits. The T3 WFST decoder is a system that performs favorably against several established decoders in the field, including the Juicer, Sphinx, and HDecode [27] in terms of RTF versus Word Accuracy [17]. In the case of the T3 WFST decoder, in addition to the existing HTK conversion tool, a tool has also been developed for converting arbitrary Sphinx format acoustic models into a format suitable for use with the T3 WFST decoder [4]. Juicer provides similar functionality to the T3 WFST decoder in terms of the model inputs it accepts. It is capable of performing decoding on both static cascades, as well as on-the-fly composition, and it has been developed to read in HTK-based acoustic models in native format [13].

This paper proposes a novel LVCSR decoder named SPREAD that is implemented by using efficient perfect-hash automata and tuple structures. The complete search network is compiled independently of the LVCSR decoder (off-line), including the needed pronunciation lexicon, language models and Look-Ahead structures, and can be fast loaded by the runtime system. The language model information is, in this way, dynamically obtained during the search within the LVCSR decoding engine. Furthermore, the proposed off-line compilation methodology of large static networks is simple and fast, even on 32-bit machines with less available RAM memory. In this LVCSR decoder, the novel implementation of the language model Look-Ahead technique (used to enhance the beam search results), based on tuple structures, is further integrated. In this way, the proposed LVCSR decoder incorporates several novel design strategies, which have not been used earlier in conventional decoders of HMM-based large vocabulary speech recognition systems. The paper starts with a motivation for using tuple structures in speech-technology-related applications, and for developing a LVCSR decoder, based on tuple structures. Firstly, the formalism behind tuple structures is presented, regarding their form and representation. Then the presentation of the LVCSR decoder technology used within the LVCSR SPREAD decoder follows. Next, the proprietary FSM tools are discussed and their application for the construction of tuple structures when developing LVCSR decoders. The main part of the paper represents the proposed compilation methodology of all the tuple structures used within a general LVCSR decoder. Additionally, the implementation of a large-scale search-network using tuple structures is described in detail. The proposed work is based on real implementation of the LVCSR decoder based on tuple structures, as used for 64k broadcast news transcription speech-recognition task for the Slovenian language. Statistics and achieved compactness of the proposed implementation are, therefore, presented in Section 8. In this way, the paper familiarizes the readers with the design solutions encountered, when building a tuple-based LVCSR decoders. The conclusion is drawn at the end.

III. MOTIVATION

The general practical issue in LVCSR speech-recognition applications concerns the size of the knowledge sources, and the size of the complete search-network. This issue can be even more crucial when the knowledge sources are consulted frequently and must, therefore, be loaded into the memory. Perfect hashing techniques based on finite-state automata can be very efficient when solving these problems [3, 23]. Namely, as will be shown, they enable compact representations without sacrificing the lookup time. In the case of LVCSR speech-recognition applications, large dictionaries are not the only space-consuming resources. Namely, several types of language models containing statistical information about the co-occurrence of words, require even more memory space, and also at the same time as fast lookup operations as possible – LVCSR systems need to be capable of working with e.g. bigram, trigram, fourgram models etc. Therefore, for speech-recognition applications, the achievable size and compactness of language models and other knowledge sources within the runtime system represent an important practical implementation issue, and also motivation for the work presented in this paper. The Slovenian language is a highly inflectional language. Therefore, the number of distinct word forms in everyday use is very large, resulting in large knowledge sources for general LVCSR speech recognition tasks. When considering this, efficient management of the data structures' size when representing knowledge sources, and the lookup efficiency, are general requirements. In this respect a very compact representation of knowledge sources and the search-network is needed, and a highly-optimized LVCSR ASR decoder must be implemented. In order to better cope with this problem, it was decided to work on a new design approach for the development of an LVCSR decoder that is based completely on perfect hash automata and the so-called tuple structures, by following the established theory on tuples

in [3]. In general, the needed knowledge sources for LVCSR ASR decoder can be represented in the form of a simple data structure that defines a mapping from some strings to some value. These data structures can be easily generalized, in which the keys are *n*-tuples of strings, for a fixed *n*. Such data structure is called tuple structure, and apart from the N-gram language models, it can also be used for the creation of large ASR search-networks including LMLA (language model look-ahead) info, as presented in the following sections. Although operations on tuple structures, like insertion and deletion, are not well supported, they can be ignored in the case of LVCSR decoders, since tuple structures can be constructed once from a given data-set (off-line), and then only loaded and used within the runtime system. At the end, also compact representation of the tuple structure is very important. Namely, by compact representation of knowledge sources using tuple structures, the time needed to load the structures into the memory is significantly reduced. The techniques used in the presented work, can be seen as applications and extensions of perfect hashing based on finite-state automata. Therefore, the proposed implementation yields to flexible and compact representation of large scale knowledge sources and also LVCSR search-networks in practice. The following section presents the basic formalism behind tuple structures.

IV. Tuple Structures

A tuple structure $T^{i,j}$ is a finite function $(W_1 \times ... \times W_i) \rightarrow Z^j$. In this finite function, $W_1 \times ... \times W_i$ are simple sets of strings, and Z are the integers [3]. This finite function can map to a tuple of integers, or to a tuple of real numbers. The so-called *word columns* contain words (e.g. lexicon words, pronunciations, diphones, triphones, language models' pairs, Look-Ahead pairs etc.). And the so-called *number columns* contain one or several integer numbers, or real numbers (e.g. N-gram probabilities, N-gram backoff-weights). Figure 1 presents part of the table forms used for the construction of tuple structures for a bigram language model, and for one of the layers within the LVCSR search network. In the first case, the word columns contain word sequences, and the number columns the bigram probabilities and backoff-weights. In the second case, the word columns contain triphones and diphones, and the number columns contain the next layer ID and the next node type (context or model node). Perfect hash finite automata are needed for the tuple structures. The perfect hash finite automaton for a finite set of words *W* is such minimal deterministic acyclic finite automaton *N* that accepts each word in *W*. And each transition within the automaton has an assigned integer number *j*. Let some word *w* represents the *i*-th word of *W*. Then the sum of the integers along an accepting path in *N* is *i*. If *N(w)* refers to the *hash key* assigned to *w* by *N*, then the time spent for its computation is $O(|w|)$ [3]. The perfect hash automata are needed in order to represent all the words in the word columns with hash keys. Furthermore, they can be used within the LVCSR decoder, when translation from the hash keys back into words is needed (e.g. ASR output results etc.). When there is enough overlap between the words from several word columns within the table forms, the same perfect hash automaton for all those columns can be used. Although, the tuple structures are able to take advantage of such shared dictionaries, it is not required that the dictionaries for different word columns are the same.

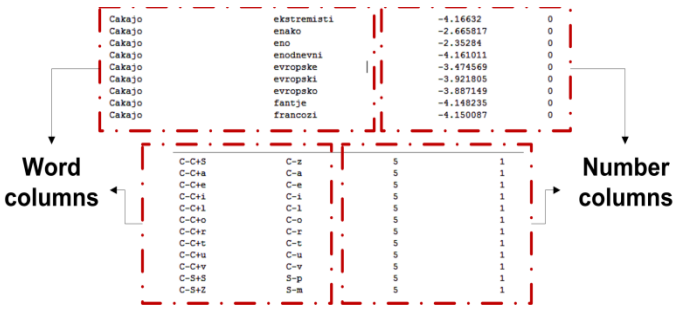


Fig. 1. Table forms consisting of word and number columns.

In general, several hash automata are used (one for each word column). Nevertheless, in the first case more space savings can be achieved. Figure 2 then shows the representation of word columns by the corresponding hash keys for the bigram language model and for one of the layers used in the ASR search-network.

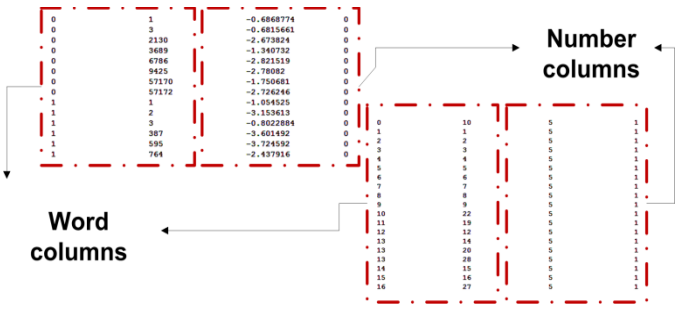


Fig. 2. Representation of word columns in table forms by using the hash keys.

A. A Table Form for Compact Representation of Tuple structures

The tuple structure $T^{i,j} : (W_1 \times ... \times W_i) \rightarrow Z^j$ is in general represented by maximal *i* perfect hash automata (when each word column has its own perfect hash automaton). Then, for each tuple structure a table form consisting of *i*+*j* rows is constructed (Figure 2). The table forms are constructed for each sequence $w_1 ... w_i$ in the domain of *T*. For *T* we have the following transformation $T(w_1 ... w_i) = (z_1 ... z_j)$. The sequences of words $w_1 ... w_n$ are converted into their hash-keys $N(w_1) .. N(w_n)$ by using perfect hash automata. In this way, each word sequence is represented by a row in the table, consisting of $N(w_1), ..., N(w_i), z_1, ..., z_j$ [3]. As can be seen in Figure 2, all the cells in the table contain numbers at the end. For compact representation it is, therefore, important that each hash-key is represented with as few bytes as are required by the largest number within individual column. An additional benefit is the machine-independency of such representation. The tables also have to be sorted in order to guarantee sorted and unique entries. At the end, the tuple structure is represented by a table of packed numbers and *i* perfect hash automata that can be used for translating words into corresponding hash-keys and vice

versa. In order to access value(s) for a given sequence, a query string is needed, consisting of hash-keys. A binary search is used to find the corresponding entry within the table. The data for a given sequence can be obtained after an unpacking of the values found in the table is performed. The time needed for calculating the hash-keys is proportional to the combined length of words within the table's entry. The binary search takes $O\left(\log\left|T^{i,j}\right|\right)$ time and is proportional to the logarithm of the number of tuples. Tuple structures $T^{i,j}$ can also be constructed when i=1 (there is only one word column, e.g. unigrams for language model). In this case, the words in the word column are unique, therefore, their hash keys are also unique numbers from $0...|W_1|-1$. Consequently, there is no need to store the hash keys of the words within the table. Instead, the hash-keys just serve as an index within the table. Also the lookup function is different. After the hash-key is obtained, it is used as the address of the numerical tuple.

B. Tree representation

The hash-key in the first column of the table can be the same for many rows (e.g. in Figure 2). On the other hand, a particular instance of initial words $w_1...w_k, k<n$ within a tuple may appear several times. The so-called *trie* structure is obtained when representing them only once, and providing a pointer towards the remaining part, and performing the same steps recursively for all the remaining columns. The corresponding edges from the root are labelled with all the hash-keys used in the first column. These edges then point towards the following vertices with outgoing edges, thus representing tuples that have the same two words at the beginning etc. In order to economize the storage space, only one copy of the hash-keys from the first few columns is kept. Additional memory for the pointers is also needed. Each vertex is represented as a vector of edges. Each edge then consists of the label (hash-key) and a pointer that always points to the first son of the vertex. In this way, the number of sons for a specific vertex can be defined as the difference between the pointer for the current vertex, and the pointer for the next one. Such representation works best if the table is dense, and if it has very few columns.

According to [3], it is necessary to only construct the trie from the word columns. Namely, the numerical columns are the corresponding output, and can be kept intact. Furthermore, the overall size of the trie structure must be minimal. Therefore, the sizes of the used pointers should be as small as possible. Each level of the trie structure corresponds to a word column of a table, and is kept separate from other word columns. Next, each word column has a separate address. Pointers only point to the next column. In this way, they represent an index within the next column (is the ordinal number of the entry within the column that they point to), and not an index in all nodes of the trie. At all trie levels (except for the last one) all vertices have at least one son. Therefore, it is possible to store a given pointer again as a difference between the index of the item it points to, and the index of the current (pointing) item. The difference will always be non-negative. The size of the pointer is defined as the smallest number of bytes needed to represent the difference between the number of items within the next column and the number of items within the current one. We don't need pointers for the last column. Namely, its indexes are the same as those in the numerical part of the tuple. Let's e.g. try to access entry $T(w_1...w_i)$. First, the value $N(w_1)$ is calculated, and then follows the search for it within the first column. When the value is not found, then the searched entry $T(w_1...w_i)$ is not stored in the tuple. On the other hand, if the value $N(w_1)$ is found, the next value $N(w_2)$ has to be calculated and then searched in the specific portion of the second column. This portion is defined by the pointer found by $N(w_1)$. The portion end is defined by the pointer at the next hash-key value in the first column. The process continues into the next columns in the same way until reaching the hash key of the last word (or fail). The index of the hash-key for the word in the last column also represents the index in the numerical part of the tuple. Binary search is best to use to find the appropriate keys in specific portions of the word columns.

A special case are those tuple structures $T^{i,j}$, where *i=1*. In this case there is no need to store the hash-keys of the words in the first word column. As the first word column is also the last one, there are also no pointers. Each hash-key of the first word column is just an index to the numerical part of the tuples.

C. Representation of real numbers

Especially in the case of N-gram language models, the number columns containing the N-gram probabilities and the backoff-weights, that demand most space. Therefore, their compact representation is even more important. Further, different computer platforms represent real numbers in a different way, using various precision. Therefore, porting numbers from one computer to another many times also results in loss of precision. The precision of a representation can be increased when we use more bytes. But in this case, the goal is also to achieve as compact a representation as possible (in the case of real numbers). Knowledge sources for the LVCSR decoder, in general, contain N-gram language models with real numbers that are frequently represented in textual form (e.g. ARPA language model). Obviously, loss of precision in this case has already happened and the precision of the representation as used in the LVCSR decoder cannot be any higher. When considering this, it is possible to specify the precision of the real numbers within the tuple data, based on the number of digits in the mantissa. In the case of ARPA language models, it is assumed that only the digits presented in the textual form of a number are significant. Then, each real number in a specific number column is decomposed into a normalized mantissa *m* and an exponent *t*, such that $r = m \cdot 2^t$, $|m| \le 0.5$ or $m = 0.0$. Let $\hat{m} = a_0.a_1...a_n$ be a representation of a mantissa *m*, where $\hat{m} = \sum_{i=0}^{n} a_i \cdot 2^{-i}$. The precision ε is then 2^{-n-1} (the biggest number, where $|\hat{m} - m| < \varepsilon$). In this way, at least $\lceil n/8 \rceil + 1$ bytes are needed to represent the mantissa with precision ε. The number of needed bytes for the exponent is also calculated, but in all practical applications it is one [3]. In the following sections, the SPREAD LVCSR decoder is presented in detail, especially by describing the implementation

of knowledge sources and the ASR search-network when using tuples.

V. LVCSR DECODER – SPREAD –

The SPREAD LVCSR decoder has been designed with a high degree of modularity. Figure 3 presents the high-level architecture of the decoder. The decoder architecture consists of four main blocks that are defined, or controlled depending on the specific application in mind. The code within each module is modularly and flexibly structured, thus enabling flexible configurations of the decoder engine.

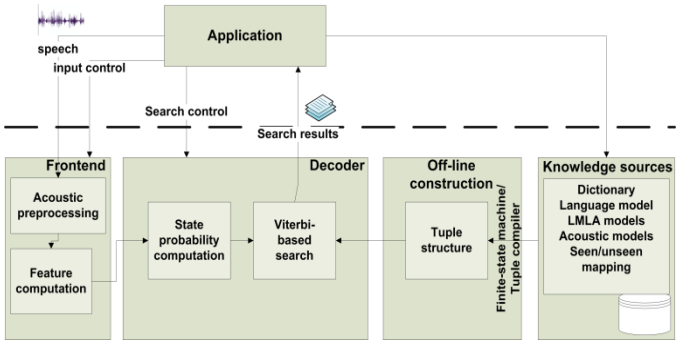


Fig. 3. Architecture of the SPREAD LVCSR decoder, based on the tuple structures.

By using tuple structures, the language-dependent knowledge sources are separated from the decoder. Furthermore, the proposed methodology of constructing one compact tuple structure (large ASR search-network, with an N-gram language model and LMLA information included), is performed off-line. Within the runtime decoder, the tuple structure is then loaded and used within the Viterbi-based search engine. Since the loading of the compact tuple structure is fast, even for large knowledge sources, the decoder is able to switch between several knowledge sources quickly and efficiently - even within a runtime system. All LVCSR decoder modules are written in C++ programming language. The off-line methodology for constructing a tuple structure from knowledge sources is performed by set of Perl[1] scripts, using several proprietary C++ FSM tools, as presented in the next sections.

Specific application defines the knowledge sources that, in general, consist of lexical, phonetic, and acoustic knowledge. The *lexical knowledge* consists of known words, along with their corresponding pronunciations. Additionally, multiple pronunciations can be included with a prior probability for each pronunciation variant. The *phonetic knowledge* consists of fundamental units within the pronunciation lexicons that are modelled in the context of their neighbours. In this way they account for the systematic and contextual variations that can be found in natural spoken speech across word boundaries. The *acoustic knowledge* is described by way of the state emission probability density functions (PDF) associated with each state of each context-dependent phoneme. Several parameters tying schemes can be used in estimation of emission PDF. The frontend module takes care for acoustic pre-processing, and the parameterisation of the speech data. The SPREAD LVCSR decoder-block then performs the recognition. The decoding problem within the system is to find the most likely word sequence $W_1^n = w_1, w_2, ..., w_n$, given a sequence of acoustic observation vectors $O_1^T = o_1, o_2, ..., o_T$, obtained from the speech signal. According to the theory in [14], this can be described by the following equation:

$$\hat{W} = \arg\max_{W_1^n} \left\{ P(W_1^n) \cdot P(O_1^T | W_1^n) \right\} = \arg\max_{W_1^n} \left\{ P(W_1^n) \cdot \sum_{S_1^T} P(O_1^T, S_1^T | W_1^n) \right\} \quad (1)$$

where W_1^n stands for the sequence of words, $S_1^T = s_1, s_2, ..., w_T$ represents any state sequence of length *T*, and $P(W_1^n)$ comprises the language model (LM) representing prior linguistic knowledge independently of the observed acoustic information. In the SPREAD LVCSR decoder, this is carried out by using a stochastic N-gram, where word probabilities are only dependent on the N-1 predecessor, and $P(O_1^T | W_1^n)$ represents the model of the lexical, phonetic, and acoustic knowledge. A complete search through such a space is still practically infeasible. Therefore, a number of approaches exist that try to solve this decoding problem. In the SPREAD LVCSR decoder, a time-synchronous search approximates the solution of the previous equation, by searching only for the most probable state sequence:

$$\hat{W} \cong \arg\max_{W_1^n} \left\{ P(W_1^n) \cdot \max_{S_1^T} P(O_1^T, S_1^T | W_1^n) \right\} \quad (2)$$

Decoding within the SPREAD LVCSR decoder performs a time-synchronous search of a network of hypotheses. At each time-step only the best hypotheses arriving at each state are retained and, in order to improve the efficiency, only the most likely hypotheses are extended to the next time-step. As already mentioned, the decoder block does not construct the ASR search-network within the runtime system. Namely, it is constructed off-line in the form of one common tuple structure that is loaded into the system during initialisation, or changed any time during the on-line process. The final tuple structure combines a standard N-gram language model, pronunciation dictionary, Look-Ahead information, and seen/unseen triphones mapping info. The decoder block is based on the token-passing algorithm with beam-search, and histogram pruning. At run time, the decoder expands the model-level tuple structure-based network into a state-level network that is suitable for finding the best state-level path. The search module requires likelihood scores for any current feature vector, in order to generate the active list. The likelihoods are computed by the state probability computation module that has access to the feature vectors.

VI. FSM TOOLS FOR THE CONSTRUCTION OF TUPLE STRUCTURES

An important advantage when using tuples for speech decoding is that they enable the integration and optimisation of several knowledge sources under the same generic representation. The proposed methodology for compiling knowledge sources into common tuple structure is performed by using proprietarily developed FSM tools, based on the theory and tuple technology as proposed in [3]. In Figure 4, the *fsmbuild*, *fsmhash*, and *fsmtuple* are those tools needed for

[1] http://www.perl.org/

compiling ASR knowledge sources into a corresponding tuple structures. Each ASR knowledge source can be split into N word columns, and M data columns. Further, the input data has to be sorted. Then perfect hash automata are built for word columns (by using the *fsmbuild* tool). In this way, a finite-state automaton is obtained that recognizes all words within individual word column (representing e.g. the triphones, diphones etc.) of the given knowledge source.

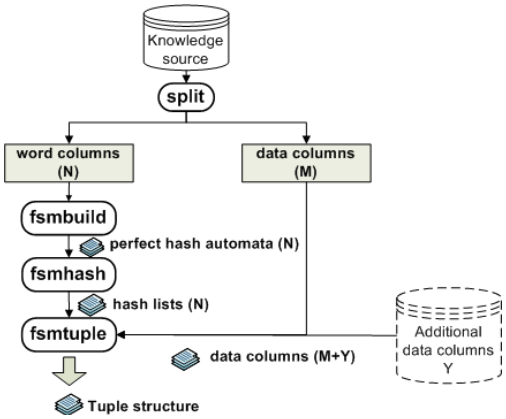


Fig. 4. FSM tools used for compiling ASR knowledge sources.

The perfect hash automata provide the mapping between words and a range of integer numbers – hash-keys. The exact numbering of the words is important for the tuple construction process. The perfect hash-automaton is at the end written into the file in binary form: a table of structures corresponding to arcs, with each arc containing a label, the number of arcs that lead from the node the arcs point to, and the index of the first arc that leads from the node the given arc points to. The *fsmhash* tool is used for translating words in specific word columns of the given knowledge source (e.g., diphones, triphones, words, etc.) into unique hash-keys. The input to the tool represents N built perfect hash automata and the corresponding N word columns' lists created beforehand. The outputs are N hash lists. In any step within the SPREAD LVCSR decoder, the mappings from hash numbers back into strings, and vice versa, can be easily and efficiently performed using these perfect hash automata. The N hash lists and additional data columns (containing integer or real numbers) are stored as table forms. At the end, the *fsmtuple* tool creates a compact structure, named the tuple structure. As can be seen from Figure 5, the input for the tuple construction process is represented in table form (*.llist), consisting of *N* columns representing words (as hash-keys), and several numbers' (integer or real) columns *M+Y*, representing tuple's data. The number of all columns *n* has to be specified, and the number of word columns *w* in a table. Furthermore, the size of the mantissa *s* can be specified (or calculated). The hash-keys for words have already been computed before, using the *fsmhash* tool. Therefore, the first step is to find the sizes of these hash-keys for each word column. Then the size of numbers in the numerical part is determined. The numerical part can contain integer or real numbers. The mantissa and exponent are calculated in the case of real numbers. The size of the whole number is, in this case, the sum of the mantissa size and the exponent size. In the case of integer numbers, the size of the numbers is just calculated. All these sizes are calculated as the number of needed bytes for storing the numbers within a specific column. In this way, only so many bytes as needed are used, to correctly represent any float or integer number within the columns of the table. Next, the tuple is constructed from the input table and written into the file. All data are written as bytes. Therefore, dedicated functions for converting numbers into bytes are used. Their input arguments are corresponding number and the calculated number of bytes that has to be used for its representation in bytes. As shown in Figure 6, the header is first written into tuple, containing e.g. the version, the word/number structure as described in the table, etc. Then, the sizes of the numbers for each column and sign vector (columns can also contain negative numbers) are written. Next, the calculated mantissa size is written for each column in the numerical part.

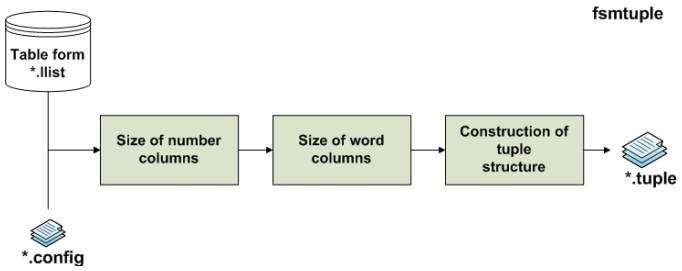


Fig. 5. The steps used during the tuple-structure construction process.

All data stored in the tuple structure is needed in order to correctly restore any number from bytes. Then, follows the construction of the tree structure: creation of the root node, with a list of pointers (1 for each child). These pointers point to records (ordered lists) of several fields e.g. hash-key, reference to a subtree etc. At the end, the indexes for the whole tree are calculated. Namely, the individual nodes of the tree are accessible via pointers from their parents. Nevertheless, in the tuple the pointers are replaced with indexes, being the ordinal numbers of the nodes within the corresponding layers. Then, the size of the whole tree is calculated (in bytes) and written into the file. Based on the indexes calculated and stored before, it is now possible to calculate and store addresses (at the byte level) of all columns in the tuple. This step is only needed when there is more than one word column in the table. At the end, the tree is also written into the tuple file (tree node's IDs, corresponding numerical part as data etc.). In this way, the tree and the corresponding numerical data are represented in the form of bytes. Such a structure is also easily stored as a binary file. In *fsmtuple* tool's configuration file, only the number of all columns has to be defined, and a number of word columns within the table. Additionally, the developer can optionally specify the size of the mantissa (can also be calculated), the desired separator between the columns in the table, the desired precision for the real numbers, the tuple's output file name, etc. All FSM tools are written in C++ programming language.

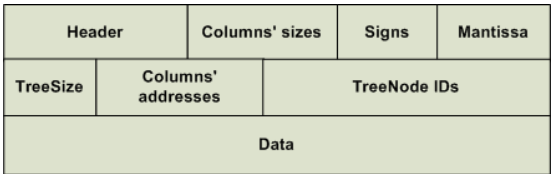


Fig. 6. Binary representation of the tuple structure.

VII. Application of Tuples To The Spread Lvcsr Decoder

In this section we propose a novel design for a one-pass LVCSR decoder engine SPREAD, based on tuple structures. The application's specific knowledge sources and ASR search-network are represented in the form of tuple structures that are combined within compact tuple-based decoding network. All these steps can be performed off-line. The detailed architecture of the LVCSR decoder SPREAD is presented in Figure 7.

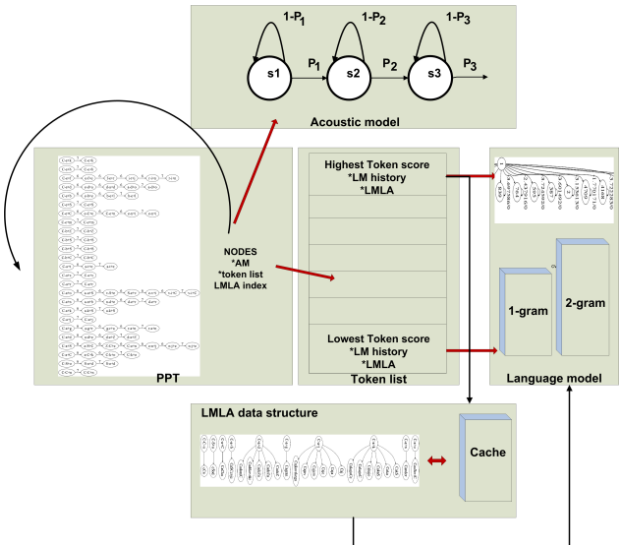


Fig. 7. The architecture of the tuple-based LVCSR decoder SPREAD.

The Viterbi search of the decoder is implemented using the token passing paradigm [27]. Hidden Markov Models (HMM) applying Gaussian Mixture Models (GMM) and the n-gram back-off language models are used to calculate the acoustic likelihoods of the context-dependent phones, and to calculate the language probabilities, respectively. The HMMs are organized within a static pronunciation prefix tree (PPT), as described in [8]. Each token contains a pointer to its LM history. Tokens coming from the leaves of the PPT are fed back into the root node of the tree after their *n*-gram history is updated. Token collisions will only occur for tokens with the same LM history. This means that each HMM state of each node in the PPT can contain a list of tokens with unique *n*-gram histories. These lists are sorted in descending order of the token probability scores.

Furthermore, decoders that make use of token-passing, restrict their search-space by various types of token pruning. In PPT-based decoders the global pruning and word-end pruning are commonly used [8]. Within the LVCSR decoder SPREAD both beam pruning methods are supported. In the case of beam pruning, tokens with a probability value between the best found probability and the best probability minus a constant beam are retained at each time-frame. All those tokens that do not fall within this beam are deleted.

During global beam pruning all tokens of the entire PPT are also compared to the best scoring token, and pruned if necessary. Word-end beam pruning is performed on all tokens that are at the leaves of the PPT, and for which the LM probabilities are incorporated into their probability scores. This pruning method is used to limit the number of tokens that are fed back into the root node of the PPT. Histogram pruning can also be used in the LVCSR decoder. Here, only the best N tokens are retained, when the number of tokens exceeds the maximum N (we significantly restrict required memory). Similar to beam pruning, histogram pruning can be performed both globally (global histogram pruning), and also in the leaves of the tree (word-end histogram pruning).

By using the proposed language model look-ahead (LMLA) technique based on tuple structures, it is possible to increase the number of tokens that can be pruned without any loss of decoding precision. It is well-known that, in the case of token-passing decoders that use PPT, full *n*-gram LMLA considerably increases the needed number of language model probability calculations. The SPREAD LVCSR decoder uses a full n-gram LMLA with a single static PPT, which is based on the tuple structures and efficient caching mechanism. Additionally, an LMLA index is assigned to each PPT's node, and an index to an LMLA field is added to each token list. The N-gram language model is also implemented in the form of a tuple structure. The language model knowledge is added to the hypothesis score at the PPT leaf nodes, and used by the LMLA mechanism.

In the following subsections the proposed methodology for constructing a compact ASR search-network based on tuple structures for a LVCSR decoder SPREAD, is presented in detail.

A. Compiling N-gram language models

Compiling N-gram language models (LM) into a tuple structure is also performed off-line. In the presented LVCSR decoder configuration, the input represents the LM N-gram language model stored in ARPA format, as shown in Figure 8. The separation into 1-gram and 2-gram data is performed first. Each file consists of word and number columns, representing unigram/bigram probabilities, and backoff weights. Next, each file is split into separate word columns and number columns, since different tasks have to be performed on each of them. Pre-processing has to be performed, in order to obtain unique and sorted lists for each word column.

The sorted word lists are then fed into the *fsmbuild* tool, and the corresponding perfect hash automata are built. In the next step the *fsmhash* tool is used, in order to translate all the words in the word columns into the corresponding lists, using hash-keys. Namely, for a final LM table form, hash keys are needed instead of words.

Furthermore, by using perfect hash automata, it is possible to translate hash-keys back into words effectively and efficiently, and vice versa, when needed. The obtained hash lists and number (data) columns are then merged into the table, by specifying the desired separator between the columns, and given to the *fsmtuple* tool. Its output then represents a LM tuple structure that has an efficient and compact trie structure. Basically, two separate tuple structures are built, and then merged into one. One structure is constructed for unigrams, and the other for bigrams.

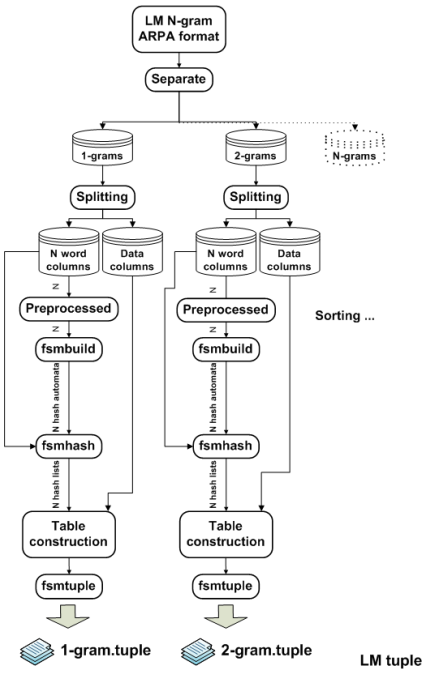


Fig. 8. Compilation process for N-gram language models.

B. Compiling the LVCSR search network

The main step within the proposed methodology for compiling ASR knowledge sources into a tuple structure represents the tuple-based construction of the ASR search-network. This structure can be constructed off-line, and is based on the idea of static PPT, and the work done in [27]. The traditional phone-level tree can be made more efficient by utilizing HMM level state tying, which has also been implemented. Cross-word triphone contexts are handled by compiling several tuple structures, with which the PPT tuple structure is merged at the end of the procedure. The obtained tuple-based network structure is in this way very compact. A general search network consists of nodes that are linked to each other with arcs.

These nodes can either correspond to one HMM state, or be dummy nodes without any acoustic probabilities associated with them. During decoding, the dummy nodes are passed immediately. They only mediate the tokens used to present the active search-network. A node can also have a word identity associated with it, which leads to the insertion of the word into the word history of the token passing that node. The proposed procedure of compiling such a LVCSR search-network into a tuple structure, assumes triphone models, where every triphone is defined in the acoustic models, and they are not tied at the triphone level. Instead, each triphone has a set of HMM states (three states in a left-to-right topology), and these states are shared amongst all triphones.

The state tying is performed using a decision tree. In this way the SPREAD LVCSR decoder is based on tuple-based network topology, including cross-word triphone models. The proposed methodology of compiling a search-network into a tuple structure follows the classical network topology idea, which is described with nodes and transition links, where the nodes are ordered in several layers [27].

Such a network also uses application specific vocabulary, and a HMM model set. The goal was to build a compact tuple representation of such a network topology, and integrate within it all the needed knowledge sources, like tuple-based N-gram language models, and tuple-based LMLA info. Construction of the LVCSR search-network to be used by the SPREAD LVCSR decoder is performed off-line, and can be repeated for any other ASR knowledge source available for application. The proposed methodology is presented in Figure 9.

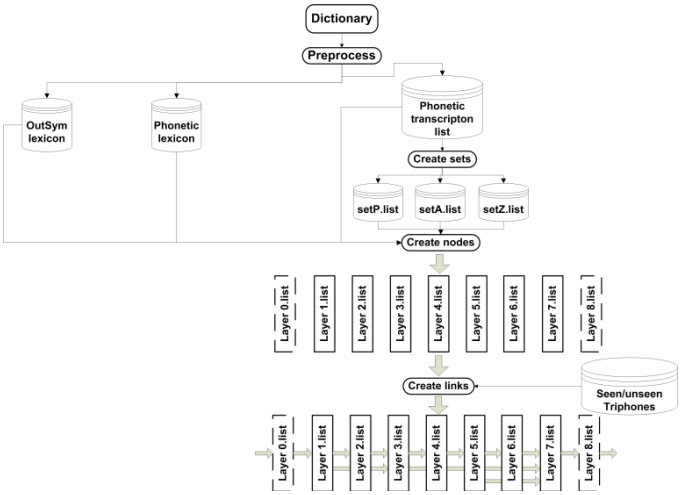


Fig. 9. The construction of the tuple-based LVCSR search network (first step).

The input represents a large dictionary. Firstly, the *outsym* and *phonetic* lexicons are built. A phonetic lexicon can be viewed as a list of word entries, where each entry contains orthography for the word and a corresponding list of pronunciations. A phonetic pronunciation in the dictionary can also contain a so-called output symbol. It is optional, but when present, the recognition output can use the specified output symbol rather than the word itself.

Therefore, an additional *outsym* lexicon can be built when this info is available. Additionally, a phonetic transcription list is built, containing only phonetic transcriptions for all the words. This list is then used for the construction of three phone sets, named *P* (all 1- phoneme transcriptions), *A* (first phonemes of all transcriptions), and ***Z*** (last phoneme of all transcriptions). All these sets are used in the next step for the construction of nodes within all layers of the network topology, here numbered from 0 to 8.

Layers 0, 2, 4, and 6 are those layers with model nodes, the other layers are used for context nodes. This step includes creation of the word final nodes (layer 0), silence (*sil/sp*) nodes, sentence start node (layer 3) and sentence end node (layer 5), word initial nodes (layer 4), PPT nodes (layer 6), word end nodes (layer 7), and other context nodes in layers 1, 3, 5, and 8. *1*-phoneme words are represented with corresponding nodes in layer 3. The model nodes are actually triphones, and other nodes are diphones. All the model nodes are firstly represented by linguistic triphones (using linguistic phonetic transcriptions from the dictionary), and then replaced by acoustic ones using seen/unseen mapping lists.

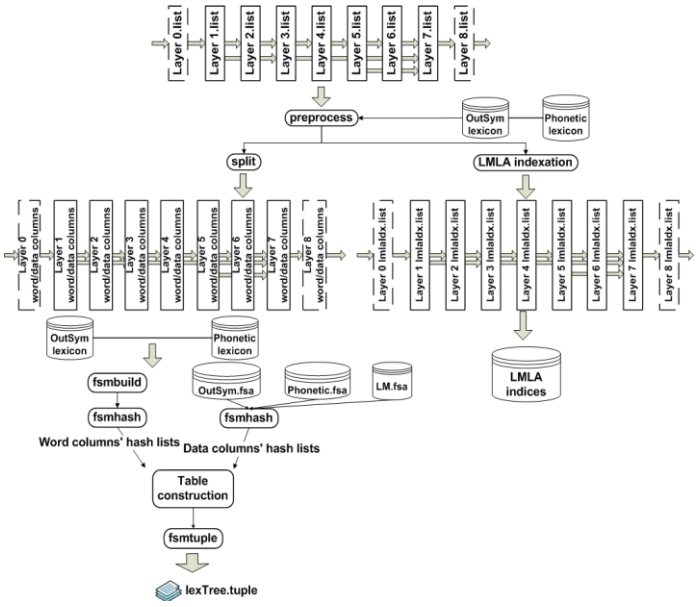


Fig. 10. The construction of the tuple-based LVCSR search network (second step).

In this way there is direct access from each model node to the corresponding HMM model stored within the HMM models' array in the runtime decoder. Context nodes are represented as diphones, and word end nodes with unique node names (layer 7). PPT (layer 6) consists of several sublayers of triphone model nodes, depending on the length of the word. As seen in Figure 9, then follows the creation of links between all these nodes in layers from 0 to 8. In this way, the last nodes within the PPT structure (layer 6) are linked with the corresponding word end nodes in layer 7. The *1*-phoneme nodes in layer 3 are linked with the corresponding word end nodes in layer 7. Furthermore, the context nodes in layer 5 are linked with the starting nodes of the PPT structure, and the model nodes in layer 4 are linked with the context nodes in layer 5. All the word end nodes are further linked with the nodes in layer 8. The final layer 8 is linked back to the corresponding model nodes in layer 0, and the model nodes in layer 0 with the corresponding context nodes in layer 1, and with the silence models in layer 2. Additionally, the context nodes in layer 1 are linked with the silence models. And the silence models in layer 1 are linked with the context nodes in layer 3. It is clear that each of such layers can be represented in the form of "word" and "number" columns. Namely, here word columns represent a node column, and a link column. Additional data (on transitions) can be added in the form of number columns, when needed. Since all layers can be represented in such a way, they can also be compiled into tuple structures. Therefore, all the constructed layers and lexicons are first split into word columns (actually diphones, triphones, ortographic, phonetic transcriptions, output symbols etc.) and translated into corresponding hash-keys (Figure 10). This is performed by using the *fsmbuild* and *fsmhash* tools. Additionally, in layers 0, 3, 4, 5, and 6, phonetic and LM information has to be added in the form of additional data columns (number columns). Next, a table form for each layer is constructed, consisting of several word and number columns (data). Finally, all the tuple structures are constructed and merged by using the *fsmtuple* tool. Additionally, LMLA indexing of all the nodes is performed. In this step all those nodes are numbered, where LMLA calculation has to be performed (the LMLA technique will be presented in the next subsection). The value 0 is assigned only to unique successors in the PPT, since in this case no LMLA calculation is needed. All nodes' LMLA indices are stored as a binary file. Within the runtime system they are loaded, and then directly accessible via hash-keys.

C. Compiling Language Model Look-Ahead Data

Calculating all the possible LM probabilities for all the tokens takes a lot of time and consumes a lot of computational resources. When the lexical network is constructed as a static tuple-based PPT, as described in the previous subsection, word identities can be determined only after there are no more branches in the tree structure. Thus, any inclusion of the language model (LM) probability is delayed until the final nodes are reached. It is well-known that by using LM probabilities in such structure as early as possible, enhances the beam pruning and, therefore, decreases the size of the search-space. This can be achieved by applying so-called language model Look-Ahead techniques. In the literature a number of methods are proposed for managing these calculations [8,16]. The least complex way for reducing the needed number of LM lookups whilst applying LMLA, is to use for the Look-Ahead only unigram probabilities. By using unigrams, the approximation of the best final LM score is less precise, but it becomes possible to integrate the corresponding Look-Ahead scores directly within the PPT. In this case, each node stores a single value: the difference between the best LM score from before and after entering the particular node. In the case of unigrams, these Look-Ahead values can be applied for all tokens, without regard to their *n*-gram history. However, it has been shown that unigram Look-Ahead is outperformed by higher order Look-Ahead systems [8]. A method that can be used for reducing the number of LM lookups has been proposed in [20]. In this case, all those PPT nodes with only one successor node are skipped when calculating the LMLA values. Their decoder used tree copies in order to incorporate the LM probabilities. Furthermore, whenever a new copy is required, the LMLA is performed on demand. In [11] at each PPT node, a special list is stored with all those words that are still reachable from that node. In the cases of small word lists, the Look-Ahead value is calculated exactly (each trigram probability is calculated, and the best one selected). Larger word lists at the PPT root node, are skipped. For all remaining lists, the intersection with the n-gram lists is calculated, before computing the corresponding LMLA values. This approach can save a considerable amount of search-time, especially for those words that do not have a trigram or bigram LM value. The proposed LMLA technique is based on tuple structures. In this approach, Look-Ahead structures are tuples that are constructed off-line. The LVCSR decoder SPREAD does not make tree copies. Instead, LM histories are stored in the tokens and the PPT tuple is shared by all the tokens. In this decoder, the language model knowledge is added to the hypothesis score at the PPT tuple's leaf nodes. Incorporating the LM model at an early stage into the tuple structure, makes it possible to compare and prune the hypotheses based on both linguistic and acoustic evidence.

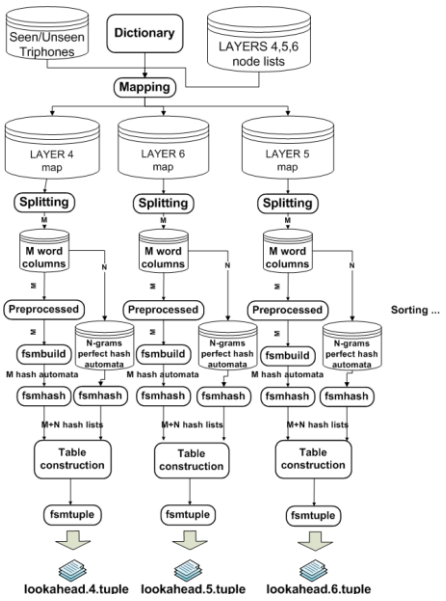


Fig. 11. Compiling LM Look-Ahead tuples for layers 4, 5, and 6.

In the SPREAD LVCSR decoder, the LMLA tuple-based mechanism in the runtime system performs calculations for each token in the tuple, the LM probabilities for all words that are reachable from that token, and temporarily adds the best one to the token's score. When the token reaches a PPT tuple's leaf node, the temporary LM probability is replaced by the probability of the word represented by the PPT tuple's leaf node. Following this procedure, sharper beams can be applied during the pruning so that fewer tokens need to be processed and, consequently, the decoding is speeded up considerably. Each node within the PPT tuple that has more than one successor, or that is a leaf node, is assigned a unique LMLA index. These indices are found in the binary file 'LMLA indices' (Figure 10). The LM Look-Ahead score is computed by finding the maximum of the LM scores over the words in the specific node's list, only when the node's LMLA index value is not 0. The words and the corresponding LM probabilities are accessed via LMLA tuple structures. Nevertheless, in order to minimize the significant amount of redundant computations involved in the LM Look-Ahead, a caching structure is also part of the LMLA process within a runtime system. The caching structure contains the Look-Ahead values for those tokens with a particular language model history. In this way, for each node the maximum LM scores of the possible follow-up words are stored for specific word histories. The LMLA index of a specific node then points to these corresponding LMLA probabilities in cache. Using this method, each node's LMLA probability is exactly calculated once. Therefore, in the case of a cache miss, the probabilities of all the words in the LM for the given word history are computed and stored to the cache. The LM Look-Ahead is applied only in those nodes where the list of possible word identities has changed from that of the previous nodes. Reducing the number of nodes in which LM Look-Ahead is applied also helps to save memory when node level caching is involved. Figure 11 illustrates the compiling of LMLA tuple structures to be used in the decoder. The input represents the seen/unseen triphones' info, the dictionary, and the nodes from layers 4, 5, and 6, where LMLA has to be performed. Seen/unseen triphones' info is needed in order to link the acoustic triphones' nodes in these layers with the dictionary entries, as used in unigrams, and bigrams. After the LVCSR search-network (Figure 10) has been built, the node lists for layers 4, 5, and 6 can be created (consisting of diphones and triphones). The next step is the mapping. Based on the seen/unseen triphones' info, the dictionary, and the node lists, for each layer corresponding maps are constructed, containing all the word identities that are reachable from each node in those layers. These maps are actually tables consisting of diphones or triphones in the first column, and corresponding possible words in the second. The next step is to split each map file into *M* separate word columns (one column contains layers' nodes, and the other corresponding words from dictionary). Next, perfect hash automata are constructed for the *M* separate word columns, using the *fsmbuild* tool. Then all the entries in the *M* word columns are translated into hash-keys by using the *fsmhash* tool. Since the LM Look-Ahead structures are constructed off-line, LMLA values cannot already be stored directly within the LMLA tuple structure. Instead, N data columns are created, containing hash-keys for the corresponding N-grams, by using N-grams perfect hash automata and the *fsmhash* tool. In this way, direct access to LM scores is possible in the online LVCSR decoder, when the LM histories are also known. Now, the corresponding tables for all layers can be constructed, containing *M* separated word columns, and *N* number columns (unigram and bigram hash keys). Finally, the tables are compiled into tuple structures. In this way, three tuple structures are obtained. Within the runtime system they are accessed in layers 4, 5, and 6.

VIII. Results

The LVCSR decoders used today employ acoustic models, pronunciation lexicon, N-gram language models, and other linguistic sources. An approach using efficient and compact tuple structures was proposed in the paper, for a construction of the LVCSR search network. As presented, tuple structures can be implemented as ordinary dictionaries. Namely, the elements within the tuple structures of a given key are concatenated with a selected separator symbol. This also means that a standard implementation of dictionaries can be employed based on perfect hash. The benefits are foremost, the important space savings and higher processing speed (automata), and the compact and reduced size of the tuple structure, especially when the structure of the key can be exploited (depending on the used knowledge sources). In this way, the time needed to load LVCSR search network into the memory is practically instantaneous. Furthermore, fast switching between several applications' specific knowledge sources is possible, since the LVCSR search network is already constructed off-line, and just loaded within the runtime system.

As presented in this paper in detail, application specific ASR knowledge sources can be compiled into tuple-based LVCSR search-network. All the needed steps are accomplished by using several Perl scripts, with proprietary FSM tools, developed in the C++ programming language. The whole procedure is completed within a matter of minutes on a PC with Intel Core 2 Quad CPU, 2.83 GHz, with a 4 GB RAM.

The largest part is spent compiling the *N*-gram language model into the tuple structure. Overall, the whole compiling procedure is simple, fast, without a large memory, or processor requirements. In the experiment, the following ASR knowledge sources were used: context-dependent acoustic models (triphones), a dictionary, and an interpolated bigram language model. The dictionary contained 64K words, and the bigram language model consisted of 64K unigrams, and of approx. 7M bigrams. The proposed methodology for compiling ASR knowledge sources into a tuple structure, can also be used in the same way for higher-order language models (if available), and for other application specific knowledge sources, and languages.

Table 1 presents the statistics about the layers' nodes of the tuple-based LVCSR search network for this speech recognition task. Table 2 then presents statistics about the nodes of the tuple-based language models, and Table 3 presents statistics about the nodes in the tuple-based LMLA structures. These data are based on table forms constructed by using available knowledge sources.

TABLE I. THE LAYERS' NODES IN THE TUPLE-BASED LVCSR SEARCH NETWORK

Layer	0	1	2	3	4	5	6	7	8
Nodes	8,155	650	651	651	7,221	497	1,254,739	64,874	650

TABLE II. THE NODES IN THE TUPLE-BASED LM MODEL

N-gram	1-grams	2-grams
Nodes	64,000	127,696

TABLE III. THE NODES IN THE TUPLE-BASED LMLA STRUCTURES

LMLA	LMLA – layer 4	LMLA – layer 5	LMLA – layer 6
Nodes	71,183	64,457	66,645

TABLE IV. THE COMPACT SIZES OF TUPLES USED FOR LVCSR SEARCH NETWORK

Layer	0	1	2	3	4	5
Tuple	304kB	3.93kB	11.2kB	274kB	182kB	81.8kB
Layer	6	7	8			
Tuple	5.87MB	380kB	47.1kB			

TABLE V. THE COMPACT SIZES OF TUPLES USED FOR LM MODEL

N-gram	1-grams	2-grams
Tuple	812kB	63.4MB

TABLE VI. THE COMPACT SIZES FOR TUPLES FOR LMLA DATA

LMLA	LMLA – layer 4	LMLA – layer 5	LMLA – layer 6
Tuple	9.43MB	438kB	440kB

All the table forms additionally contain several data columns (number columns) that are used within the ASR system. The tables 4-6 then represent the achieved compact sizes of the tuples after compiling constructed table forms. The sizes reported in the tables are the sizes of the final compiled files. The overall size of the merged tuple structure loaded for the specific task by the SPREAD LVCSR decoder is 81.234 MB for the 64k LVCSR task.

The same task was also tested by HDecode [27]. In the case of HDecode, the loading of knowledge sources prepared in their format and construction of internal ASR structures, took 50 times longer (since all the structures for the LM, LMLA and LVCSR search-network has to be constructed during initialisation). Furthermore, a set of 100 audio files was recognized by using both decoders in order to evaluate whether the tuple-based decoder also showed any benefits regarding the processing speed. In both systems the same configuration was performed in order to compare the obtained results. In the case of the SPREAD LVCSR decoder, approx. 20% higher processing speed was achieved, without loss of recognition accuracy. All the experiments were performed on a PC with Intel Core 2 Quad CPU, 2.83 GHz, with a 4 GB RAM.

IX. CONCLUSION

This paper presented the novel design of a LVCSR decoder engine, named SPREAD. This LVCSR decoder is based on a time-synchronous beam search approach. The ASR search network includes statically expanded cross-word triphone contexts. An approach using efficient tuple structures was proposed and presented, for constructing a complete ASR search-network. These data structures were motivated by practical applications in speech and language processing. The used technique for compact representation of tuple structures can be seen as an application and extension of perfect hashing by means of finite-state automata. Therefore, the benefits are foremost the important space savings and higher processing speed. Furthermore, the advantage of the proposed LVCSR decoder implementation based on tuple structures is the compact and reduced size of the tuple structure, especially when exploiting the structure of the key (*n*-tuples of strings). Therefore, the time needed to load an ASR search-network into the memory is also significantly reduced. Further, in the paper the complete methodology of compiling general ASR knowledge sources into a tuple structure (representing an ASR search-network) was proposed and presented. It has been shown that ASR knowledge sources can be implemented by ordinary dictionaries, where the elements in the tuple of a given key are concatenated with a specific separator symbol of our choice. Therefore, a standard implementation of dictionaries can be employed, typically a hash table or perfect hash.

Furthermore, the beam search was enhanced with a novel implementation of bigram language model Look-Ahead technique, by using a tuple structure and a caching scheme. The SPREAD LVCSR decoder is based on a token-passing algorithm and is able to restrict search-space by several types of token pruning. By using the presented language model look-ahead (LMLA) technique, it is possible to increase the number of tokens that can be pruned without any decoding precision loss.

REFERENCES

[1] Aubert, X. An overview of decoding techniques for large vocabulary continuous speech recognition. Computer Speech & Language, Volume 16, issue 1, pp. 89-114, 2002.

[2] Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., Stolcke, A. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Trans. Speech Lang. Process. Volume 5, issue 1, Article 3 (December 2007), pp. 1-29, 2007.

[3] Daciuk, J., van Noord, G. Finite automata for compact representation of tuple dictionaries, Theoretical Computer Science, Volume 313, Issue 1, pp. 45-56, 16 February 2004.

[4] Dixon, P., Caseiro, D., Oonishi, T., Furui, S. The Titech Large Vocabulary WFST Speech Recognition System. In Proc. ASRU, pp. 1301-1304, 2007.

[5] Evermann, G., Woodland, P. C. Design of fast LVCSR systems. In Proc. ASRU'03, 2003, pp. 7–12, 2003.

[6] Fujii, Y., Yamamoto, K., Nakagawa, S. Large vocabulary speech recognition system: SPOJUS++. In Proceedings of the 11th WSEAS international conference on robotics, control and manufacturing technology, and 11th WSEAS international conference on Multimedia systems & signal processing (ROCOM'11/MUSP'11). S. Chen, Nikos Mastorakis, Francklin Rivas-Echeverria, and Valeri Mladenov (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, pp.110-118, 2011.

[7] Hoffmeister, B., Heigold, G., Rybach, D., Schlüter, R., Ney, H. WFST Enabled Solutions to ASR Problems: Beyond HMM Decoding. IEEE Transactions on Audio, Speech, and Language Processing. Volume 20, number 2, pp. 551-564, February 2012.

[8] Huijbregts, M., Ordelman, R., de Jong, F. Fast N-gram Language Model Look-Ahead for Decoders with Static Pronunciation Prefix Trees. In Interspeech 2008, 9th Annual Conference of the International Speech Communication Association, pp. 1582-1585, Brisbane, Australia, September 22-26, 2008.

[9] Lee, A., Kawahara, T., Shikano, K. Julius - an open source real-time large vocabulary recognition engine. In proceedings of Eurospeech 2001, pp. 1691-1694, 2001.

[10] Liu, X., Gales, M. J. F., Woodland, P. C. Improving LVCSR system combination using neural network language model cross adaptation. In Interspeech 2011, pp. 2857-2860, Florence, Italy, August 2011.

[11] Massonie, D., Nocera, P., Linares, G. Scalable language model look-ahead for LVCSR. In proceedings Interspeech 2005, Lisbon, Portugal, pp. 569-572, 2005.

[12] Mohri, M., Pereira, F., Riley, M. Weighted finite-state transducers in speech recognition. In: Computer Speech and Language 16, pp. 69-88, 2002.

[13] Moore, D., Dines, J., Magimai Doss, M., Vepa, O., Cheng, O., Hain, T. Juicer: AWeighted Finite State Transducer Speech Decoder. In Proc. Interspeech, pp. 241-244, 2005.

[14] Moore, D., Dines, J., Magimai-Doss, M., Vepa, J., Cheng, O., Hain, T. Juicer: A Weighted Finite-State Transducer Speech Decoder. In MLMI'06, 3rd joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, pp. 285-296, 2006.

[15] Nguyen, P. Techware: Speech recognition software and resources on the web. IEEE signal processing magazine. Volume 26, number 3, pp. 102-105, 2009.

[16] Nolden, D., Schlüter, R., Ney, H. Acoustic Look-Ahead for More Efficient Decoding in LVCSR. In Interspeech 2011, pp. 893-896, Florence, Italy, August 2011.

[17] Novak, J. R., Dixon, P. R., Furui, S. An empirical comparison of the t3, juicer, HDecode and sphinx3 decoders. In Interspeech 2010, pp. 1890-1893, 2010.

[18] Novak, J. R., Minematsu, N., Hirose, K. Painless WFST cascade construction for LVCSR – transducersaurus. In Interspeech 2011, pp. 1537-1540, Florence, Italy, August 2011.

[19] Novak, J. R., Minematsu, N., Hirose, K. Open Source WFST Tools for LVCSR Cascade Development. Finite-State Methods and Natural Language Processing, 9th International Workshop, FSMNLP 2011, pp. 65-73, Bois, France, July 12-16, 2011.

[20] Ortmanns, S., Ney, H., Eiden, A., Coenen, N. Look-ahead techniques for improved beam search. In proceedings of the CRIM-FORWISS Workshop, Montreal, pp. 10-22, 1996.

[21] Parada, C., Dredze, M., Sethy, A., Rastrow, A. Learning Sub-Word Units for Open Vocabulary Speech Recognition. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

[22] Pylkkönen, J. An Efficient One-pass Decoder for Finnish Large Vocabulary Continuous Speech Recognition, Proceedings of the 2nd Baltic Conference on Human Language Technologies (HLT'2005), Tallinn, Estonia, pp. 167-172, April 4-5, 2005.

[23] Rojc, M., Kačič, Z. Time and Space-Efficient Architecture for a Corpus-based Text-to-Speech Synthesis System, Speech Communication, Vol. 49 (3), pp. 230-249, 2007.

[24] Rybach, D., Hahn, S., Lehnen, P., Nolden, D., Sundermeyer, M., Tüske, Z., Wiesler, S., Schlüter, R., Ney, H. RASR - The RWTH Aachen University Open Source Speech Recognition Toolkit. In IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Hawaii, USA, December 2011.

[25] Sixtus, A., Ney, H. From within-word model search to across-word model search in large vocabulary continuous speech recognition. In: Computer Speechamd Language 16, pp. 245-271, 2002.

[26] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, K. Sphinx-4: A flexible open source framework for speech recognition. Sun Microsystems Technical Report, No. TR-2004-139, Sun Microsystems Laboratories, 2004.

[27] Young, S., Everman, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V.,Woodland, P. The HTK Book. Cambridge University Engineering, 2006.

Prediction of Satellite Motion under the Effects of the Earth's Gravity, Drag Force and Solar Radiation Pressure in terms of the KS-regularized Variables

Hany R. Dwidar
Astronomy, Meteorology and Space Science Dept.
Faculty of Science - Cairo University
Giza - EGYPT

***Abstract*— This paper is concerned with an orbit prediction using one of the best regular theories (KS-regularized variables). Perturbations due to the Earth's gravitational field with axial symmetry up to the fourth order zonal harmonic, atmospheric drag (variation in density model with height) and solar radiation pressure are considered. Applications of the problem with a comparison between the perturbations effect will be illustrated by numerical and graphical example.**

Keywords—KS-regularized variables; orbit determination; Numerical Modeling

I. INTRODUCTION

It is well known that the solutions of the Classical Newtonian Equations of motion are unstable and these equations are not suitable for long-term integrations. Many transformations have emerged in the literature in the recent past to stabilize the equations of motion either to reduce the accumulation of local numerical errors or allowing of using a larger integration step size, in the transformed space, or both.

Examples of such transformations include the use of a new independent variable-time transformation, transformation to orbital parameter space which tends to decouple fast and slow variables, and the use of integrals as control terms. One of such transformation, known as the KS-transformation, is due to Kustaa-neimo and Stiefel, who regularized the non-linear Kepler motion and reduced it to linear differential equations of a harmonic oscillator of constant frequency. Reference [29] further developed the application of the KS-transformation to problems of perturbed motion, producing a perturbational equations version ([1] ; [3] ; [4] ; [13] ; [14] ; [15] ; [20] ; [21] ; [23] ; [28] ; [30]; [31]; [32] ; and [33]).

Space vehicles (including artificial Earth satellites) are subjected to a number of disturbing forces which are classed as non-gravitational forces. These non-gravitational forces are, for example, atmospheric drag, solar radiation pressure, drag on a charged satellite and meteorite collisions. Aside from the effects of the Earth's imperfect shape, the largest perturbative force on a space vehicle close to the Earth is caused by the atmosphere. Whenever a space vehicle passes within about 800 Km of the Earth's surface, it is subjected to a dissipative force induced by motion through the Earth's atmosphere.

Most of the other non-gravitational forces acting upon a space vehicle are negligible with respect to the effect of the Earth's oblateness and atmosphere when the vehicle is close to the Earth.

Getting high in the atmosphere (above 600 Km) the solar radiation pressure force is more important than atmospheric drag. As the vehicle enters inter-planetary space, the previously neglected perturbations become increasingly more important as the space vehicle leaves the region of the Earth's influence.

The drag acceleration causes a distortion in the shape of the orbit and a continuous loss of the kinetic energy of the satellite, to the atmosphere (e.g., [11]). If the atmosphere were stationary, the orientation angles would have not been affected. But due to the rotation of the atmosphere the velocity of the satellite relative to the atmosphere differs from its initial velocity. Consequently, the drag force vector will not lie in the plane of the unperturbed motion and therefore, all six orbital elements will be affected. The net result is:

1) a secular variation of the orbital elements, and

2) a drop in orbital altitude which increases the potential energy to compensate the drop in kinetic energy.

This effect is largest at perigee where the density of the atmosphere is maximum (along the orbit), and is reflected as a decrease in altitude at the next apogee passage. The result is that apogee altitudes decrease more rapidly than do the perigee altitudes. Thus an elliptic orbit will tend to become circular, while an initially circular orbit with uniform drag over its entire path will tend to remain nearly circular and decays through a nearly spiraling trajectory.

The interest in studying the effects of radiation pressure on the motion of artificial satellites has been initiated by the discrepancies between theory and observations of the balloon-type satellites. The effect due to direct solar radiation pressure exceeds that of atmospheric drag at a height of 800 Km with a force magnitude of 10^{-5} dyne/cm ([27] and [26]) and is particularly emphasized for balloon-type satellites for which the area to mass ratio is large. Certain such satellites changes shape from spherical to spheroidal shapes, producing a component of force at right angles to the Sun-satellite direction ([18]).

The solar radiation pressure force becomes a discontinuous function of time when the satellite enters the Earth's shadow.

Reference [19] derived first order expressions for the rates of change in the osculating elements caused by solar radiation pressure by the method of variation of vector elements, shadow effects were not taken into account. References [5] and [12] used Lagrange's planetary equations to find first order solutions, with the integrations performed between the times of exit and entry into the shadow. The resonance effects produced by the commensurabilities between the different mean motions gave good field for detailed theoretical studies (e.g., [9] and [22]). The effect of solar radiation pressure are analyzed in four very useful and interesting expositions given by [10] , [25] and [26] who discussed it (as one of the non-gravitational forces) from all its different aspects and [17] which analyzed in great detail the effects produced both by the direct and albedo radiation pressures on both spherical satellites as well as those of complex shapes.

Reference [24] derived the components of the force in the directions of the radius, normal to it in the direction of motion and normal to the orbit plane, the shadow effect is considered and the effect of diffuse radiation pressure were to be about 1/100 of the direct solar radiation pressure.

References [8] and [26] pointed out the practical use of a shadow function is limited by the number of terms we need to take into account which makes the integration process extremely laborious.

Further, numerical integrations show that the shadow functions give inaccurate results outside of the shadow cylinder since in this region the function is no longer equal to one and the effect is as though the satellite is in the shadow.

Reference [7] studied the behavior of a particle moving under the effect of central attraction and perturbed by the constant radiation pressure. He obtained evidence for the existence of a surface of stable circular orbits with centers on an axis through the primary body and derived the necessary & sufficient conditions for the existence of stable circular orbits when taking the primary's shadow into account.

Also, [16] studied the Kepler problem including radiation pressure and drag, the secular and vector integrals of motion are obtained and [8] pointed out the importance of both solar radiation pressure and atmospheric drag in a first order theory of some satellites.

In this paper, we use the method of fourth order Rung-Kutta method to predict the motion of a satellite under the perturbation effects the Earth's gravitational field with axial symmetry up to the fourth order zonal harmonic, atmospheric drag (variation in density model with height) and solar radiation pressure by using KS-regularized differential equation. we compare graphically the influence of each perturbation.

II. Formulate The Problem

The equations of motion of an artificial satellite are given generally as

$$\ddot{\vec{x}} + \frac{\mu}{r^3}\vec{x} = -\frac{\partial V}{\partial \vec{x}} + \vec{P}, \qquad (2.1)$$

where $\vec{x}$ is the position vector in a rectangular frame (the physical frame), $r = |\vec{x}|$ is the distance from the origin, μ is the Earth's gravitational constant, V is the perturbed time independent potential and $\vec{P}$ is the resultant of all non-conservative perturbing forces and forces derivable from a time dependent potential.

The potential of the Earth's gravity with axial symmetry can be written as

$$V = \mu \sum_{i=2}^{\infty} R^i J_i \left(\frac{1}{r}\right)^{i+1} P_i(x_3/r), \qquad (2.2)$$

where R is the Earth's equatorial radius, J_i is the non-dimensional coefficient of the Earth's oblateness and $P_i(x_3/r)$ is the Legendre polynomial of order i. In the present paper we shall assume that the potential of the Earth's gravity of the axial symmetry is taken up to the fourth order zonal harmonics J_4, then Eq.(2.2) rewrite as

$$V = \frac{3}{2}Q_2 x_3^2 r^{-5} - \frac{1}{2}Q_2 r^{-5} + \frac{5}{2}Q_3 x_3^3 r^{-7} - \frac{3}{2}Q_3 x_3 r^{-5} + \frac{35}{8}Q_4 x_3^4 r^{-9} - \frac{15}{4}Q_4 x_3^2 r^{-7} + \frac{3}{8}Q_4 r^{-5}, \qquad (2.3)$$

where $Q_i = \mu R^i J_i$, i = 2(1)4

and $r = \sqrt{x_1^2 + x_2^2 + x_3^2}$.

Since the perturbing acceleration due to air drag is expressed as

$$\vec{D} = -\frac{1}{2} C_D \frac{A}{M} \rho |\vec{v}| \vec{v} \qquad (2.4)$$

where - C_D is the non-dimensional drag coefficient depending on the satellite geometry and in most cases its value lies between 2.1 & 2.3;

- A is the effective cross-sectional area, M is the satellite mass;

ρ is the density function of the ambient gas (the atmosphere) and depends primarily on the altitude and to a lesser extent on the solar and geomagnetic activity. In this paper we'll take the most famous models of air density which is

$$\rho = \rho_0 \left[\frac{r_0 - \eta}{r - \eta}\right]^{\tau}, \qquad (2.5)$$

where ρ_0 is the value of ρ at the reference level r_0, while η and τ are two adjustable parameters. They can be adapted to the estimated or observed variations of the solar activity and periodically updated so that the dynamics of the atmosphere is taken into account. The value of η is approximately equal to the mean Earth's equatorial radius and τ equals the inverse of gradient of the density scale height and can take values in the range from 3 to 9 ([6]).

- $\vec{v}$ is the velocity of the satellite relative to the atmosphere.

Also, since the perturbing acceleration due to solar radiation pressure can be expressed as ([16])

$$F_{solar} = \beta \frac{\mu}{r^3} \vec{r} \qquad (2.6)$$

where $\vec{r}$ is the radius vector and β is a constant associated with the radiation pressure effect. The range of physically possible β, for a repulsive force, is $0<\beta<1$. For $\beta = 0$ the attracting center does not radiate at all. But for $\beta>1$ the resultant of the collinear force turns from attraction to repulsion, with the consequence that the problem is quite different from the initially stated ([16]).

Finally, the equations of motion of an artificial satellite in KS-regularized variables are

$$\vec{u}'' + \alpha_k \vec{u} = \frac{r}{2} \vec{\lambda}, \qquad (2.7.1)$$

$$\alpha_k = -<\vec{u}', \vec{\lambda}>, \qquad (2.7.2)$$

$$t' = r, \qquad (2.7.3)$$

$$r'' + 4\alpha_k r = \mu + r <\vec{u}, \vec{\lambda}>, \qquad (2.7.4)$$

where

- α_k is one-half of the negative Keplerian energy as
$$\alpha_k = \left(\frac{\mu}{2} - <\vec{u}', \vec{u}'> \right) / r ;$$
- $\vec{\lambda} = L^T(\vec{u})\vec{b} = L^T(\vec{u})\left(-\frac{\partial V}{\partial \vec{x}} + \vec{P} \right)$,
- $$L(\vec{u}) = \begin{pmatrix} u_1 & -u_2 & -u_3 & u_4 \\ u_2 & u_1 & -u_4 & -u_3 \\ u_3 & u_4 & u_1 & u_2 \\ u_4 & -u_3 & u_2 & -u_1 \end{pmatrix},$$

and

- $$r = -<\vec{u}, \vec{u}> \quad , \quad r' = 2<\vec{u}, \vec{u}'> ;$$

hence $<\vec{a}, \vec{b}>$ is used to denote the scalar product of two vectors $\vec{a}$ and $\vec{b}$. Denoting differentiation with respect to the new time *s* (knowing as the fictitious time) by a prime (′), since the independent variable is changed from time (*t*) to fictitious time (*s*) according to ([29])

$$\frac{dt}{ds} = r,$$

then for any variable ζ we have

$$\zeta' = r\dot{\zeta}.$$

III. Equations of Motion

The differential equations of motion for the satellite in KS-regularized variables under the perturbations of the Earth's gravity and air drag are

$$u_1'' = -\alpha_k u_1 + \frac{r}{2} \lambda_1, \qquad (3.1)$$

$$u_2'' = -\alpha_k u_2 + \frac{r}{2} \lambda_2, \qquad (3.2)$$

$$u_3'' = -\alpha_k u_3 + \frac{r}{2} \lambda_3, \qquad (3.3)$$

$$u_4'' = -\alpha_k u_4 + \frac{r}{2} \lambda_4, \qquad (3.4)$$

$$\alpha_k' = -u_1' \lambda_1 - u_2' \lambda_2 - u_3' \lambda_3 - u_4' \lambda_4, \qquad (3.5)$$

$$t' = r, \qquad (3.6)$$

$$r'' = \mu + r\left(-4\alpha_k + u_1\lambda_1 + u_2\lambda_2 + u_3\lambda_3 + u_4\lambda_4\right), \qquad (3.7)$$

where

$$\lambda_1 = u_1 b_1 + u_2 b_2 + u_3 b_3,$$
$$\lambda_2 = -u_2 b_1 + u_1 b_2 + u_4 b_3,$$
$$\lambda_3 = -u_3 b_1 - u_4 b_2 + u_1 b_3,$$
$$\lambda_4 = u_4 b_1 - u_3 b_2 + u_2 b_3;$$

and we have two forms of *b's*; one with drag force only and the second with drag and solar radiation pressure; of course under the Earth's gravity.

The first form of b_i (i=1,3) are

$$b_1 = \frac{15}{2}Q_2 x_1 x_3^2 r^{-4} - \frac{3}{2}Q_2 x_1 r^{-5} + \frac{35}{2}Q_3 x_1 x_3^3 r^{-9} - \frac{15}{2}Q_3 x_1 x_3 r^{-7} + \frac{315}{8}Q_4 x_1 x_3^4 r^{-11} - \frac{105}{4}Q_4 x_1 x_3^2 r^{-9} + \frac{15}{8}Q_4 x_1 r^{-7} - \gamma\rho v v_1,$$

$$b_2 = \frac{15}{2}Q_2 x_2 x_3^2 r^{-4} - \frac{3}{2}Q_2 x_2 r^{-5} + \frac{35}{2}Q_3 x_2 x_3^3 r^{-9} - \frac{15}{2}Q_3 x_2 x_3 r^{-7} + \frac{315}{8}Q_4 x_2 x_3^4 r^{-11} - \frac{105}{4}Q_4 x_2 x_3^2 r^{-9} + \frac{15}{8}Q_4 x_2 r^{-7} - \gamma\rho v v_2,$$

$$b_3 = -\frac{9}{2}Q_2 x_3 r^{-5} + \frac{15}{2}Q_2 x_3^3 r^{-7} - 15\,Q_3 x_3^2 r^{-7} + \frac{35}{2}Q_3 x_3^4 r^{-9} + \frac{3}{2}Q_3 r^{-5} - \frac{175}{4}Q_4 x_3^3 r^{-9} + \frac{315}{8}Q_4 x_3^5 r^{-11} + \frac{75}{8}Q_4 x_3 r^{-7} - \gamma\rho v v_3.$$

The second form of b_i (i=1,3) are

$$b_1 = \frac{15}{2}Q_2 x_1 x_3^2 r^{-4} - \frac{3}{2}Q_2 x_1 r^{-5} + \frac{35}{2}Q_3 x_1 x_3^3 r^{-9} - \frac{15}{2}Q_3 x_1 x_3 r^{-7} + \frac{315}{8}Q_4 x_1 x_3^4 r^{-11} - \frac{105}{4}Q_4 x_1 x_3^2 r^{-9} + \frac{15}{8}Q_4 x_1 r^{-7} - \gamma\rho v v_1 - \beta\mu x_1 r^{-3},$$

$$b_2 = \frac{15}{2}Q_2 x_2 x_3^2 r^{-4} - \frac{3}{2}Q_2 x_2 r^{-5} + \frac{35}{2}Q_3 x_2 x_3^3 r^{-9} - \frac{15}{2}Q_3 x_2 x_3 r^{-7} + \frac{315}{8}Q_4 x_2 x_3^4 r^{-11} - \frac{105}{4}Q_4 x_2 x_3^2 r^{-9} + \frac{15}{8}Q_4 x_2 r^{-7} - \gamma\rho v v_2 - \beta\mu x_2 r^{-3},$$

$$b_3 = -\frac{9}{2}Q_2 x_3 r^{-5} + \frac{15}{2}Q_2 x_3^3 r^{-7} - 15\,Q_3 x_3^2 r^{-7} + \frac{35}{2}Q_3 x_3^4 r^{-9} + \frac{3}{2}Q_3 r^{-5} - \frac{175}{4}Q_4 x_3^3 r^{-9} + \frac{315}{8}Q_4 x_3^5 r^{-11} + \frac{75}{8}Q_4 x_3 r^{-7} - \gamma\rho v v_3 - \beta\mu x_3 r^{-3};$$

and

$$\gamma = \frac{1}{2} C_D \frac{A}{M}.$$

IV. SOLUTION TECHNIQUE

In this section, the solution technique of the formulations of section 3 will be applied by two steps. The first step is to transform Eqs.(3.1) to (3.7) into first order differential equations by the following substitutions

$$y_i = u_i, \qquad y_{i+4} = u_i', \quad \text{i = 1(1)4,}$$
$$y_9 = \alpha_k, \qquad y_{10} = t, \qquad y_{11} = r$$

and $y_{12} = r'$.

Then the first order system of the problem becomes

$$y_1' = y_5, \tag{4.1}$$
$$y_2' = y_6, \tag{4.2}$$
$$y_3' = y_7, \tag{4.3}$$
$$y_4' = y_8, \tag{4.4}$$
$$y_5' = -y_9\, y_1 + \tfrac{1}{2} y_{11} b_1, \tag{4.5}$$
$$y_6' = -y_9\, y_2 + \tfrac{1}{2} y_{11} b_2, \tag{4.6}$$
$$y_7' = -y_9\, y_3 + \tfrac{1}{2} y_{11} b_3, \tag{4.7}$$
$$y_8' = -y_9\, y_4 + \tfrac{1}{2} y_{11} b_4, \tag{4.8}$$
$$y_9' = -y_5\, b_1 - y_6\, b_2 - y_7\, b_3 - y_8\, b_4, \tag{4.9}$$
$$y_{10}' = y_{11}, \tag{4.10}$$
$$y_{11}' = y_{12}, \tag{4.11}$$
$$y_{12}' = \mu + y_{11}\left(y_1 b_1 + y_2 b_2 + y_3 b_3 + y_4 b_4 - 4 y_9\right). \tag{4.12}$$

Also, the accuracy checks were need in the solution could be obtained. The accuracy of the computed values of the *y's* variables at any fictitious time *s* (corresponding to the time *t*) could be checked by the bilinear relation (BI)

$$BI = y_4\, y_1' - y_3\, y_2' + y_2\, y_3' - y_1\, y_4',$$

and it must be equal to zero in excellent accuracy. The second step is solving the above system by using the fourth-order Runge-Kutta method with a fixed step size in the next section.

V. RESULTS AND CONCLUSION

We'll take as the numerical example the Explorer 19 at 750 Km height ([2]). So, the initial position and velocity components are

$$\bar{x}_0 = (3538.646, -2902.799, -5483.478)\ Km,$$
$$\dot{\bar{x}}_0 = (5.842408, -1.772259, 4.707377)\ Km/sec,$$

at epoch 14 February 1976, where one orbital revolution is elapsed in 111 min., it has the ratio *A/m = 13.04E-07 Km/Kg*.

Since the adopted physical constant are

$R = 6378.135\ Km$, $\mu = 398600.8\ Km^3/sec^2$,

and the coefficients of the four order zonal harmonic are

$J_2 = 1.0826157 \times 10^{-3}$,

$J_3 = -2.53648 \times 10^{-6}$,

$J_4 = -1.6233000 \times 10^{-6}$,

where $C_D = 2.2$ ([27]), also we'll chose τ equals 4, and finally β equals 0.5 .

We'll use all the above values to compute the position and velocity components, i.e., the six elements; especially (the elements a, e, i) because of these elements are much affected by our studied forces. Also, we'll get the accuracy check (bilinear relation, *BI*) at any time (days); and we get the following figures and supplemented tables. The figures show the variations of the classical orbital elements with the time over one hundred, one thousand and two thousand revolutions (as an example). All the Figures show the effects of the Earth's gravitational field with axial symmetry up to the four order zonal harmonic, air drag and solar radiation force. Also, all the Figures show a significant difference in *a,i*; but in *e* show the slightly difference, that is because the height of satellite about 750 Km. All Tables give the bilinear relation (*BI*) under the studied forces at any time (days), which indicates a good prediction for the numerical solution. The numerical results are just only as an example, since this method could be applied to any orbit. To get more accurate prediction of the motion of the artificial satellite we will be taken into account the whole other forces affecting on the motion.

TABLE I. THE VALUES OF BILINEAR RELATION CORRESPOND TO THEIR PERTURBATION FORCES, OVER ONE HUNDRED REVOLUTIONS.

Time (Days)	The bilinear relation (BI)		
	Only gravity	*With pert. and without SRP*	*With pert. and with SRP*
0.0	9.094947018E-13	9.094947018E-13	9.094947018E-13
0.768729642	-2.119122655E-10	-2.437445801E-10	-2.037268132E-10
1.537459283	-4.147295840E-10	-4.174580681E-10	-4.110916052E-10
2.306188925	-6.075424608E-10	-6.511982065E-10	-6.184563972E-10
3.074918567	-7.621565601E-10	-9.304130799E-10	-8.494680515E-10
3.843648208	-9.813447832E-10	-1.155967766E-09	-1.048647391E-09
4.61237785	-1.190528565E-09	-1.263288141E-09	-1.218722900E-09
5.381107492	-1.396074367E-09	-1.469743438E-09	-1.429725671E-09
6.149837133	-1.651642378E-09	-1.701664587E-09	-1.690750651E-09
6.918566775	-1.876287570E-09	-1.972694008E-09	-1.965418051E-09
7.687296417	-2.097294782E-09	-2.193701221E-09	-2.183696779E-09

TABLE II. THE VALUES OF BILINEAR RELATION CORRESPOND TO THEIR PERTURBATION FORCES, OVER ONE THOUSAND REVOLUTIONS.

Time (Days)	The bilinear relation (BI)		
	Only gravity	*With pert. and without SRP*	*With pert. and with SRP*
69.10879479	-1.300668373E-08	-1.282751327E-08	-1.281568984E-08
69.87752443	-1.291800800E-08	-1.270018402E-08	-1.277567208E-08
70.64625407	-1.278158379E-08	-1.255057214E-08	-1.275566319E-08
71.41498371	-1.259513738E-08	-1.254329618E-08	-1.268517735E-08
72.18371335	-1.251191861E-08	-1.239686753E-08	-1.255466486E-08
72.95244299	-1.232592695E-08	-1.229727786E-08	-1.244961823E-08
73.72117264	-1.223497748E-08	-1.209718903E-08	-1.233775038E-08
74.48990228	-1.212993084E-08	-1.196940502E-08	-1.219359547E-08
75.25863192	-1.194939614E-08	-1.186663212E-08	-1.204853106E-08
76.02736156	-1.177386366E-08	-1.179296305E-08	-1.197349775E-08
76.7960912	-1.170656105E-08	-1.166154107E-08	-1.182388587E-08

TABLE III. THE VALUES OF BILINEAR RELATION CORRESPOND TO THEIR PERTURBATION FORCES, OVER TWO THOUSAND REVOLUTIONS.

Time (Days)	The bilinear relation (BI)		
	Only gravity	*With pert. and without SRP*	*With pert. and with SRP*
145.981759	3.085688149E-09	3.104332791E-09	2.578872227E-09
146.7504886	3.039986041E-09	3.069544618E-09	2.542719812E-09
147.5192182	2.985871106E-09	3.016566552E-09	2.493834472E-09
148.2879479	2.934029908E-09	2.949946065E-09	2.428350854E-09
149.0566775	2.875594873E-09	2.886963557E-09	2.359001883E-09
149.8254072	2.803062671E-09	2.816250344E-09	2.294882506E-09
150.5941368	2.724959813E-09	2.752358341E-09	2.222009243E-09
151.3628664	2.638671504E-09	2.673573363E-09	2.130605026E-09
152.1315961	2.550564204E-09	2.580577529E-09	2.056481208E-09
152.9003257	2.444039637E-09	2.470642357E-09	1.969851837E-09
153.6690554	2.341835170E-09	2.371393748E-09	1.866283128E-09

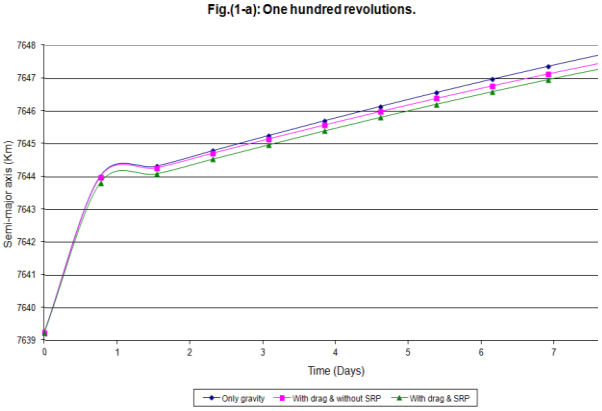


Fig. 1.a Semi-major axis of One hundred revolutions

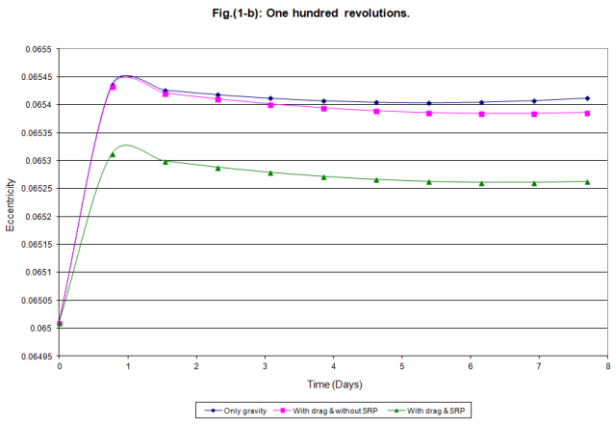


Fig. 1.b Eccentricity of One hundred revolutions

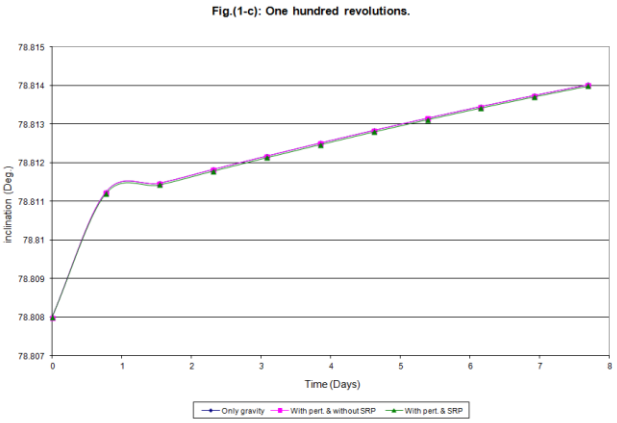


Fig. 1.c Inclination of One hundred revolutions

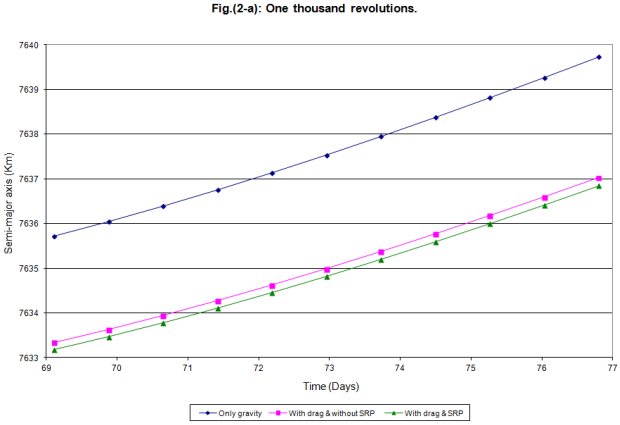


Fig. 2.a Semi-major axis of One thousand revolutions

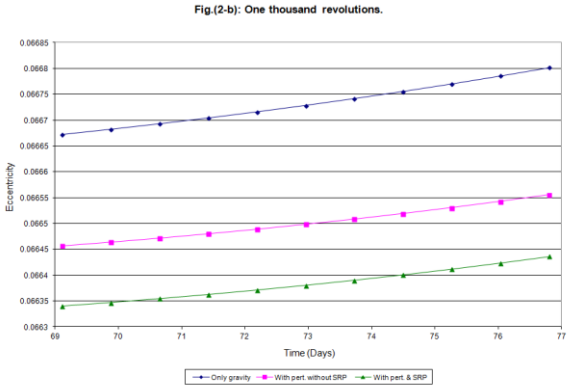


Fig. 2.b Eccentricity of One thousand revolutions

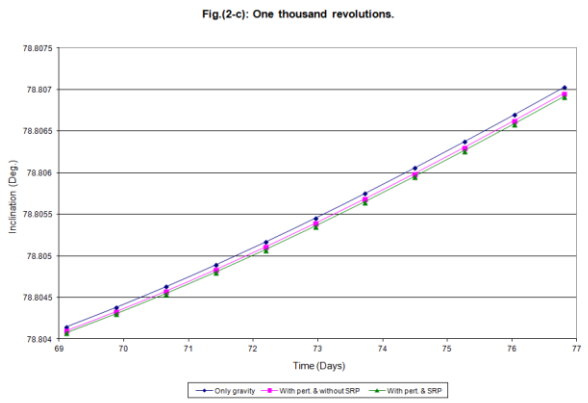


Fig. 2.c Inclination of One thousand revolutions

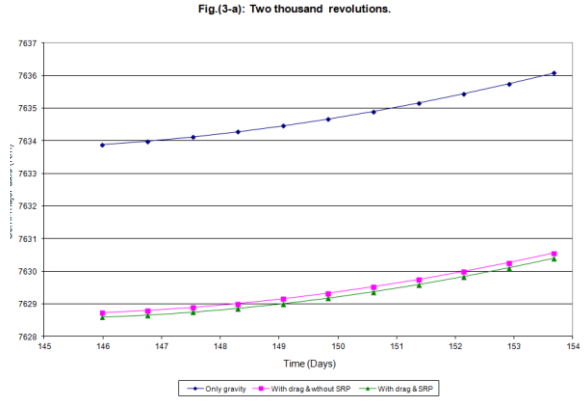


Fig. 3.a Semi-major axis of Two thousand revolutions

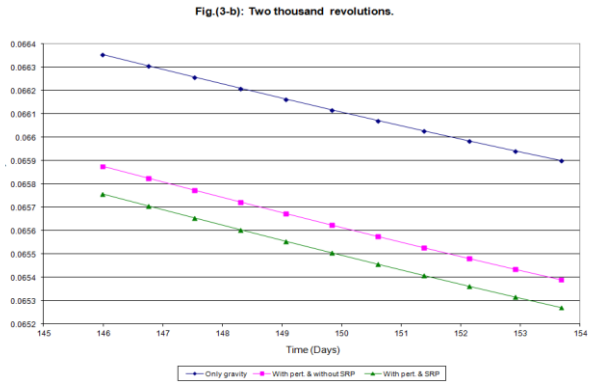


Fig. 3.b Eccentricity of Two thousand revolutions

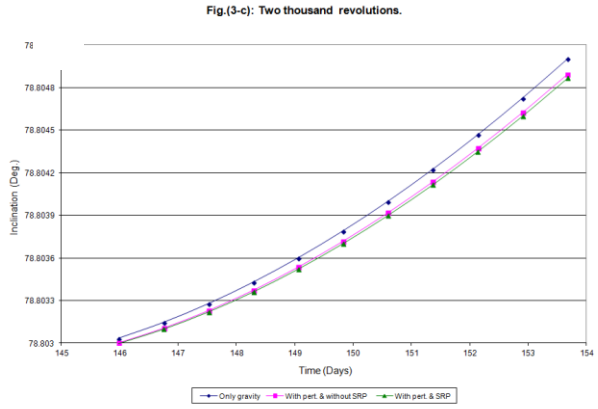


Fig. 3.c Inclination of Two thousand revolutions

REFERENCES

[1] R. F. Arenstorf, "New regularization of the restricted problem of three bodies," Astron. J. , vol. 68, p. 548, Oct 1963.

[2] C. J. Brookes and F. C. E. Rylan, "Perturbation of the orbit of Explorer 19 due to solar radiation," Celestial Mechanics, vol. 27, pp. 339-352, Aug 1982.

[3] R. Broucke, "Solution of the N-Body Problem with Recurrent Power Series," Celestial Mechanics, vol. 4, pp. 110-115, Sep 1971.

[4] C. J. Cohen and R. H. Lyddane, "Radius of convergence of Lie series for some elliptic elements," Celestial Mechanics, vol. 25, pp. 221-234, 1981.

[5] G. E. Cook, "Luni-Solar Perturbations of the Orbit of an Earth Satellite," Geophysical Journal International, vol. 6, pp. 271-291, Apr 1962.

[6] F. Delhaise, "Analytical treatment of air drag and earth oblateness effects upon an artificial satellite," Celestial Mechanics and Dynamical Astronomy, vol. 52, pp. 85-103, Mar 1991.

[7] H. Dankowicz, "Some special orbits in the two-body problem with radiation pressure," Celestial Mechanics and Dynamical Astronomy, vol. 58, pp. 353-370, Apr 1994.

[8] R. V. De Moraes, "Solar radiation pressure and balloon type artificial satellites," in Satellite Dynamics, 1975.

[9] G. Hori, "The Effect of Radiation Pressure On The Motion Of An Artificial Satellite," in Space Mathematics Part III, vol. 5, J. B. Rosser, Ed., Amer. Math. Soc., 1966, p. 167.

[10] B. Kampos, "General Perturbation theory," 1968.

[11] D. King-Hele, Satellite orbits in an atmosphere. Theory and applications., D. King-Hele, Ed., Glasgow, UK.: Blackie and Son Ltd., 1987.

[12] Y. Kozai, "Effects Motion of an Artificial Satellite," SAO Special Report, vol. 56, p. 25, Jan 1961.

[13] E. Leimanis and N. Minorsky, Dynamics and nonlinear mechanics. Some recent advances in the dynamics of rigid bodies and celestial mechanics, New York, USA: Wiley, 1958.

[14] T. Levi-Civita, "Sur la résolution qualitative du problème restreint des trois corps," in Opere Matematiche, vol. 2, Bologna, 1956.

[15] R. H. Lyddane, "Small eccentricities or inclinations in the Brouwer theory of the artificial satellite," Astron. J., vol. 68, p. 555, Oct 1963.

[16] A. G. Mavraganis and D. G. Michalakis, "The two-body problem with drag and radiation pressure," Celestial Mechanics and Dynamical Astronomy, vol. 58, pp. 393-403, Apr 1994.

[17] A. Milani, A. M. Nobili and P. Farinella, Non-gravitational perturbations and satellite geodesy., Bristol,UK.: Adam Hilger Ltd., 1987.

[18] P. Moore, "Perturbations of a spheroidal satellite due to direct solar radiation pressure," Celestial Mechanics, vol. 20, pp. 125-142,Aug 1979.

[19] P. Musen, "The Influence of the Solar Radiation Pressure on the Motion of an Artificial Satellite," J. Geophys. Res., vol. 65, p. 1391, Sep 1960.

[20] D. O'Mathuna, "Satellite Prediction Formulae for Vinti's Model," Celestial Mechanics, vol. 1, pp. 467-478, Sep 1970.

[21] R. Ramnath, "Gravitational Perturbations of Equatorial Orbits," Celestial Mechanics, vol. 8, pp. 85-98, Aug 1973.

[22] M. Roy, Dynamics of satellites, I. U. of Theoretical and A. Mechanics, Eds., Berlin: Springer Heidelberg, 1963.

[23] G. Scheifele, "On Nonclassical Canonical Systems," Celestial Mechanics, vol. 2, pp. 296-310, Sep 1970.

[24] L. Sehnal, "The influence of the re-radiation of the Earth on the motion of the artificial satellites," in The Theory of Orbits in the Solar System and in Stellar Systems, 1966.

[25] L. Sehnal, "Radiation pressure effects in the motion of artificial satellites.," in Dynamics of Satellites, 1970.

[26] L. Sehnal, "Non-gravitational forces in satellite dynamics," in Satellite Dynamics, 1975.

[27] M. A. Sharaf and M. E. Awad, "Prediction of trajectories in Earth's gravitational field with axial symmetry.," Proc. Math. Phys. Soc. Egypt, vol. 60, 1985.

[28] T. E. Sterne, "The gravitational orbit of a satellite of an oblate planet," Astron. J., vol. 63, p. 28, Jan 1958.

[29] E. L. Stiefel and G. Scheifele, Linear and regular celestial mechanics; perturbed two-body motion, numerical methods, canonical theory, Berlin: Springer-Verlag, 1971.

[30] V. Szebehely and G. E. O. Giacaglia, "On the elliptic restricted problem of three bodies," Astron. J., vol. 69, p. 230, Apr 1964.

[31] V. Szebehely, "Selection of regularizing functions," Astron. J., vol. 72, p. 370, Apr 1967.

[32] B. D. Tapley and V. Szebehely, "The regularization of Optimal Trajectories," in JPL Space Programs Summary, vol. 4, Pasadena, Calif., Jet Propulsion Laboratory, 1967, pp. 37-46.

[33] E. T. Whittaker, A Treatise on the Analytical Dynamics, of Particles and Rigid Bodies, Fourth Edition ed., E. T. Whittaker, Ed., London and New York.: Cambridge University Press, 1961.

OJADEAC: An Ontology Based Access Control Model for JADE Platform

Ban Sharief Mustafa
Computer Sciences Department, Mosul University
Mosul , Iraq

Najla Aldabagh
Computer Sciences Department, Mosul University
Mosul , Iraq

***Abstract*—Java Agent Development Framework (JADE) is a software framework to make easy the development of Multi-Agent applications in compliance with the Foundation for Intelligent Physical Agents (FIPA) specifications. JADE propose new infrastructure solutions to support the development of useful and convenient distributed applications. Security is one of the most important issues in implementing and deploying such applications. JADE-S security add-ons are one of the most popular security solutions in JADE platform. It provides several security services including authentication, authorization, signature and encryption services. Authorization service will give authorities to perform an action based on a set of permission objects attached to every authenticated user. This service has several drawbacks when implemented in a scalable distributed context aware applications. In this paper, an ontology-based access control model called (OJADEAC) is proposed to be applied in JADE platform by combining Semantic Web technologies with context-aware policy mechanism to overcome the shortcoming of this service. The access control model is represented by a semantic ontology, and a set of two level semantic rules representing platform and application specific policy rules. OJADEAC model is distributed, intelligent, dynamic, context-aware and use reasoning engine to infer access decisions based on ontology knowledge.**

Keywords*—Java Agent Development Framework (JADE); JADE-S; Ontology-Based Access Control Model; Web Ontology Language (OWL)

I. INTRODUCTION

Currently Multi-Agent system provides a platform to build open distributed systems including e-commerce, web-services and pervasive computing environments. Security is an important issue in most of these applications and must be quarantined or they will face significant deployment problems. JADE is a popular Multi-Agent platform used in many commercial, academic and scientific agent-based Projects.

There are a number of extensions to JADE that provide a security platform to the system in particular S-Agent and the JADE-S plug-in. JADE-S is a security Add-ons component providing secure platform with authentication, authorization, encryption and public key infrastructure [1]. The authorization in JADE-S depends on Java Security model and Java access controller. JADE-S extends this controller to act with JADE platform architecture and permissions [2]. It structures the platform as a multi-user environment where every agent or container will be owned by an authenticated user, who is authorized to perform several privileged critical actions. However these permissions are an extension to Java Permission objects to support JADE platform actions. Its access control is depends on Identity Based Access control, where permissions are given based on the identity of the user. Thus it cannot support policy, attributes, and context aware access decisions. Also, a fine grained access rule cannot be adopted in such model.

By introducing Semantic Web and Semantic Web technologies from Berners-Lee [3], a new intelligent and semantic vision is introduced for building security services. The new Semantic Web technologies shows great promising in building semantic based security services especially in building an access control model.

In this paper, OJADEAC model, an ontology based access control based on Semantic Web technologies is proposed. OJADEAC relies on a proposed JMASO ontology that models the JADE Multi-Agent system knowledge with any information needed to support access decisions. OJADEAC is a policy model where an access is taken according to access control policy rules. Policy rules are specified at two levels: platform and application. Complete authorization architecture is provided by building a kernel service that automatically enforces access control policies on request to JADE commands. OJADEAC model shows great advantages over JADE-S authorization model. However it also suffers from some drawbacks including model performance and JADE shortcomings.

The paper is structured as follows: section 2 presents the related works. Section 3 introduces JADE and JADE-S platforms. Section 4 gives a brief introduction to OWL. Section 5 deals with model and model implementation, introducing JADE multi-agent ontology and the model architecture.

II. RELATED WORKS

Several projects build a secure agent platform based on JADE-S secure platform extending and substituting its authorization service. In other side, using Semantic Web technologies in access control mechanisms has taken considerable attention from different researchers who build an ontology based access control for different domains. Related works are reviewed in these two dimensions. The first dimension on extending JADE-S authorization service, Vila and Schuster [4] built an Intelligent Learning Management System called EUME, based on JADE-S. Permission service is extended to define a fine-grained access to specific

functionality provided by special agents by implementing a two authorization layers. Vitabile [5] proposed a new access control model which merges the advantages of the three classical models: Role Based Access Control (RBAC) model, Mandatory Access Control (MAC) model and Credential Based Access Control (CBAC). The new model replaces JADE-S authorization service.

In using Semantic Web technologies to build access control model, Giunchiglia [6] presented Relation Based Access Control (RelBAC) which is a model and logic to deal with the problem of access control in Web 2.0 applications representing access control rules and policies as DL formulas and reasoning about them using reasoner. Masoumzadeh [7] proposes social network system ontology, ontology based access control model and authorization policy rules to address the protection of semantic rich information in the knowledge base ontology for social network.

Shen [8] proposes semantic context-based access control model to be applied in a mobile Web services environment by combining Semantic Web technologies with context-based access control mechanism. His work focuses on the context aware access decision due to the characteristic of the open mobile Web services pervasive applications.

III. JADE AND JADE-S

JADE is a platform that provides basic middleware–layer functionalities which are independent of the specific application and which simplify the realization of distributed applications that exploit the software agent abstraction [9]. A JADE platform is composed of agent containers that can be distributed over the network. A special container called main container represent the bootstrap point of a platform [10]. JADE based on distributed coordinated filters architecture. According to this architecture every agent based operation will be forwarded to the service responsible for implementing it [10].

JADE supports defining a new kernel service to implements new agent operations. The kernel Service composed form several components including: outgoing and incoming sinks responsible for implementing the commands belonging to this service, service slice which represents a service proxy in other nodes. At last a two chain filters work as incoming and outgoing filters. Service can use outgoing filter to intercept any command issued from another service and react in a service specific way. Service use incoming filter to intercept any vertical commands issued by service slices in that node and react in a service specific way [10]. Figure 1 shows the filter architecture for message kernel service.

JADE-S is formed from standard JADE with JADE Security plug-in. JADE-S provide some security features to JADE platform. It extending Java security model provides the advantages of Java Authentication and Authorization Service (JAAS), Java Cryptography Extension (JCE) and Java Secure Socket Extension (JSSE). It allows exchanging critical information through a network using a secure data transmission (SSL) [3] .

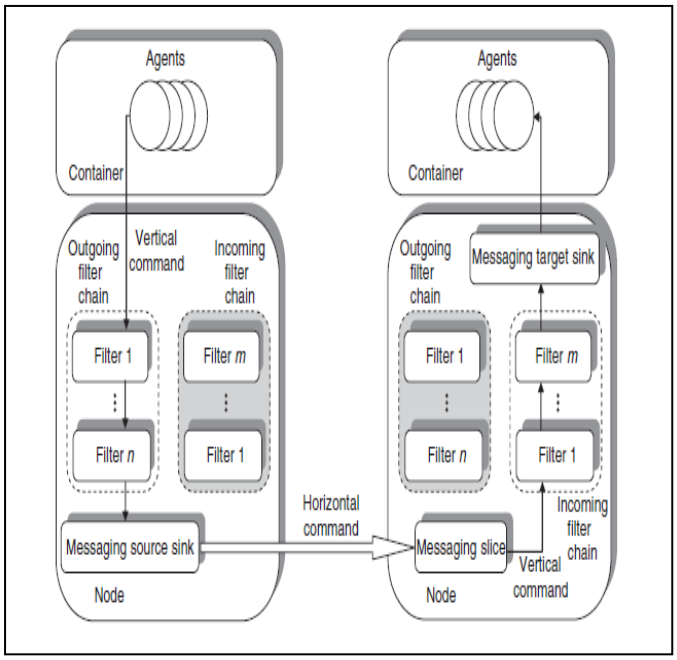


Fig. 1. Filter architecture for message kernel service [8]

JADE-S structures the agent platform as a multi-user environment where every agents and containers will belong to authenticated users. A permissions file contain a set of actions that each user is authorized to perform will represent the security policy of the platform [2]. JADE-S provides four JADE kernel services: ***Security Service*** which represents the base service to run JADE-S, and carried the authentication process. ***Signature and Encryption Services*** which provides Java cryptography message signing and encryption methods to assure message integrity, non-repudiation and confidentiality [1]. ***Permission Service*** based on an access control list saved in a policy file that follows the Java/JAAS syntax to make a decision on all actions that agents can perform in the platform [4]. The main policy file resides in main container, while other containers will have their own local policy file. Right given to an authenticated user will be translated to all its owned agents [4].

Access rights will enforced during the platform execution in a point that based on the Permission kernel service filters. The check command method performs the authorization check on received command. The main drawback for this service is that it is based on Identity Based Access Control (IBAC) model. Where every access decision is taken based on the permissions given to authenticated username. Such model is not suitable to be implemented in an open and distributed applications, thus many research projects that based on JADE-S platform tend to replace this service with another one to overcome its deficiency as mentioned in related work.

IV. WEB ONTOLOGY LANGUAGE

OWL is a Semantic Web language proposed by W3C [11], and is probably the most popular language for creating ontologies today.

OWL defines classes, properties, and their hierarchies. OWL gives a more expressivity in expressing a complex and richer relationships with greatly enhanced reasoning ability [12]. OWL comes in three different versions: Lite, DL and FULL. The most suitable one in building ontologies and reasoning over it is DL, because it is designed to support existing description logics, and has properties that are desirable for reasoning systems [13].

V. Model Implementation

An ontology-based access control model called OJADEAC is proposed based on Semantic Web technologies. OJADEAC is a fine grained semantic aware model that protects resource access in JADE platform and service access in domain specific application. The model relies on an ontology called (JMASO) that model the JADE Multi-Agent platform knowledge by storing keys entities and their relationships typically found in JADE and any information related to access control purpose including inferences based on access control policy rules.

A. JADE Multi-Agent System Ontology

JADE ontology called JMASO is proposed that modelsmain concepts with its relationships found in JADE platform. JAMSO can be extended to cover other aspects in this domain. The current version of JMASO contains 6 concepts, 10 object properties and 8 data properties. Protégé editor has been used to create JMASO ontology. Figure 2 show the complete ontology graph. Node concept represents any container in JADE platform with two descendents: MainContainer and Container.

The Principle concept represent any entity that can be responsible about their actions, it specialize into two concepts: Agent and Owner. Owner represent all users begin a new node. Agent is the class of all agents created in platform. Node is linked to Owner with OwnedBy object properties. Agent is linked to Node with hasLocation object properties. NamedAction is an important class that models all action in JADE that can be requested and checked against access control policy rules. Service concept models all services registered in DF agent. NamedAction concept has three descendents subclasses: AgentAction, ContainerAction and ServiceAction. NamedAction relates to Agent with hasSubject property and relates to Agent or Service with hasObject property.

PermissionAction is specialized to two disjoint classes ProhibiedAction and PermittedAction. Access check is taken according to the inferred type of NamedAction object which must be either PermittedAction or ProhibiedAction.

B. OJADEAC model architecture

OJADEAC authorization architecture is shown in Figure 3. The diagram reflects the following logical actors involved in OJADEAC model:

Knowledge Base: is composed form JMASO ontology and a set of two levels of policy rules (system and application specific). In OJADEAC model, Jena rules are used to express a policy rules. For example, the following platform policy rule says that if agent created in container that is owned by action requester, then it is permitted action.

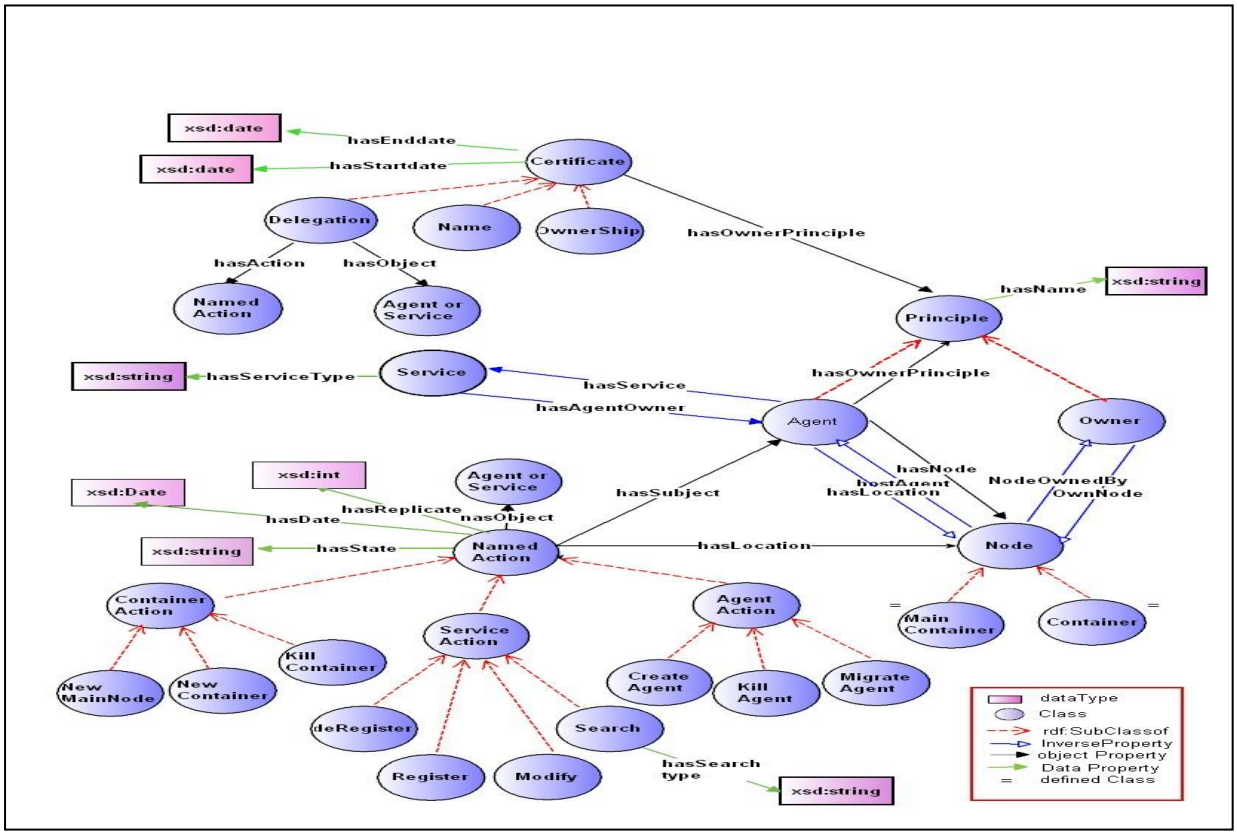


Fig. 2. JAMSO ontology graph

[permittedCreate: //name of rule
(?x rdf:type ac:CreateAgent)

(?x ac:hasSubject ?y)

(?x ac:hasLocation ?z)

(?z ac:NodeOwnedBy ?y)

->

(?x rdf:type ac:permittedAction)]

Ontology manager: It is responsible for gathering and updating A-Box knowledge parts. New individuals and its related properties with other individuals and data values are asserted to knowledge base during runtime. Ontology manager is distributed over nodes gathering its knowledge during runtime. Main container will contain knowledge about the whole platform by sharing and exchanging it with other node's knowledge. RDQL is used in knowledge sharing and exchanging between different ontology knowledge scattered on different nodes in platform. When node knowledge need an information from other node, it send an ACL message with RDQL query as message content to this node or to main container and gets response as sub model to be added to its ontology. For example, the following RDQL query asks for triples that have subject (?x) which own Container-1:

" SELECT ?x WHERE {?x ont:OwnNode ont:Container-1}"

ont: is the prefix name space of JMASO ontology. The query answer is a set of triples with subject zaid binding to ?x variable:

[http://Onto.owl#zaid, http://Onto.owl#hasJob, "Faculty"]

[http://Onto.owl#zaid,http://Onto.owl#hasRole, ContainerAdmin"]

[http://Onto.owl#zaid, http://Onto.owl#hasDegree, "null"]

[http://Onto.owl#zaid,http://Onto.owl#hasEmail, "sss@yahoo.com"]

[http://Onto.owl#zaid, rdf#type, http://Onto.owl#Owner]

[http://Onto.owl#zaid, rdf#type, owl#Thing]

[http://Onto.owl#zaid, rdf#type, http://Onto.owl#Principle]

[http://Onto.owl#zaid, rdf#type, http://Onto.owl#Agent]

Policy Enforcement Point (PEP): OntSecure service is a new kernel service that is added to JADE platform to support OJADEAC model implementation. The PEP will be placed in OntSecure incoming and outgoing filters. PEP is different form one command to other depending on the sequence steps executed to implement this command. Figure 4 shows sequence steps including OJADEAC PEP and ontology assertions when REQUEST-CREATE command is issued.

Policy Decision Point (PDP): Decision is made by inference over JAMSO ontology and policy rules to infer action type. The action will belong either to PermittedAction or ProhibiedAction. A method to add a new action to ontology will activate the inference process, and then action type will be checked. If action is prohibited, then a security exception is raised and command is blocked, else command will continue its implementation.

Policy Administrator Point (PAP): platform administrator and application vendors are responsible for creating policy and policy rules set.

VI. DISCUSSION

OJADEAC model is an ontology based access control for JADE platform. It substitutes its authorization service providing several advantages over it. OJADEAC features over JADE-S authorization service can be summarized in following points:

1) Using ontology provides reasoning ability for access control decision. Also access control information can be accessed, queried and discovered automatically including owner attributes which can be deduced from knowledge by reasoning.

2) The proposed model has a higher degree of interoperability compared with other approaches to access control. This is because of the nature of ontologies in providing semantic interoperability.

3) Fine-grained access control policies are expressed using a set of Jana rules. Policy rules are formulated either as a system level policy rules or application level policy rules.

4) In addition to supporting ABAC decisions and RBAC decisions, OJADEAC model can support context aware decisions because policy rules can take into account any resource, object and environment conditions. These constraints can be evaluated by reasoning over OMASJ ontology model. Date, time and location in addition to others can be included in decision making.

5) OJADEAC model can store past agent/user behaviors. So it can adapt a trust and reputation model and detect any abnormal behavior to act as intrusion detection.

VII. CONCLUSIONS

As OJADEAC model has several features, it also suffers from some draw backs that need to be addressed and solved to apply this model in real applications implementation. The first drawback is its performance due to time spent in reasoning and in managing ontology. Another problem is JADE shortcomings that affect the model behavior. Some commands are implemented in a way that cannot be intercepted in service filters (example, DF service commands). Other problem is the missing parameters attached to some commands (example, KILL-CONTAINER requestor parameter is always NULL) which effect making a decision. At last, the model is a good start to overcome the shortcoming of JADE-S permission service and to reduce the gap with Semantic Web applications.

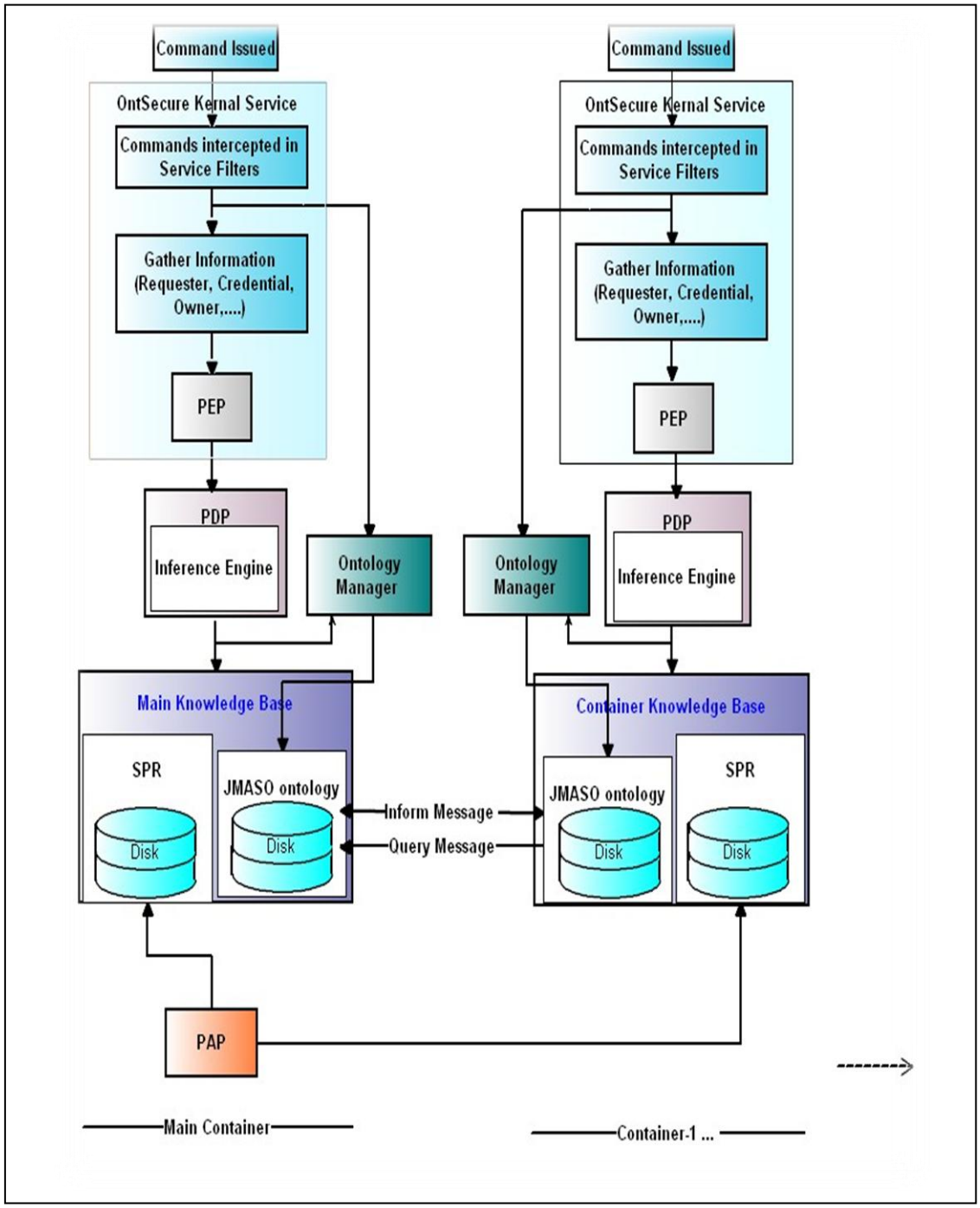


Fig. 3. OJADEAC model architecture

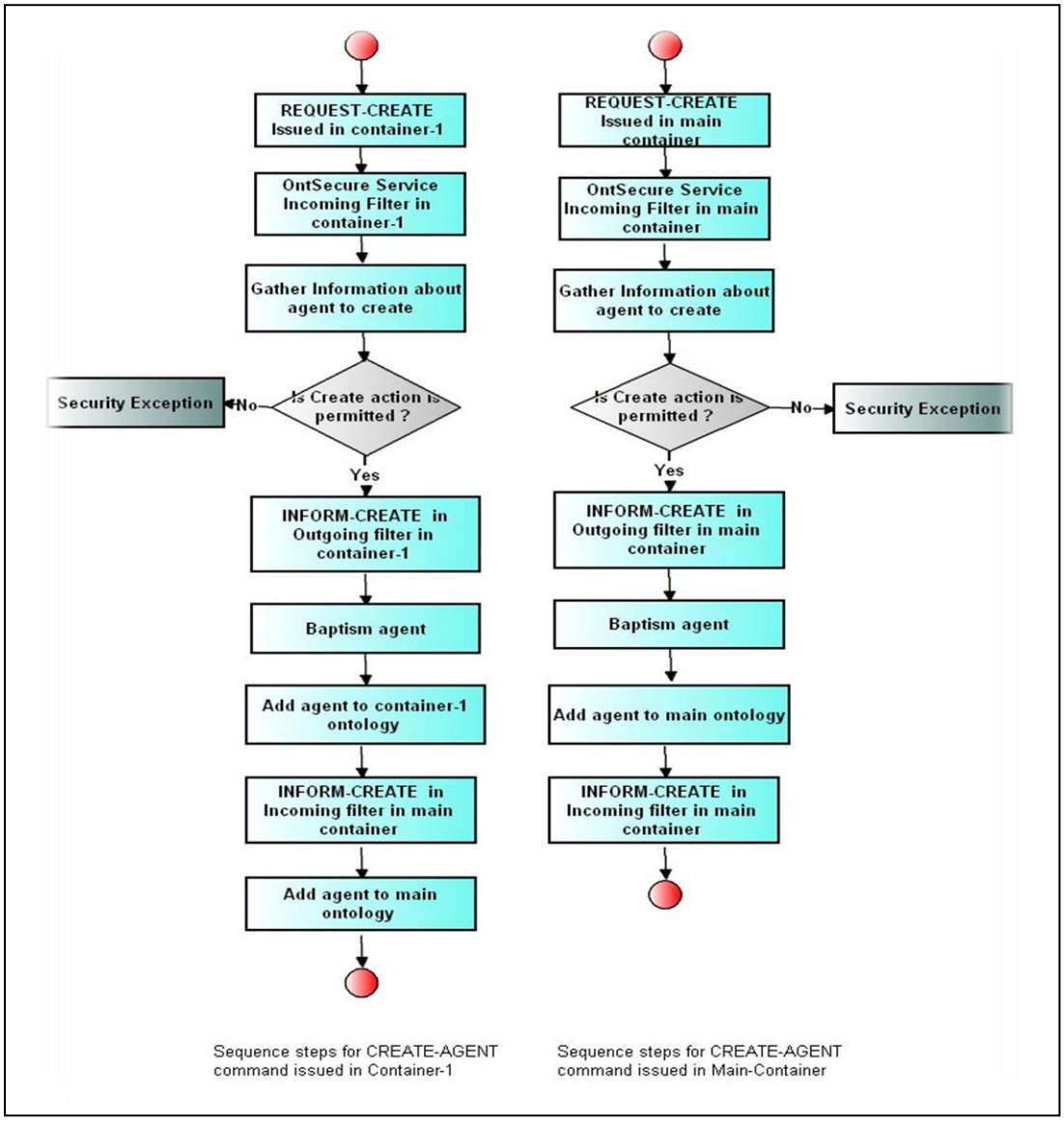


Fig. 4. Sequence steps for CRAETE-AGENT command

REFERENCES

[1] G. Vitaglione, "JADE Tutorial, Security Administrator Guide" 1, 19 Sep. 2002.

[2] A. Moreno, D. Sanchez, and D. Isern, "Security measures in a Medical Multi-Agent System", in Proceeding Artificial intelligence and Applications, 2005.

[3] JADE Board, "JADE Security Guide", Technical Report, TILab, Telecom Italia, 28 February 2008.

[4] X. Vila, A. Schuster, and A. Riera, "Security for a Multi-Agent System based on JADE", Computer and Security, Vol. 26, 2007, pp.391-400.

[5] S. Vitabile, V. Conti, C. Militello and F. Sorbello, (September 2009), "An extended JADE-S based framework for developing secure Multi-Agent Systems," Computer Standards & Interfaces, Vol. 31, No. 5, September 2009, pp. 913-930.

[6] F. Giunchiglia and R. Zhang, "Ontology Driven Community Access Control", Technical Report, University of Trento, Dec. 2008.

[7] A. Masoumzadeh, and J. Joshi, "Ontology based Access Control for Social Network Systems", Information Privacy, Security and Integrity, Vol. 1, No. 1,2011 .

[8] H. Shen and Y. Cheng, "A Semantic Context-Based Model for Mobile Web Services Access Control", Computer Network and Information Security, Vol. 1, 2011, pp. 18-25.

[9] A. LUPAŞC, "A Multi-Agent Platform for Developments of Accounting Intelligent Applications", Economics and Applied Informatics, No. 1, 2008, pp. 79-86..

[10] F. Bellifemine, ,G. Caire and D. Greenwood, "Developing multi-agent systems with JADE", John Wily & Sons, Ltd, 2007.

[11] W3C OWL Working Group, "OWL 2 Web Ontology Language: Document Overview". W3C Recommendation, Internet: http://www.w3.org/TR/owl2-overview/.

[12] N. Aldabagh, and B. Mustafa, "A Comparative Study between Using OWL Technology and Jess Rule Based For Applying Knowledge to Agent Based System", International Journal of Computer Science and Information Security, Vol. 10, No. 7. July 2012.

[13] I. Horrocks, F. Peter and P. Schneider, "KR and Reasoning on the Semantic Web: OWL", In Handbook of Semantic Web Technologies, Ed. Domingue, J., Fensel, D., and Hendler, A., Chapter 9, 2011, pp. 365-398. Springer.

Proposal for Two Enhanced NTRU

Ahmed Tariq Sadiq
Computer Science Department
University of Technology
Baghdad, Iraq

Najlaa Mohammad Hussein
Computer Science Department
Baghdad University
Baghdad, Iraq

Suha Abdul Raheem Khoja
Electronic and communication Engineering Department
Baghdad University
Baghdad, Iraq

***Abstract*—Sound is very widely used in communication. In order to ensure secure communication a cryptographic data scheme is used. Secure sound is needed in many fields such as military, business, banking and electronic commerce. There is also an increasing demand for secured sound in network communication. Several symmetric and asymmetric algorithms are used for sound encryption. In this work, NTRU, the last in line public key cryptosystem is enhanced in two methods and used for encrypting sound files after converting the sound into text. In the proposed methods the message is encrypted one character at a time, since NTRU encrypts only prime numbers, thus 7 bits of each character is encrypted and the eighth bit is left without encryption. In method I NTRU algorithm is enhanced by adding the result obtained from calculating a mathematical equation of one variable to the message and then the resulted encrypted bit is fed-back and added to the next bit of the message in the next step; this procedure is repeated for the subsequent bits of the message. In method II NTRU algorithm is enhanced by adding the subsequent states of LFSR (Linear Feedback Shift Register) to the subsequent bytes of the message. The proposed methods are tested on several sound files; the results show that the proposed methods I and II maintain approximately the same original method encryption and decryption time while generating more complex encryption.**

***Keywords*—NTRU; security; sound**

I. Introduction

NTRU (Number theory Research unit) algorithm is a public key cryptosystem invented by three professors of mathematics from brown university of America Jeffrey Hoffstin, Jill Pipher and Joseph H. Silverman in 1996. [1] NTRU is built on polynomial algebra. The basic objects are truncated polynomials in the ring $R=Z[X] / (X^N-1)$ and the basic tool is the reduction of polynomials with respect to two relatively prime modulo. The security of the system is (hoped to be) based on the difficulty of finding a "short" factorization for such polynomials. This latter problem is equivalent to finding a short vector in a certain 2N dimensional lattice, a commonly known and also widely studied hard problem. [2] Since NTRU is a ring based public key cryptosystem and is therefore quite different from the group based cryptosystems whose security relies on the integer factorization problem or the discrete logarithm problem. This extra structure can be exploited to obtain a very fast cryptosystem; to encrypt/decrypt a message block of length N, NTRU only requires $O(N^2)$ time, whereas the group based schemes like RSA etc. requires $O(N^3)$ time. Furthermore, NTRU also has a very short key size of O(N) and very low memory requirements, which makes it ideal for constrained devises such as smart cards. [3]

The rest of this paper is organized as follows: related work is given in section II, section III provides a brief description of NTRU algorithm, the proposed methods are described in section IV, section V presents the experimental results and finally conclusions and future work are given in section VI.

II. Related Work

Jaspreet Kaur and Er. Kanwal preet Singh [4] use three different kinds of algorithms NTRU, RSA and RINGDAEL for speech encryption and decryption by first converting the speech into text then further the text is converted into cipher text. The performances are analyzed of these three approaches respectively the parameters calculated are encryption, decryption, delay time, complexity, packet lost and security levels. In these three approaches, encryption decryption and delay time are varied according to the number of bits per second.

On the other hand, complexity and packet lost are approximately the same. There is no packet lost during transmitting and receiving the data. Also, Jaspreet Kaur and Er. Kanwal preet Singh [5] use three different kind of techniques i.e. MD-5, SHA-2 and RINGDAEL for speech encryption, where the speech is first converted into text then the text is converted into cipher text. At the end, the performances of these three approaches are analyzed, respectively.

III. Brief Description of Ntru Algorithm

A. Prameters

NTRU has three integer parameters N, p and q. N represents the degree of the polynomials at most N-1, p and q are used to reduce the coefficients of the polynomials, p is smaller than q and they have no common divisor. [6, 7, 8]

B. Key generation

Sending a secret message from Bob to Alice requires the generation of a public and private key. The public key is known by both Alice and Bob and the private key is only known by Alice. To generate the key pair two polynomials f and g with coefficients much smaller than q, with degree at most N-1 and with coefficients in {-1, 0, 1} are required.

The polynomial f must satisfy the requirement that the inverses modulo q and modulo p exist, which means that $f*fp = 1 \pmod p$ and $f*fq = 1 \pmod q$ must hold. So when the chosen f is not invertible Alice has to go back and try another f. Both f and fp is Alice private key. The public key h is generated by computing $h = fq * g \pmod q$. [9]

C. Encryption

When Bob wants to send a secret message to Alice, he puts his message in the form of a polynomial m with coefficients between $-1/2p$ and $1/2p$ .Next Bob randomly chooses another small polynomial r. This is the blinding value which is used to obscure the message. Bob uses the message m; randomly chosen polynomial r and Alice's public key h to compute the polynomial e = r * h + m (mod q). The polynomial e is the encrypted message which Bob sends to Alice. [10, 11]

D. Decryption

In addition to the publically available information Alice knows her own private key, on receiving Bob's cipher text, Alice start the decryption process by computing the polynomial a = f * e (mod q). [12] She then shift the coefficient of polynomial a to the range $\left(-q/2, q/2\right)$[13] and does a mod p computation to obtain: d = fp * a (mod p). Assuming that the parameters have been chosen properly then the polynomial d must be equal to Bob plain text m. [12]

IV. Proposed Methods

A. methodI

In this method Alice and Bob agree on a mathematical equation of one variable say (x), the value of this variable is send via one of the key establishment protocols. [14]

Bob start the encryption process by calculating the result of the mathematical equation and assigning it to a variable say (v), the value of this variable is added to the message (m), then for each bit of the encrypted message (e) the value of the previous e is assigned to the variable v, this means for each bit the value of the encrypted message is fed-back and added to the new value of m. Adding the mathematical equation to the message makes the encryption process more complex especially if the degree of this equation is high. The pseudo code of encryption process in the original NTRU algorithm [15, 13] is enhanced in this method and listed in pseudo code (1) as follows:

Pseudo code 1 Encode (N, q, r, m, h, e, x)

```
Require: N, q, Public Key h, message m, and random polynomial r.
1:  v = calculate the result of mathematical equation of one
        variable
2:  Star Multiply (r, h, e, N, q)
3:  for i = 0 to N - 1 do
4:    e[i] = e[i] + m[i] + v mod q
5:    v = e[i]
6:  end for
7:  {Encode returns the encrypted message, e, through the argument list.}
```

When Alice starts the decryption process, she also calculates the result of the mathematical equation and assigns it to a variable, but instead of adding the result of the mathematical equation to the message she subtracted it form the encrypted message, then for each bit the value of the encrypted message is fed-back and subtracted from the new value of the message. The pseudo code of the decryption process in the original NTRU algorithm is enhanced in this method and listed in pseudo code (2) as follows:

Pseudo code 2 Decode (N, q, p, f, fp, e, d, x)

```
Require: N, q, p, secret key f, inverse polynomial fp, and encrypted message e.
1:  v = calculate the result of mathematical equation of one
        variable
2:  for i = 0 to N - 1 do
3:    vv = e[i]
4:    e[i] = e[i] – v mod q
5:    v= vv
6:  end
7:  Star Multiply (f, e, a, N, q)
8:  for i = 0 to N - 1 do
9:    if a[i] < 0 then
10:     a[i] = a[i] + q {Make all coefficients positive}
11:   end if
12:   if a[i] > q/2 then
13:     a[i] = a[i] - q {Shift coefficients of a into range (-q/2,
          q/2)}
14:   end if
15: end for
16: Star Multiply (a, fp, d, N, p)
17: {Decode returns the decrypted message, d, through the argument list.}
```

B. methodII

In this method pseudo random bits are generated with a LFSR [16], Alice and Bob agree on the initial state of the LFSR. The generation of LFSR is shown in pseudo code (3)

Pseudo code 3 LFSR

```
1:  for i = N down to 2 do
2:    lfsr (i) = lfsr (i-1)
3:  end for
4:  lfsr (1) = xor (lfsr (3), lfsr (5))
```

To encrypt the message Bob adds the initial state of the LFSR to the first byte of the message. The subsequence states of the LFSR are then added to the subsequent bytes of the message. The changes that are made to the pseudo code of encryption process in the original NTRU algorithm are shown in pseudo code (4) as follows:

Pseudo code 4 Encode (N, q, r, m, h, e, lfsr)

```
Require: N, q, Public Key h, message m, and random polynomial r.
1:  Star Multiply (r, h, e, N, q)
2:  for i = 0 to N - 1 do
3:    e[i] = e[i] + m[i] + lfsr[i] mod q
4:  end for
5:    {Encode returns the encrypted message, e, through the argument list.}
```

To decrypt the message Alice repeats the same steps that are followed by Bob but instead of adding the states of the LFSR to the message, she subtracted the states from the subsequent bytes of the message. The changes that are made to the pseudo code of the decryption process in the original NTRU are shown in pseudo code (5) as follows:

Pseudo Code 5 Decode (N, q, p, f, fp, e, d, lfsr)

```
Require: N, q, p, secret key f, inverse polynomial fp, and encrypted message e
1:  for i = 0 to N - 1 do
2:    e[i] = e[i] – lfsr[i] mod q
3:  end
4:  Star Multiply (f, e, a, N, q)
5:  for i = 0 to N - 1 do
```

```
6:      if a[i] < 0 then
7:        a[i] = a[i] + q {Make all coefficients positive}
8:      end if
9:      if a[i] > q/2 then
10:       a[i] = a[i] - q {Shift coefficients of a into range (-
          q/2, q/2)}
11:     end if
12:   end for
13:   Star Multiply (a, fp, d, N, p)
14:   {Decode returns the decrypted message, d, through the argument list.}
```

V. Experimental Results

In this work the sound is first converted into text then further the text is converted into cipher text. This method can be applied to any kind of sound files after storing the file in a text editor such as note pad. This method is applied to the original NTRU algorithm (namely original method) and to the two proposed methods. The sound is converted into text via ISO-8859-1: 8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1, is part of the ISO/IEC 8859 series of ASCII-based standard character encodings. It is generally intended for "Western European" languages [17].

In the original method and proposed methods I and II the message is partitioned into characters, each character is encrypted separately. Since N must be a prime number, 7 bit of the character is encrypted and the eighth bit is left without encryption. The method is tested for different values of N, the results show that the maximum value for N in this method is 47; it is also shown that if the value of N is increased, encryption and decryption time will also increase. The original method and the proposed methods are tested on 25 wave sound files of sizes ranging from 10 KB to 1MB. The encryption and decryption time in seconds is computed for each one of the 25 files 25 times and then average of computation is taken to increase the accuracy of calculation.

Fig. 1, Fig. 2 and Fig. 3 respectively, displays the effect of file size on the time of encryption and decryption of the original method proposed method I and proposed method II respectively. Fig. 4 displays a comparison for the effect of file size on the time of encryption of the original method, proposed method I and proposed method II. Fig. 5 displays a comparison for the effect of file size on time of decryption of the original method proposed method I and proposed method II.

VI. Conclusions and Future Work

Proposed method I enhanced the original NTRU algorithm by adding the results obtained from calculating a mathematical equation of one variable to the message. This led to a more complex encryption while maintaining approximately the same original algorithm encryption and decryption time. Proposed method II enhanced the original NTRU algorithm by adding the states of the LFSR to the bytes of the message. This maintains approximately the same original method encryption and decryption time while generating more complex cipher. The time needed for encryption and decryption in proposed method I and II is approximately the same in spite of the different values that are added to the message in each method. In future research, Apply the proposed methods on Field Programmable Gate Array (FPGA) which has higher speed when compared to the standard processors.

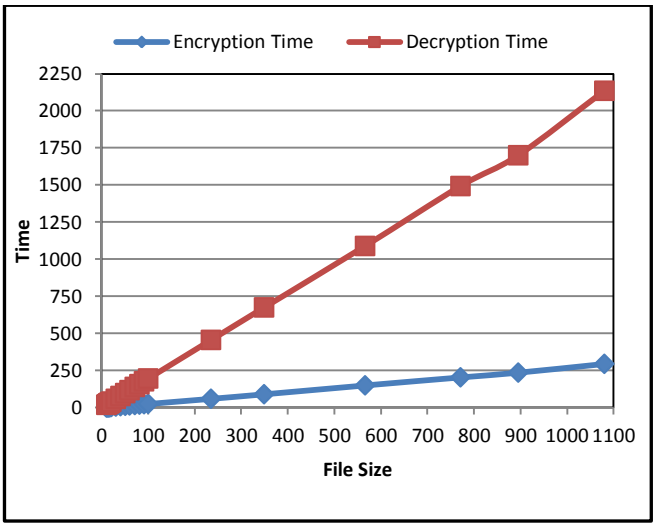


Fig. 1. The Effect Of File Size On The Time Of Encryption And Decryption Of The Original Method

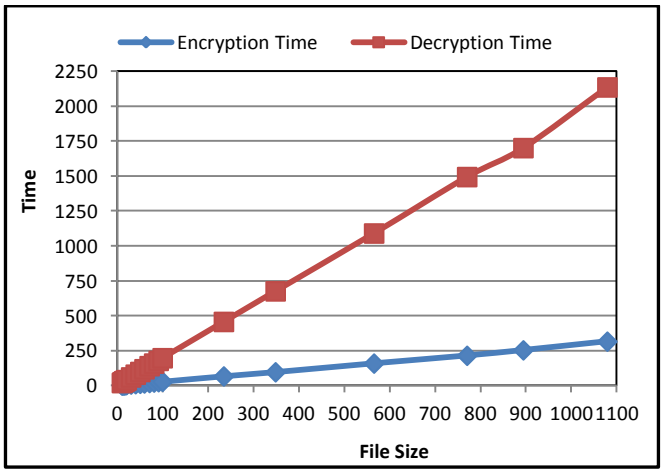


Fig. 2. The Effect Of File Size On The Time Of Encryption And Decryption Of Proposed Method I

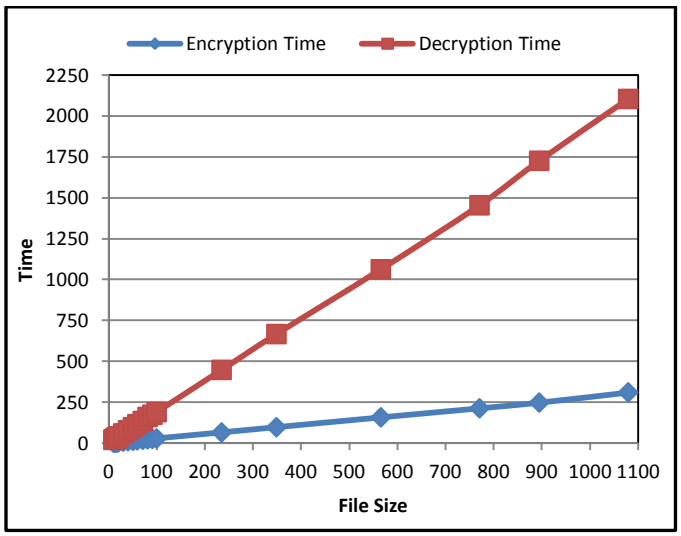


Fig. 3. The Effect Of File Size On The Time Of Encryption And Decryption Of Proposed Method Ii

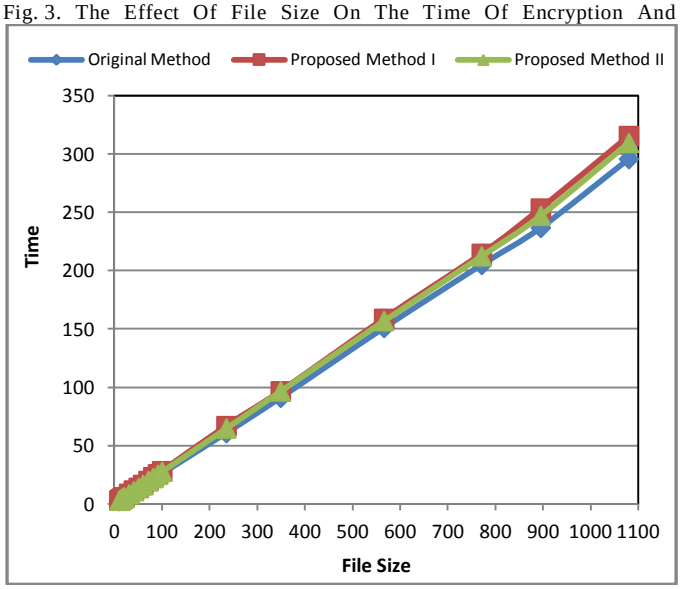


Fig. 4. Compression For The Effect Of File Size On Time Of Encryption Of The Original Method And The Two Proposed Methods

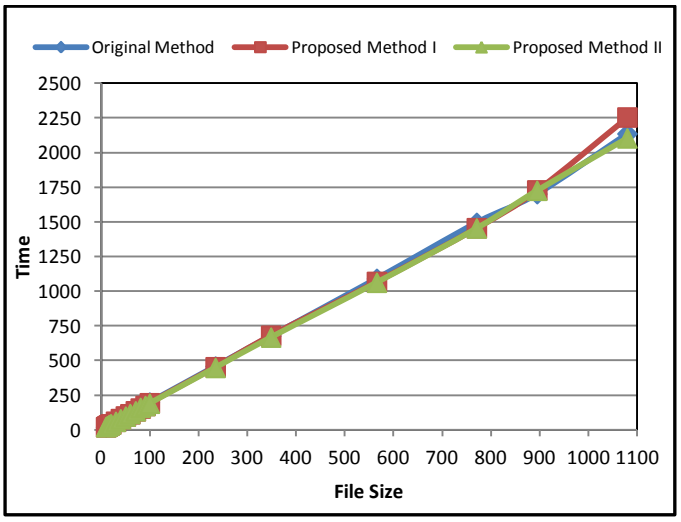


Fig. 5. Compression For The Effect Of File Size On The Time Of Decryption Of The Original Method And The Two Proposed Methods

REFERENCES

[1] Yue, B. S., Yan, Z. H., Rruchuan, W., "A algorithm of dynamic patterns for NTRU", IEEE, 2008.

[2] Nayak, R., Pradhan, J., Sastry, C. V."A matrix formulation for NTRU cryptosystem", IEEE 2008.

[3] Jha, R., Saimi, Anil, K., "A comparative analysis and enhancement of NTRU algorithm for network security and performance improvement" International Conference on Communication Systems and Network Technologies, 2011.

[4] Kaur, J., Singh, K. p. "Comparative study of speech encryption algorithms using mobile applications", International Journal of Computer Trends and Technology, Vol.4, No. 7, Jully 2013, pp. 2346 - 2350.

[5] Kaur, J., Singh, K. p. "Speech to text encryption using cryptography techniques", International Journal of Innovative Research and Development, Volume 2, N0. 4, pp. 274 - 283, April 2013.

[6] Shen, X., Du, Z., Chen, R., "Research on NTRU algorithm for mobile java security" IEEE 2009, pp.366 - 369.

[7] Ramajaneyulu, S., Nayak, R., "Secure mobile system using NTRU encrypt", International Journal of Computer Trends and Technology, Volume 4, Issue 2, 2013.

[8] Manasa, C., Masheswar, M.V.S.N. "Secure mobile IM system using NTRU", Intrernation Jornal of Engineering Research and Technology, Volume 1, No.9, November 2012.

[9] Kumar, R. G. V. S., Jumar, N. K., Sekhar, C. P., Numma, B. V. V. S., Kumar, V. B., "Modified mutual authentication and key agreement protocol based on NTRU cryptography for wireless communications", International Journal of Computer Science and Network (IJCSN), Volume 1, Issue 4, August, 2012.

[10] Reddy, A. N., Nayak, R., Baboo, S., "Analysis and performance characteristics of cryptosystem using image files", International Journal of Computer Applications , pp.0975 – 8887, Volume 51 – No. 22, August 2012.

[11] Narasimham, C., Pradhan, A., "Evaluation of performance characteristics of cryptosystem using text files" Journal of Theoretical and Applied Information Technology, JATIT, 2008.

[12] Jeffrey, H., Pipher, J., Silverman, J. H., "An introduction to mathematical cryptography", Springer, New York, 2008.

[13] O'Rourke, C. M., "Efficient NTRU implementations" Thesis, April, 2002.

[14] Paar, C., Pelzl, J., "Understanding cryptography: a textbook for student and pracittionners", Springer-Verlag Berlin Heidelberg, 2010.

[15] Yadav, S. K., Bhardwaj, K.,"On NTRU implementation: an algorithmic approach",Proceedings of the 4th National Conference; INDIA, 2010.

[16] Beker, H., Piper, F., "Cipher systems the protection of communications", Northwood Books: London, 1982.

[17] ISO /IEC JTC 1/SC 2/WG 3 7bit and 8bit codes and their extension SECRETARIAT: ELOT, 1998.

Building BTO System in the Sanitary Materials Manufacturer with the Utilization of the High Accuracy Forecasting

Hirotake Yamashita
[1]College of Business Administration and Information Science, Chubu University, 1200 Matsumoto-cho Kasugai, Aichi,487-8501, Japan

Kazuhiro Takeyasu
[2]College of Business Administration, Tokoha University, 325 Oobuchi, Fuji City, Shizuoka, 417-0801, Japan

***Abstract*—In recent years, BTO (Build to Order) system is prevailing. It pursues short lead time, minimum stocks, and thereby minimum cost. But the high accuracy demand forecasting is inevitable for the parts manufacturers. In this paper, well organized BTO system in the sanitary materials manufacturer is seek with the aid of high accuracy demand forecasting, which is newly developed by us. Focusing that the equation of ESM is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in ESM was derived. Trend removal method was also devised. AR model is also used for forecasting. After removing trend, AR model is utilized and forecasting is executed. Better one in the forecasting accuracy between them was chosen for the final forecasting.**

Thus, we could obtain the high accuracy demand forecasting. These methods are examined by the data of sanitary materials manufacturer and the BTO system is newly built by utilizing this method. Further development of this system should be performed hereafter.

Keywords*—BTO; forecasting; lead time; stock; sanitary materials; AR model

I. INTRODUCTION

Supply Chain Management has developed in various aspects. In recent years, BTO (Build to Order) is becoming a big trend. BTO is a system that manufacturers produce goods after they take the order from the customers. The manufacturers can handle the order with the stockless production system. The success story of BTO is found in Dell, where the direct sales to customers have been built. It has the advantage that it has low risk of holding stocks and it can make flexible manufacturing and it can also cope with the customers detailed requests. It is called "Mass Customization" because it has realized the low cost production with mass production and enabled to cope with the customers' requests at the same time. Stockless production with minimum lead time bears a good cost benefit but it has a lot of problems and issues. Even if manufacturer takes a BTO system, parts suppliers must have certain stocks in order to meet the request of manufacturer who gather parts and make assembling. Minimum stocks including parts manufacturer should be pursued. Otherwise, the system would not work long.

Reviewing past papers, there are many researches made on this. G. Parry and A. Graves (2008)[1] and D. M. Anderson (2004)[2] made a versatile overview on BTO. G. Parry and A. Graves (2008) especially focused on the possibility of BTO for automobile manufacturing. Fukushima et al. (2006)[3], Fukushima et al. (2007)[4] made a research on Dell China at Amoy and detailed production scheme including suppliers was analyzed.

In this paper, we are trying to build at the sanitary materials manufacturer, which has a particular circumstance of suppliers and users. Therefore it would be a rare and precious research. One of the most biggest hazard point is how to make correct forecasting. Stock level including parts manufacturers depends much on this.

We proposed a new method of estimation of smoothing constant in ESM before[5]. Focusing that the equation of ESM is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in ESM was derived. Trend removal method was also devised. In making forecast such as production data, trend removing method is devised. Trend removing by the combination of linear and 2nd order non-linear function and 3rd order non-linear function is executed to the data of sanitary materials manufacturer. The weights for these functions are set 0.5 for two patterns at first and then varied by 0.01 increment for three patterns and optimal weights are searched. For the comparison, monthly trend is removed after that. Theoretical solution of smoothing constant of ESM is calculated for both of the monthly trend removing data and the non-monthly trend removing data. Then forecasting is executed on these data. AR model is also used for forecasting. After removing trend, AR model is utilized and forecasting is executed. Better one in the forecasting accuracy between them was chosen for the final forecasting.

In this paper, utilizing above stated method, BTO system for the sanitary materials manufacturer is newly built.

The rest of the paper is organized as follows. In section 2, current status and issues of sanitary materials manufacturer are stated. Section 3 through 6 are the description of forecasting method. In section 3, ESM is stated by ARMA model and estimation method of smoothing constant is derived using

ARMA model identification. The combination of linear and non-linear function is introduced for trend removing in section 4. The Monthly Ratio is referred in section 5. AR model is described in section 6. Forecasting is executed in section 7, and estimation accuracy is examined. In section 8, building BTO system is stated, which is followed by the conclusion of section 9.

II. Current Status and Issues of Sanitary Materials Manufacturer

A. Outline of Sanitary Materials Manufacturer

Outline of Sanitary Materials Manufacturer (Company O) is as follows.

/ Established: 1936

/ Business Domain: Production and sales of Sanitary Materials

/ Business Department and Product:

A. Medical Department

absorbent gauze, absorbent cotton, products for maternity clinic, products for gynecological department, products for surgery

B. Consumer Department

hygiene cotton, products for nursing and nursing care

/ Base:

A. Headquarter

Nagoya

B. Sales Base

10 places in Japan

Shanghai

C. Factory

3 factories in Japan

Indonesia (Affiliated)

/ Number of Product Items: 3943 (Real Moving one: about 1600)

/ Lead Time for the Production:

A. Sterilized Product: 10~12 days

B. Non Sterilized Product: 5~7 days

/ ABC Analysis in the Sales Amount (September 2011 through August 2012)

Total items: 3943

Rank A items (Upper 70%) 577

Rank B items (Upper 90%) 727

Rank C items 2639

B. Problems and Issues

Under the severe competition with other companies, Company O has took every means in order to reduce cost. The methods they took are as follows.

/ Increase Sales Base so as to respond to the customers' needs

/ Increase distribution center

/ Increase purchase from abroad

/ Build factory in abroad

Partially, these made contribution of cost reduction, but caused disadvantage of the increase of stocks and prolonged lead time.

Imported parts from abroad consist of nearly 20 % and these take more than one month's lead time, which causes increase of stocks. Each department made every effort in each department and that did not make any optimization in the total.

There are several means for the total optimization. Among them, correct forecasting is inevitable in the total supply chain management.

From now on, the newly developed forecasting method by us is stated in section 3 through 6 and forecasting is executed in section 7.

This time, we set Rank A items (Share is 15% in the total items) as important items to be controlled and make forecasting for them. The base data for forecasting is each sales amount of September 2009 through August 2012.

III. Description of ESM using ARMA model

A. ESM and ARMA mode

In ESM, forecasting at time $t+1$ is stated in the following equation.

$$\hat{x}_{t+1} = \hat{x}_t + \alpha\left(x_t - \hat{x}_t\right) \quad (1)$$

$$= \alpha x_t + (1-\alpha)\hat{x}_t \quad (2)$$

Here,

$\hat{x}_{t+1}$: forecasting at $t+1$

x_t : realized value at t

α : smoothing constant $(0 < \alpha < 1)$

(2) is re-stated as

$$\hat{x}_{t+1} = \sum_{l=0}^{\infty} \alpha(1-\alpha)^l x_{t-l} \quad (3)$$

By the way, we consider the following (1,1) order ARMA model.

$$x_t - x_{t-1} = e_t - \beta e_{t-1} \quad (4)$$

Generally, (p,q) order ARMA model is stated as

$$x_t + \sum_{i=1}^{p} a_i x_{t-i} = e_t + \sum_{j=1}^{q} b_j e_{t-j} \quad (5)$$

Here,

$\{x_t\}$: Sample process of Stationary Ergodic Gaussian Process $x(t)\ t = 1,2,\cdots,N,\cdots$

$\{e_t\}$: Gaussian White Noise with 0 mean σ_e^2 variance

MA process in (5) is supposed to satisfy convertibility condition.

Utilizing the relation that

$$E\left[e_t \middle| e_{t-1}, e_{t-2}, \cdots\right] = 0$$

we get the following equation from (4).

$$\hat{x} = x_{t-1} - \beta e_{t-1} \quad (6)$$

Operating this scheme on t+1, we finally get

$$\begin{aligned} \hat{x}_{t+1} &= \hat{x}_t + (1-\beta)e_t \\ &= \hat{x}_t + (1-\beta)(x_t - \hat{x}_t) \end{aligned} \quad (7)$$

If we set $1-\beta = \alpha$, the above equation is the same with (1), i.e., equation of ESM is equivalent to (1,1) order ARMA model, or is said to be (0,1,1) order ARIMA model because 1st order AR parameter is -1 .

B. Estimation of Smoothing Constant utilizing System Identification of ARMA model

Comparing with (4) and (5), we obtain.

$$\begin{cases} a_1 = -1 \\ b_1 = -\beta = \alpha - 1 \end{cases}$$

From (1), (7),

$$\alpha = 1 - \beta .$$

Therefore, we get

$$\begin{cases} a_1 = -1 \\ b_1 = -\beta = \alpha - 1 \end{cases} \quad (8)$$

From above, we can get estimation of smoothing constant after we identify the parameter of MA part of ARMA model. But, generally MA part of ARMA model become non-linear equations which are described below.

Let (5) be

$$\tilde{x}_t = x_t + \sum_{i=1}^{p} a_i x_{t-i} \quad (9)$$

$$\tilde{x}_t = e_t + \sum_{j=1}^{q} b_j e_{t-j} \quad (10)$$

We express the autocorrelation function of $\tilde{x}_t$ as $\tilde{r}_k$ and from (9), (10), we get the following non-linear equations which are well known.

$$\tilde{r}_k = \begin{cases} \sigma_e^2 \sum_{j=0}^{q-k} b_j b_{k+j} & (k \le q) \\ 0 & (k \ge q+1) \end{cases} \quad (11)$$

$$\tilde{r}_0 = \sigma_e^2 \sum_{j=0}^{q} b_j^2$$

For these equations, recursive algorithm has been developed. In this paper, parameter to be estimated is only b_1, so it can be solved in the following way.

From (4) (5) (8) (11), we get

$$\left.\begin{aligned} & q = 1 \\ & a_1 = -1 \\ & b_1 = -\beta = \alpha - 1 \\ & \tilde{r}_0 = (1 + b_1^2)\sigma_e^2 \\ & \tilde{r}_1 = b_1 \sigma_e^2 \end{aligned}\right\} \quad (12)$$

If we set

$$\rho_k = \frac{\tilde{r}_k}{\tilde{r}_0}$$

the following equation is derived.

$$\rho_1 = \frac{b_1}{1 + b_1^2} \quad (13)$$

We can get b_1 as follows.

$$b_1 = \frac{1 \pm \sqrt{1 - 4\rho_1^2}}{2\rho_1} \quad (14)$$

In order to have real roots, ρ_1 must satisfy

$$|\rho_1| \le \frac{1}{2} \quad (15)$$

From invertibility condition, b_1 must satisfy

$$|b_1| < 1 \quad (16)$$

From (14), using the next relation,

$$(1-b_1)^2 \geq 0$$
$$(1+b_1)^2 \geq 0$$

(16) always holds.

As

$$\alpha = b_1 + 1$$

b_1 is within the range of

$$-1 < b_1 < 0$$

Finally we get

$$\left.\begin{aligned} b_1 &= \frac{1-\sqrt{1-4\rho_1^2}}{2\rho_1} \\ \alpha &= \frac{1+2\rho_1-\sqrt{1-4\rho_1^2}}{2\rho_1} \end{aligned}\right\} \quad (17)$$

which satisfy above condition. Thus we can get a theoretical solution by a simple way.

Focusing that the equation of ESM is equivalent to (1,1) order ARMA model equation, we can estimate smoothing constant after estimating ARMA model parameter.

It can be estimated only by calculating 0th and 1st order autocorrelation function

IV. Trend Removal Method

As a trend removal method, we describe the combination of linear and non-linear function.

[1] Linear function

We set:

$$y = a_1 x + b_1 \quad (18)$$

as a linear function.

[2] Non-linear function

We set:

$$y = a_2 x^2 + b_2 x + c_2 \quad (19)$$

$$y = a_3 x^3 + b_3 x^2 + c_3 x + d_3 \quad (20)$$

as a 2nd and a 3rd order non-linear function.

[3] The combination of linear and non-linear function

We set:

$$y = \alpha_1(a_1 x + b_1) + \alpha_2(a_2 x^2 + b_2 x + c_2) \quad (21)$$

$$y = \beta_1(a_1 x + b_1) + \beta_2(a_3 x^3 + b_3 x^2 + c_3 x + d_3) \quad (22)$$

$$\begin{aligned} y = {} & \gamma_1(a_1 x + b_1) + \gamma_2(a_2 x^2 + b_2 x + c_2) \\ & + \gamma_3(a_3 x^3 + b_3 x^2 + c_3 x + d_3) \end{aligned} \quad (23)$$

as a combination of linear and 2nd order non-linear and 3rd order non-linear function. Here, , , . Trend is removed by dividing the original data by (21),(22) and (23). Comparative discussion concerning (21), (22) and (23) is described in section 5.

V. Monthly Ratio

For example, if there is the monthly data of L years as stated bellow:

$$\{x_{ij}\}\ (i = 1, \cdots, L)\ (j = 1, \cdots, 12)$$

where, $x_{ij} \in R$ in which j means month and i means year and x_{ij} is a shipping data of i-th year, j-th month, then, monthly ratio $\tilde{x}_j\ (j = 1, \cdots, 12)$ is calculated as follows.

$$\tilde{x}_j = \frac{\frac{1}{L}\sum_{i=1}^{L} x_{ij}}{\frac{1}{L}\cdot\frac{1}{12}\sum_{i=1}^{L}\sum_{j=1}^{12} x_{ij}} \quad (24)$$

Monthly trend is removed by dividing the data by (24). Numerical examples both of the monthly trend removal case and the non-removal case are discussed in section 7.

VI. AR Model by Using Delay Operator

p - th order AR model is stated as follows.

$$x_t + a_1 x_{t-1} + \cdots + a_p x_{t-p} = e_t \quad (21)$$

VII. Forecasting the Shipping Data of Manufacturer

A. Analysis Procedure

The shipping data of sanitary materials manufacture for 4 cases from September 2009 through August 2012 are analyzed. Contents of these are as follows.

Case 1: Sum data of all Rank A items

Case 2: Product A (absorbent cotton)

Case 3: Product B (Set product for birth)

Case 4: Product C (Material of set product for birth)

Analysis procedure is as follows. There are 36 monthly data for each case. We use 24 data (1st to 24th) and remove trend by the method stated in section 4. Then we calculate monthly ratio by the method stated in section 5. After removing monthly trend, the method stated in section 3 is applied and Exponential Smoothing Constant with minimum variance of forecasting error is estimated. Then 1 step forecast is executed. Thus, data is shifted to 2nd to 25th and the forecast for 26th data is executed consecutively, which finally reaches forecast of 36th data. To examine the accuracy of forecasting, variance of forecasting error is calculated for the data of 25th to 36th data. Final forecasting data is obtained by multiplying monthly ratio and trend.

Forecasting error is expressed as:

$$\varepsilon_i = \hat{x}_i - x_i \tag{25}$$

$$\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \tag{26}$$

Variance of forecasting error is calculated by:

$$\sigma_\varepsilon^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\varepsilon_i - \bar{\varepsilon}\right)^2 \tag{27}$$

These schemes are similarly adapted to AR model stated in section 6. We set Forecasting Accuracy Ratio (FAR) as follows.

$$FAR = \left(1 - \sum_{i=1}^{N}\frac{\left|\hat{x}_i - x_i\right|}{x_i}\right) \times 100(\%) \tag{28}$$

The method that has better PAR value is adopted between them.

B. *Forecasting Results*

Forecasting results for 4 cases are exhibited in Figure 1 through Figure 4.

Forecasting results for 4 cases are exhibited in Figure 1 through Figure 4.

We can observe that each case has a very good forecasting accuracy.

We have confirmed the FAR for all the Rank A items and materials of set product for birth in Rank A. Summary of them is exhibited at Table 2.

Summary of them is exhibited at Table 2.

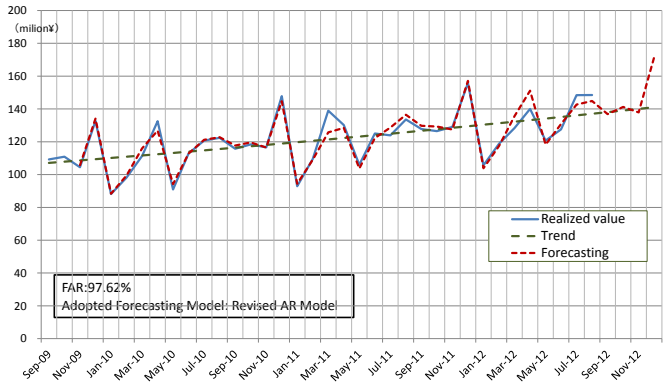


Fig. 1. Forecasting result for Case 1

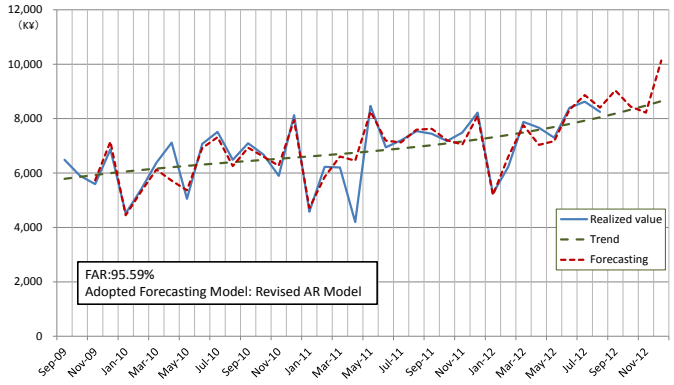


Fig. 2. Forecasting result for Case 2

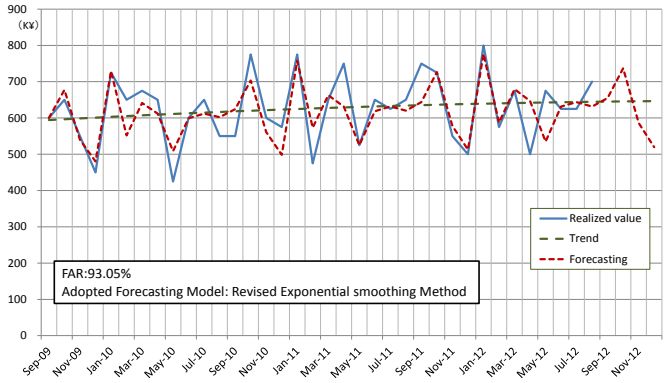


Fig. 3. Forecasting result for Case 3

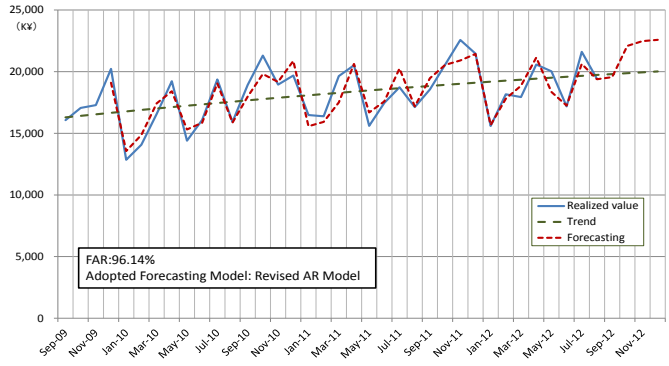


Fig. 4. Forecasting result for Case 4

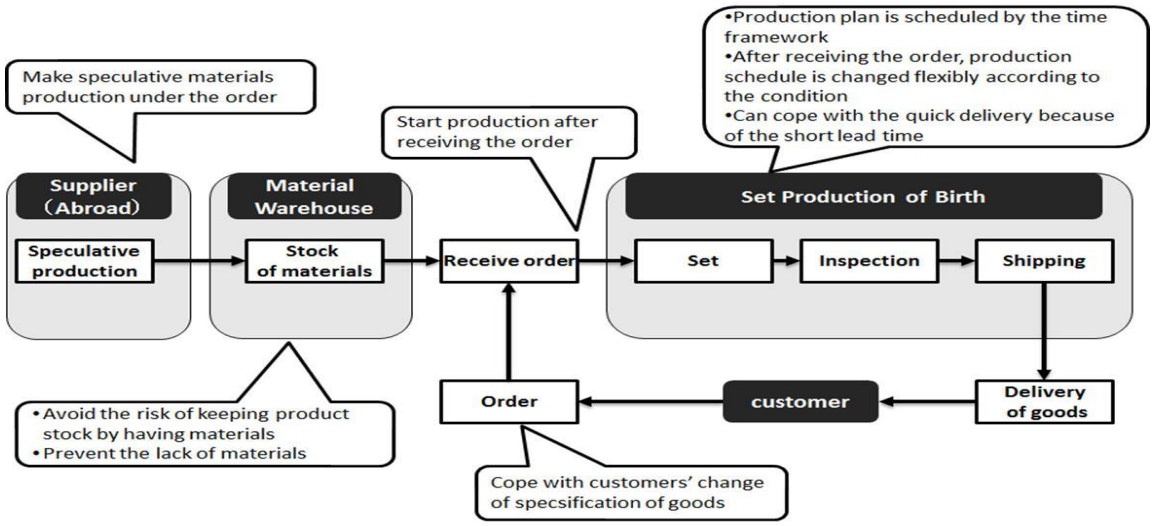


Fig. 5. The Scheme of BTO for the Set Production of Birth

TABLE I. THEIR SUMMARY IS EXHIBITED IN TABLE 1

	Selected Model	FAR (%)
Case1	Revised AR Mode	97.62
Case2	Revised Exponential Smoothing Method	95.59
Case3	Revised AR Mode	93.05
Case4	Revised AR Mode	96.14

TABLE II. FAR FOR ALL THE RANK A ITEMS

	Number of items	Distribution (FAR) (%)			Mode
		~85%	~80%	~75%	
All the Rank A items	577	70.2	86.5	92.7	90.0
Materials of Set Product for Birth	124	97.6	97.6	98.4	96.0

The mode is 90% for all the Rank A items.

We take up materials of set product for birth in Rank A and the mode is 96%. These are sufficiently enough to utilize these forecasting methods to the actual daily business.

VIII. BUILDING THE BTO SYSTEM FOR THE SET PRODUCTION OF BIRTH

The construction of materials for the set product of birth differs by customer. Therefore procurement becomes difficult. Exquisite forecasting of materials of set product for birth enables procurement to be a stable and correct one, which would contribute to keep the appropriate stock level. From this condition, building the BTO system for the set product of birth is a good way to this set product. In Figure 5, the scheme of BTO system for the set product of birth is exhibited.

Factory does not have product stocks basically. Planner makes order sheet based upon the forecasting results of materials, and he announce unofficially to the suppliers (Abroad, for example, Indonesia) for 3 through 4 months' preliminary order quantities. Suppliers make production under this preliminary order. Thus, it helps the shortening of procurement lead time and lack of materials. After receiving the order from the customer, they soon start production and make delivery after it is completed. The operation is mainly consisted by the hand work of making set, therefore replacing the product with other products can easily be done.

By shortening the lead time, they can cope with the request of customers flexibly. Improvement of cash flow is also achieved by decreasing stock level.

Now, we summarize the advantage of forecasting in this company.

/ Exquisite production planning can be achieved by utilizing the result of precise forecasting. Demand forecasting is the basis and the starting point of Supply Chain Management.

/ Exquisite preliminary order can be built, which causes less stocks and shorter lead time in suppliers.

/ Smooth BTO is established with short lead time and stockless production by co-operating with suppliers, which will derive win-win relationship among maker and suppliers..

IX. CONCLUSION

In recent years, BTO (Build to Order) system is prevailing. It pursues short lead time, minimum stocks, and thereby minimum cost. But the high accuracy demand forecasting is inevitable for the parts manufacturers.

In this paper, well organized BTO system in the sanitary materials manufacturer is seek with the aid of high accuracy demand forecasting, which is newly developed by us. Focusing that the equation of ESM is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in ESM was derived. Trend removal method was also devised. AR model is also used for forecasting. After removing trend, AR model is utilized and forecasting is executed. Better one in the forecasting accuracy between them was chosen for the final forecasting. Thus, we could obtain the high accuracy demand forecasting. These methods are examined by the data of sanitary materials manufacturer and the BTO system is newly built by utilizing this method. Smooth BTO is established with short lead time

and stockless production by co-operating with suppliers, which will derive win-win relationship among maker and suppliers. By shortening the lead time, they can cope with the request of customers flexibly. Improvement of cash flow is also achieved by decreasing stock level..

X. FUTURE WORKS

Future works of this paper are as follows.

/ To examine the advantage in the performed case

/ To further develop the BTO system in the performed case

/ To Extend the BTO system in various cases

Further development of this system should be performed hereafter. In the end, we appreciate Mr. Norio Funato for his helpful support of our study.

REFERENCES

[1] G. Parry and A. Graves Editors, Build To Order: The Road to the 5-Day Car, Springer, 2008.

[2] D. M. Anderson, Build-to-Order & Mass Customization, CIM Press, 2004.

[3] Fukushima, Kazunobu / Kohmura, Toshitaka / Oshima, Taku / Zhang, Jixun / Kiuchi, Masamitsu, “Developing Supply Chains of Japanese Manufacturers in China”, Josai management review, Vol.2,No.1,pp.97-110,2006.

[4] Fukushima, Kazunobu / Kohmura, Toshitaka / Oshima, Taku / Zhang, Jixun / Kiuchi, Masamitsu, “Developing Supply Chains of Japanese Manufacturers in China (Part 2)”, Josai management review, Vol.3, No.1, pp.47-60, 2007.

[5] Kazuhiro Takeyasu and Keiko Nagata, “Estimation of Smoothing Constant of Minimum Variance with Optimal Parameters of Weight”, International Journal of Computational Science Vol.4,No.5, pp. 411-425,2010

A Hybrid Method to Improve Forecasting Accuracy in the Case of Sanitary Materials Data

Daisuke Takeyasu[1]
[1]Graduate School of Culture and Science, The Open University of Japan, 2-11 Wakaba, Mihama-District, Chiba City, 261-8586, Japan

Hirotake Yamashita[2]
[2]College of Business Administration and Information Science, Chubu University, 1200 Matsumoto-cho Kasugai, Aichi, 487-8501, Japan

Kazuhiro Takeyasu[3]
[3]College of Business Administration, Tokoha University, 325 Oobuchi, Fuji City, Shizuoka, 417-0801, Japan

***Abstract*—Sales forecasting is a starting point of supply chain management, and its accuracy influences business management significantly. In industries, how to improve forecasting accuracy such as sales, shipping is an important issue. In this paper, a hybrid method is introduced and plural methods are compared. Focusing that the equation of exponential smoothing method (ESM) is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in exponential smoothing method is proposed before by Takeyasu et.al. which satisfies minimum variance of forecasting error. Firstly, we make estimation of ARMA model parameter and then estimate smoothing constants. In this paper, combining the trend removing method with this method, we aim to improve forecasting accuracy. Trend removing by the combination of linear and 2nd order non-linear function and 3rd order non-linear function is carried out to the manufacturer's data of sanitary materials.**

The new method shows that it is useful for the time series that has various trend characteristics and has rather strong seasonal trend. The effectiveness of this method should be examined in various cases.

***Keywords*—component; minimum variance; exponential smoothing method; forecasting; trend; sanitary materials**

I. INTRODUCTION

The needs for sales forecasting is prevailing among companies, but the contents of such needs are undergoing significant changes because of the rapid changes in the recent business environment. Correct forecasting along with supply chain management is required that leads to the shortened lead time and less stocks.

Time series analysis is often used in such themes as sales forecasting, stock market price forecasting etc. Sales forecasting is inevitable for Supply Chain Management. But in fact, it is not well utilized in industries. It is because there are so many irregular incidents therefore it becomes hard to make sales forecasting. A mere application of method does not bear good result. The big reason is that sales data or production data are not stationary time series, while linear model requires the time series as a stationary one. In order to improve forecasting accuracy, we have devised trend removal methods as well as searching optimal parameters and obtained good results. We created a new method and applied it to various time series and examined the effectiveness of the method. Applied data are sales data, production data, shipping data, stock market price data, flight passenger data etc.

Many methods for time series analysis have been presented such as Autoregressive model (AR Model), Autoregressive Moving Average Model (ARMA Model) and Exponential Smoothing Method (ESM)[1]－[4]. Among these, ESM is said to be a practical simple method.

For this method, various improving method such as adding compensating item for time lag, coping with the time series with trend [5], utilizing Kalman Filter[6], Bayes Forecasting[7], adaptive ESM[8], exponentially weighted

Moving Averages with irregular updating periods [9], making averages of forecasts using plural method [10] are presented. For example, Maeda[6] calculated smoothing constant in relationship with S/N ratio under the assumption that the observation noise was added to the system. But he had to calculate under supposed noise because he could not grasp observation noise.

It can be said that it does not pursue optimum solution from the very data themselves which should be derived by those estimation. Ishii [11] pointed out that the optimal smoothing constant was the solution of infinite order equation, but he didn't show analytical solution. Based on these facts, a new method of estimation of smoothing constant in ESM was proposed before [12]. Focusing that the equation of ESM is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in ESM was derived. Furthermore, combining the trend removal method, forecasting accuracy was improved, where shipping data, stock market price data etc. were examined [13] －[19].

In this paper, utilizing above stated method, a revised forecasting method is proposed. A mere application of ESM does not make good forecasting accuracy for the time series which has non-linear trend and/or trend by month. A new method to cope with this issue is required. Therefore, utilizing above stated method, a revised forecasting method is proposed in this paper to improve forecasting accuracy. In making forecast such as production data, trend removing method is

devised. Trend removing by the combination of linear and 2^{nd} order non-linear function and 3^{rd} order non-linear function is executed to the manufacturer's data of sanitary materials. The weights for these functions are set 0.5 for two patterns at first and then varied by 0.01 increment for three patterns and optimal weights are searched. For the comparison, monthly trend is removed after that. Theoretical solution of smoothing constant of ESM is calculated for both of the monthly trend removing data and the non-monthly trend removing data. Then forecasting is executed on these data. This is a revised forecasting method. Variance of forecasting error of this newly proposed method is assumed to be less than those of previously proposed method. The new method shows that it is useful especially for the time series that has stable characteristics and has rather strong seasonal trend and also the case that has non-linear trend. The rest of the paper is organized as follows. In section 2, the new method is described. ESM is stated by ARMA model and estimation method of smoothing constant is derived using ARMA model identification. The combination of linear and non-linear function is introduced for trend removing and the Monthly Ratio is also referred. Forecasting is executed in section 3, and estimation accuracy is examined, which is followed by the Discussion of section 4

II. Description of The New Method

A. Description of ESM Using ARMA Model[12]

In ESM, forecasting at time $t+1$ is stated in the following equation.

$$\begin{aligned}\hat{x}_{t+1} &= \hat{x}_t + \alpha\left(x_t - \hat{x}_t\right) \\ &= \alpha x_t + (1-\alpha)\hat{x}_t\end{aligned} \tag{1}$$

Here,

$\hat{x}_{t+1}$: forecasting at $t+1$

x_t : realized value at t

α : smoothing constant $(0<\alpha<1)$

(1) is re-stated as

$$\hat{x}_{t+1} = \sum_{l=0}^{\infty} \alpha(1-\alpha)^l x_{t-l} \tag{2}$$

By the way, we consider the following (1,1) order ARMA model.

$$x_t - x_{t-1} = e_t - \beta e_{t-1} \tag{3}$$

Generally, (p,q) order ARMA model is stated as

$$x_t + \sum_{i=1}^{p} a_i x_{t-i} = e_t + \sum_{j=1}^{q} b_j e_{t-j} \tag{4}$$

Here,

$\{x_t\}$:

Sample process of Stationary Ergodic Gaussian Process $x(t)\ \ t = 1,2,\cdots,N,\cdots$

$\{e_t\}$: Gaussian White Noise with 0 mean σ_e^2 variance

MA process in (4) is supposed to satisfy convertibility condition. Utilizing the relation that

$$E\left[e_t \middle| e_{t-1}, e_{t-2}, \cdots\right] = 0$$

we get the following equation from (3).

$$\hat{x}_t = x_{t-1} - \beta e_{t-1} \tag{5}$$

Operating this scheme on $t+1$, we finally get

$$\begin{aligned}\hat{x}_{t+1} &= \hat{x}_t + (1-\beta)e_t \\ &= \hat{x}_t + (1-\beta)\left(x_t - \hat{x}_t\right)\end{aligned} \tag{6}$$

If we set $1-\beta=\alpha$, the above equation is the same with (1), i.e., equation of ESM is equivalent to (1,1) order ARMA model, or is said to be (0,1,1) order ARIMA model because 1st order AR parameter is -1 [1][3]. Focusing that the equation of exponential smoothing method (ESM) is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in exponential smoothing method is derived.

Finally we get:

$$\left.\begin{aligned} b_1 &= \frac{1-\sqrt{1-4\rho_1^2}}{2\rho_1} \\ \alpha &= \frac{1+2\rho_1-\sqrt{1-4\rho_1^2}}{2\rho_1} \end{aligned}\right\} \tag{7}$$

Thus we can obtain a theoretical solution by a simple way.

Here ρ_1 must satisfy

$$-\frac{1}{2} < \rho_1 < 0 \tag{8}$$

in order to satisfy $0<\alpha<1$.

Focusing on the idea that the equation of ESM is equivalent to (1,1) order ARMA model equation, we can estimate smoothing constant after estimating ARMA model parameter.

It can be estimated only by calculating 0th and 1st order autocorrelation function.

B. Trend Removal Method[12]

As ESM is a one of a linear model, forecasting accuracy for the time series with non-linear trend is not necessarily good. How to remove trend for the time series with non-linear

trend is a big issue in improving forecasting accuracy. In this paper, we devise to remove this non-linear trend by utilizing non-linear function.

As trend removal method, we describe the combination of linear and non-linear function.

[1] Linear function

We set

$$y = a_1 x + b_1 \tag{9}$$

as a linear function.

[2] Non-linear function

We set

$$y = a_2 x^2 + b_2 x + c_2 \tag{10}$$

$$y = a_3 x^3 + b_3 x^2 + c_3 x + d_3 \tag{11}$$

as a 2nd and a 3rd order non-linear function.

[3] The combination of linear and non-linear function

We set

$$y = \alpha_1 (a_1 x + b_1) + \alpha_2 (a_2 x^2 + b_2 x + c_2) \tag{12}$$

$$y = \beta_1 (a_1 x + b_1) + \beta_2 (a_3 x^3 + b_3 x^2 + c_3 x + d_3) \tag{13}$$

$$y = \gamma_1 (a_1 x + b_1) + \gamma_2 (a_2 x^2 + b_2 x + c_2) + \gamma_3 (a_3 x^3 + b_3 x^2 + c_3 x + d_3) \tag{14}$$

as the combination of linear and 2nd order non-linear and 3rd order non-linear function. Here, $\alpha_2 = 1 - \alpha_1$, $\beta_2 = 1 - \beta_1$, $\gamma_3 = 1 - (\gamma_1 + \gamma_2)$. Comparative discussion concerning (12), (13) and (14) are described in section 5.

C. Monthly Ratio[12]

For example, if there is the monthly data of L years as stated bellow:

$$\{x_{ij}\} \ (i = 1, \cdots, L) \ (j = 1, \cdots, 12)$$

Where, $x_{ij} \in R$ in which j means month and i means year and x_{ij} is a shipping data of i-th year, j-th month. Then, monthly ratio $\tilde{x}_j \ (j = 1, \cdots, 12)$ is calculated as follows.

$$\tilde{x}_j = \frac{\frac{1}{L}\sum_{i=1}^{L} x_{ij}}{\frac{1}{L} \cdot \frac{1}{12} \sum_{i=1}^{L} \sum_{j=1}^{12} x_{ij}} \tag{15}$$

Monthly trend is removed by dividing the data by (15). Numerical examples both of monthly trend removal case and non-removal case are discussed in 5.

III. Forecasting The Shipping Data Of Manufacturer

A. Analysis Procedure

Manufacturer's data of sanitary materials from September 2009 to August 2012 are analyzed. First of all, graphical charts of these time series data are exhibited in Fig. 1, 2,3.

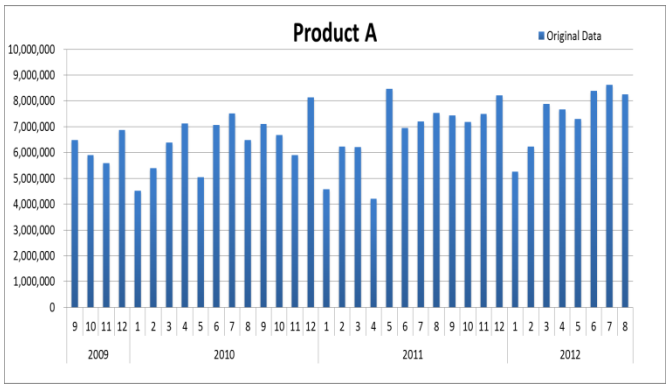


Fig. 1. Product A

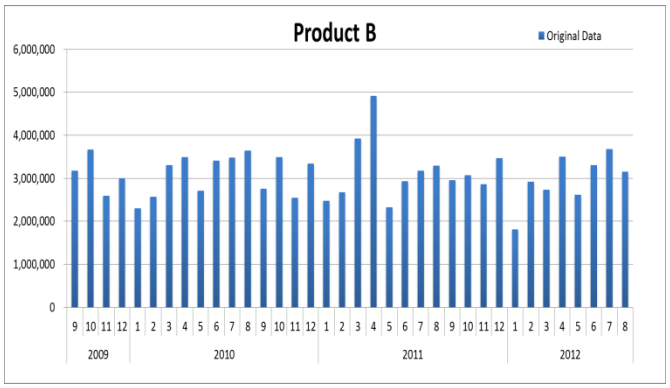


Fig. 2. Product B

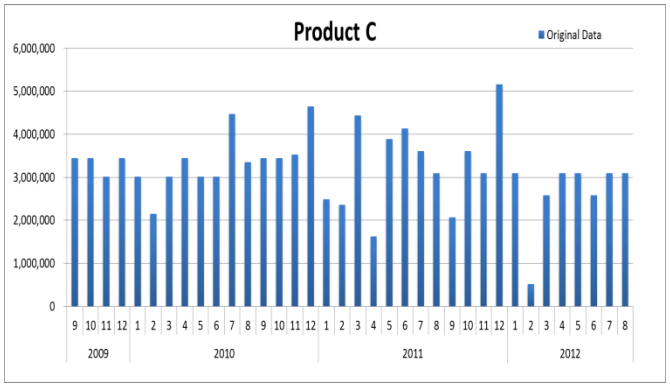


Fig. 3. Product C

Analysis procedure is as follows. There are 36 monthly data for each case. We use 24 data(1 to 24) and remove trend by the method stated in 2.2. Then we calculate monthly ratio by the method stated in 2.3. After removing monthly trend, the method stated in 2.1 is applied and Exponential Smoothing Constant with minimum variance of forecasting error is estimated.

Then 1 step forecast is executed. Thus, data is shifted to 2nd to 25th and the forecast for 26th data is executed consecutively, which finally reaches forecast of 36th data. To examine the accuracy of forecasting, variance of forecasting error is calculated for the data of 25th to 36th data. Final forecasting data is obtained by multiplying monthly ratio and trend. Forecasting error is expressed as:

$$\varepsilon_i = \hat{x}_i - x_i \tag{16}$$

$$\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \tag{17}$$

B. Trend Removing

Trend is removed by dividing original data by (12),(13),(14). The patterns of trend removal are exhibited in Table1.

In pattern1 and 2, the weight of α_1, α_2, β_1, β_2 are set 0.5 in the equation (12),(13). In pattern3, the weight of α_1 is shifted by 0.01 increment in (12) which satisfy the range $0 \le \alpha_1 \le 1.00$. In pattern4, the weight of β_1 is shifted in the same way which satisfy the range $0 \le \beta_1 \le 1.00$. In pattern5, the weight of γ_1 and γ_2 are shifted by 0.01 increment in (14) which satisfy the range $0 \le \gamma_1 \le 1.00$, $0 \le \gamma_2 \le 1.00$ .The best solution is selected which minimizes the variance of forecasting error. Estimation results of coefficient of (9), (10) and (11) are exhibited in Table 2. Estimation results of weights of (12), (13) and (14) are exhibited in Table 3.

Graphical chart of trend is exhibited in Fig. 4, 5, 6 for the cases that monthly ratio is used.

TABLE I. COEFFICIENT OF (9),(10) AND (11)

	1st		2nd			3rd			
	a_1	b_1	a_2	b_2	c_2	a_3	b_3	c_3	d_3
Product A	43209.10957	5856643.297	1333.16432	9880.001554	6001069.432	567.0224981	-19930.17936	226879.5116	5503507.19
Product B	12078.85565	2983007.054	50.89560657	10806.46549	2988520.745	-249.7081938	9414.952874	-84756.86028	3207639.685
Product C	12459.34348	3155302.373	-397.7965947	22404.25835	3112207.742	-256.3640436	9215.855041	-75706.26115	3337167.19

TABLE II. THE PATTERNS OF TREND REMOVAL

Pattern1	α_1, α_2 are set 0.5 in the equation (12)
Pattern2	β_1, β_2 are set 0.5 in the equation (13)
Pattern3	α_1 is shifted by 0.01 increment in (12)
Pattern4	β_1 is shifted by 0.01 increment in (13)
Pattern5	γ_1 and γ_2 are shifted by 0.01 increment in (14)

TABLE III. WEIGHTS OF (12), (13) AND (14)

	Monthly ratio	Pattern1		Pattern2		Pattern3		Pattern4		Pattern5		
		α_1	α_2	β_1	β_2	α_1	α_2	β_1	β_2	γ_1	γ_2	γ_3
Product A	Used	0.5	0.5	0.5	0.5	0	1	1	0	0	1	0
	Not used	0.5	0.5	0.5	0.5	0	1	1	0	0	1	0
Product B	Used	0.5	0.5	0.5	0.5	0.1	0.9	0.92	0.08	0.1	0.9	0
	Not used	0.5	0.5	0.5	0.5	0.37	0.63	1	0	0.37	0.63	0
Product C	Used	0.5	0.5	0.5	0.5	0.47	0.53	1	0	0.47	0.53	0
	Not used	0.5	0.5	0.5	0.5	1	0	1	0	1	0	0

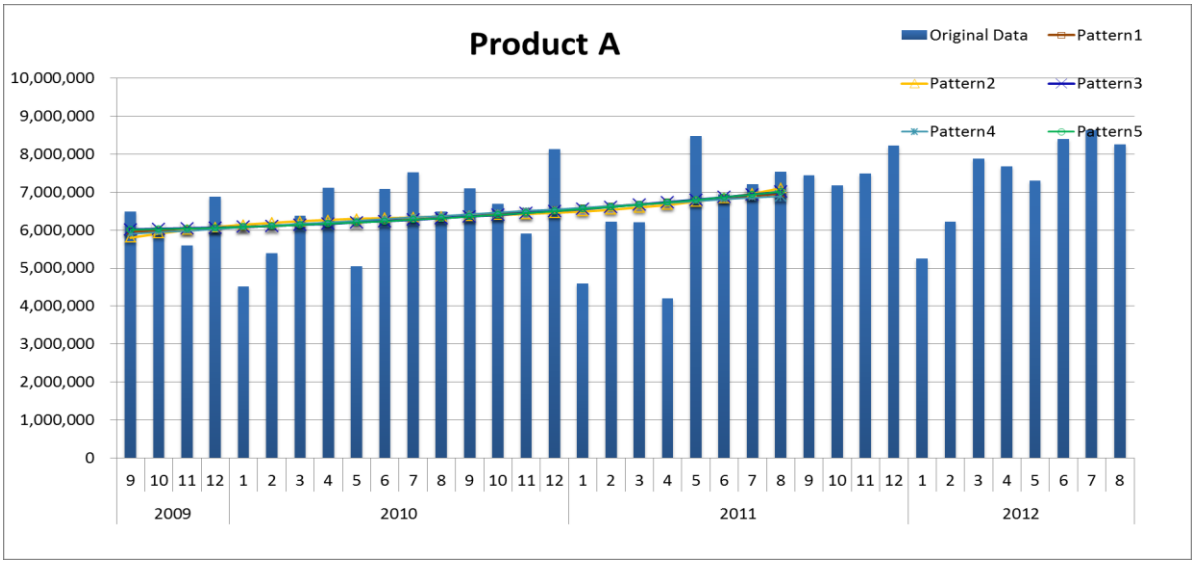


Fig. 4. Trend of Product A

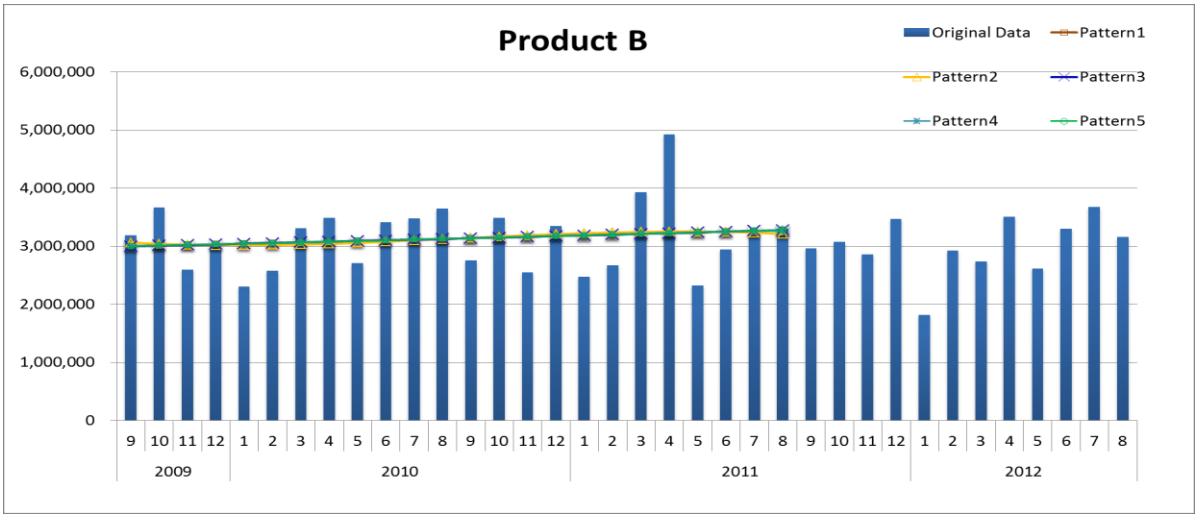


Fig. 5. Trend of Product B

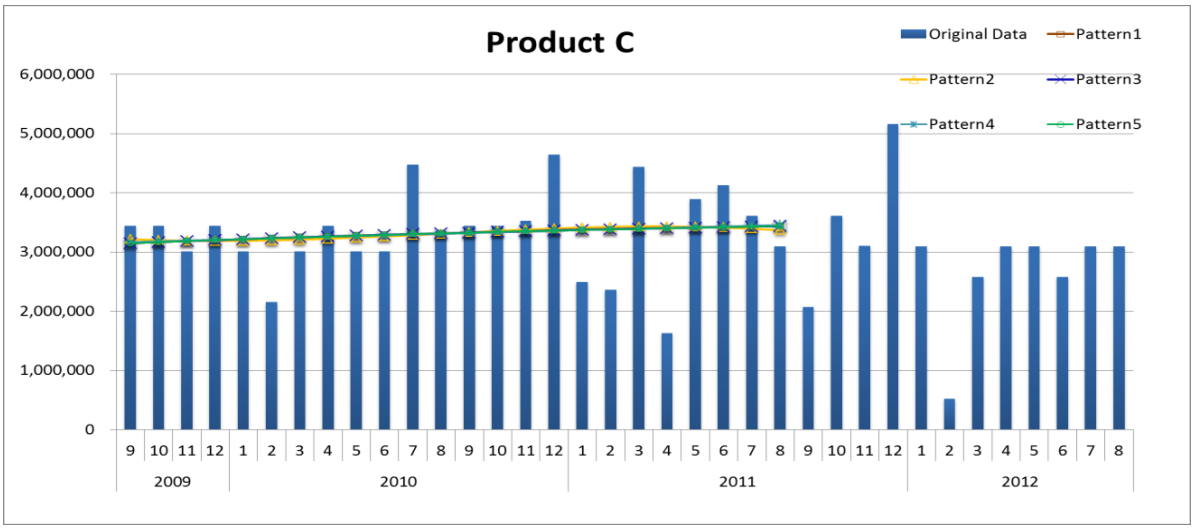


Fig. 6. Trend of Product C

C. Removing trend of monthly ratio

After removing trend, monthly ratio is calculated by the method stated in 2.3.

Calculation result for 1st to 24th data is exhibited in Table 4 through 8.

D. Estimation of Smoothing Constant with Minimum Variance of Forecasting Error

After removing monthly trend, Smoothing Constant with minimum variance of forecasting error is estimated utilizing (7). There are cases that we cannot obtain a theoretical solution because they do not satisfy the condition of (A-9).

In those cases, Smoothing Constant with minimum variance of forecasting error is derived by shifting variable from 0.01 to 0.99 with 0.01 interval. Calculation result for 1st to 24th data is exhibited in Table 9.

E. Forecasting and Variance of Forecasting Error

Utilizing smoothing constant estimated in the previous section, forecasting is executed for the data of 25th to 36th data. Final forecasting data is obtained by multiplying monthly ratio and trend.

Variance of forecasting error is calculated by (18). Forecasting results are exhibited in Fig. 7, 8, 9 for the cases that monthly ratio is used.

TABLE IV. MONTHLY RATIO (PATTERN1)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Product A	1.10	1.01	0.92	1.19	0.72	0.91	0.98	0.89	1.03	1.07	1.12	1.05
Product B	0.97	1.16	0.83	1.02	0.77	0.84	1.15	1.33	0.80	1.00	1.05	1.09
Product C	1.06	1.06	1.00	1.23	0.84	0.68	1.12	0.77	1.03	1.06	1.20	0.96

TABLE V. MONTHLY RATIO (PATTERN2)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Product A	1.12	1.02	0.93	1.20	0.72	0.91	0.98	0.88	1.03	1.07	1.11	1.04
Product B	0.96	1.15	0.83	1.02	0.77	0.84	1.15	1.33	0.80	1.01	1.05	1.10
Product C	1.05	1.05	0.99	1.22	0.84	0.68	1.12	0.77	1.03	1.07	1.21	0.97

TABLE VI. MONTHLY RATIO (PATTERN3)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Product A	1.10	1.01	0.92	1.19	0.72	0.91	0.89	1.03	1.07	1.12	1.19	1.05
Product B	0.97	1.16	0.83	1.02	0.77	0.84	1.15	1.33	0.80	1.00	1.05	1.09
Product C	1.06	1.06	1.00	1.23	0.84	0.68	1.12	0.77	1.03	1.06	1.20	0.96

TABLE VII. MONTHLY RATIO (PATTERN4)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Product A	1.10	1.01	0.92	1.19	0.72	0.91	0.98	0.89	1.03	1.07	1.12	1.06
Product B	0.97	1.16	0.83	1.02	0.77	0.84	1.15	1.33	0.80	1.00	1.05	1.09
Product C	1.06	1.06	1.00	1.23	0.84	0.68	1.12	0.77	1.03	1.06	1.20	1.00

TABLE VIII. MONTHLY RATIO (PATTERN5)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Product A	1.10	1.01	0.92	1.19	0.72	0.91	0.99	0.89	1.03	1.07	1.12	1.05
Product B	0.97	1.16	0.83	1.02	0.77	0.84	1.15	1.33	0.80	1.00	1.05	1.09
Product C	1.06	1.06	1.00	1.23	0.84	0.68	1.12	0.77	1.03	1.06	1.20	0.96

TABLE IX. ESTIMATED SMOOTHING CONSTANT WITH MINIMUM VARIANCE

	Monthly ratio	Pattern1		Pattern2		Pattern3		Pattern4		Pattern5	
		ρ_1	α	ρ_1	α	ρ_1	α	ρ_1	α	ρ_1	α
Product A	Used	-0.411	0.476	-0.4423	0.3966	-0.3976	0.5050	-0.4210	0.4530	-0.3976	0.5050
	Not used	-0.2679	0.7095	-0.2864	0.6852	-0.2648	0.7135	-0.2691	0.7080	-0.2648	0.7135
Product B	Used	-0.2245	0.7628	-0.2357	0.7496	-0.2246	0.7628	-0.2264	0.7606	-0.2246	0.7628
	Not used	-0.066	0.9337	-0.0749	0.9247	-0.0660	0.9337	-0.0661	0.9336	-0.0660	0.9337
Product C	Used	-0.4232	0.4476	-0.4211	0.4526	-0.4235	0.4469	-0.4178	0.4607	-0.4235	0.4469
	Not used	-0.266	0.7119	-0.2671	0.7106	-0.2656	0.7125	-0.2656	0.7125	-0.2656	0.7125

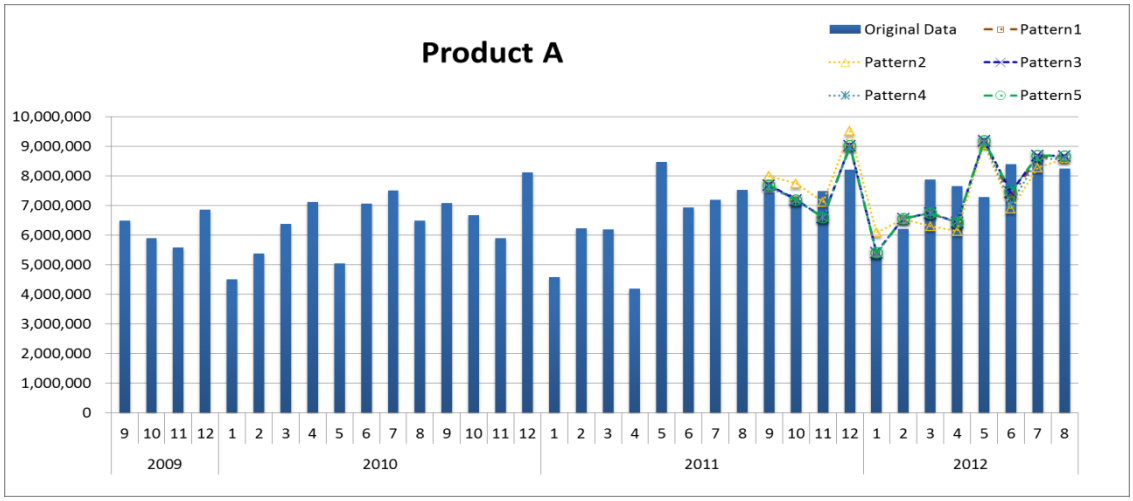


Fig. 7. Forecasting Results of Product A

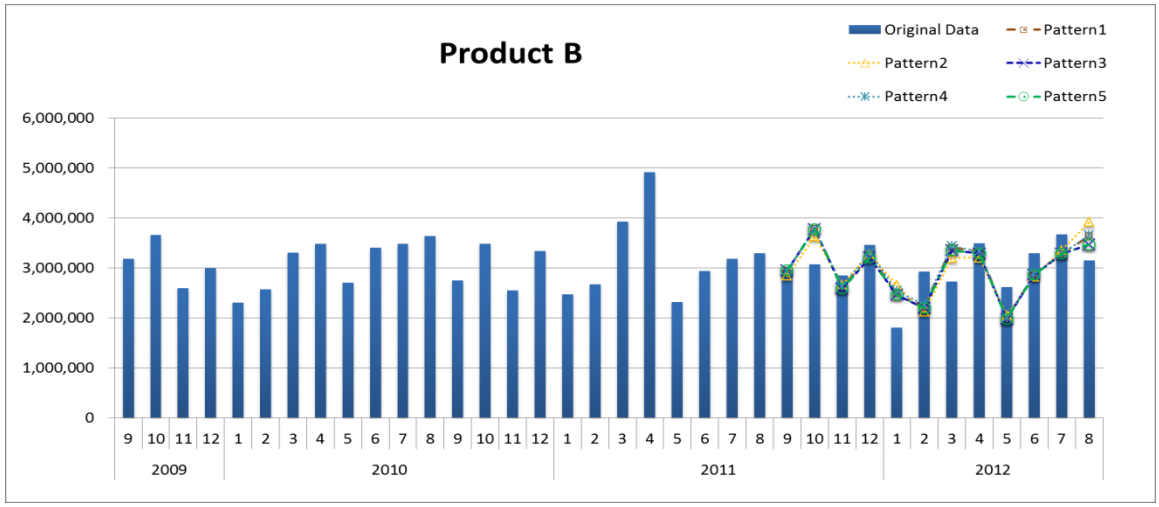


Fig. 8. Forecasting Results of Product B

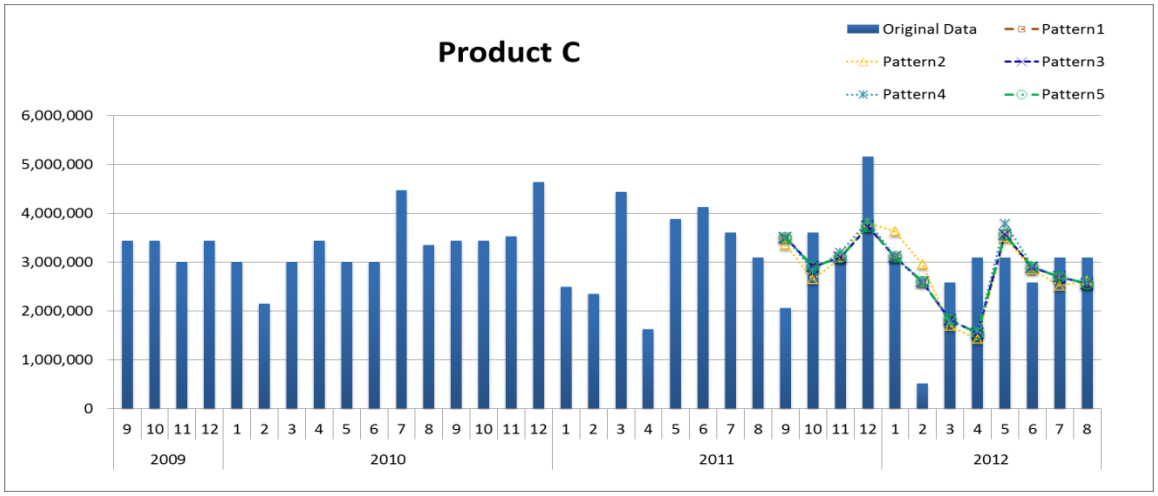


Fig. 9. Forecasting Results of Product C

Variance of forecasting error is exhibited in Table 10

TABLE X. VARIANCE OF FORECASTING ERROR

	Monthly ratio	Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
Product A	Used	841,383,645,691.7740	1,253,890,581,301.6900	828013457221.089	878641598931.531	828013457221.089
	Not used	1,233,162,627,302.1600	1,476,091,369,104.5700	1228777654815.98	1239890302140.87	1228777654815.98
Product B	Used	272,535,112,508.2970	298,632,461,408.8090	258441818256.66	276267182227.964	258441818256.66
	Not used	530,376,395,688.7940	578,434,563,503.2370	529971383044.765	539573173959.263	529971383044.765
Product C	Used	1,144,959,209,811.1200	1,361,652,912,074.6500	1144904100567.48	1161683354954.07	1144904100567.48
	Not used	2,104,397,068,344.4300	2,357,366,912,250.4600	2075183264170.54	2075183264170.54	2075183264170.54

F. Remarks

These time series have non-linear trend and trend by month. Applying only an ESM does not make good forecasting accuracy.

All cases had a good result in 1st+2nd order with the case that monthly ratio is used. We can observe that monthly trend is rather apparent in these cases. Therefore the method has selected the monthly trend removing case.

IV. DISCUSSION

Correct sales forecasting is inevitable in industries. Poor sales forecasting accuracy leads to a gap between the sales plan and result, which in turn generates a gap between the sales plan and the production plan. The condition in which the quantity in a production plan exceeds that in a sales plan (excess production) pushes up cost caused by increased finished and intermediate product inventory. Increased inventory and prolonged dwell time of product in inventory will lead to increased waste loss as well as extended lead-time, affecting customer satisfaction. In order to improve forecasting accuracy, we have devised trend removal methods as well as searching optimal parameters and obtained good results. We created a new method.

V. CONCLUSION

Focusing on the idea that the equation of exponential smoothing method(ESM) was equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in exponential smoothing method was proposed before by Takeyasu et.al.[12] which satisfied minimum variance of forecasting error. Combining the trend removal method with this method, we aimed to improve forecasting accuracy.

A mere application of ESM does not make good forecasting accuracy for the time series which has non-linear trend and/or trend by month. A new method to cope with this issue is required. Therefore, utilizing above stated method, a revised forecasting method is proposed in this paper to improve forecasting accuracy. An approach to this method was executed in the following method. Trend removal by a linear function was applied to the manufacturer's data of sanitary materials. The combination of linear and non-linear function was also introduced in trend removing. For the comparison, monthly trend was removed after that. Theoretical solution of smoothing constant of ESM was calculated for both of the monthly trend removing data and the non-monthly trend removing data.

Then forecasting was executed on these data. Product Ⅰ and Product Ⅱ had a good result in $1^{st}+2^{nd}$ order with the case that monthly ratio is used, while Product Ⅲhad a good result in $1^{st}+2^{nd}$ order with the case that monthly ratio is not used.

VI. FUTURE WORKS

It is our future works to investigate much further cases to confirm the effectiveness of our new method. Various cases should be examined hereafter.

In the end, we appreciate Mr. Norio Funato for his helpful support of our study.

REFERENCES

[1] Box Jenkins, Time Series Analysis Fourth Edition, Wiley,2011.

[2] Kazuhiro Takeyasu and Yuki Higuchi, Time Series Analysis and its Applications, Osaka Municipals Universities Press,2007.

[3] Toru Katayama, Introduction to System Identification, Asakura Shoten,1998.

[4] Kazuhiro Takeyasu, Yasuo Ishii and Yuki Higuchi, Time Series Analysis and its Applications Ⅲ, Osaka Municipals Universities Press,2009.

[5] Peter R.Winters, "Forecasting Sales by Exponentially Weighted Moving Averages", Management Science, Vol6, No.3, pp. 324-343,1984.

[6] Katsuro Maeda, "Smoothing Constant of Exponential Smoothing Method", Seikei University Report Faculty of Engineering, No.38, pp. 2477-2484,1984.

[7] M.West and P.J.Harrison, Baysian Forecastingand Dynamic Models, Springer-Verlag, New York,1989.

[8] Steinar Ekern, "Adaptive Exponential Smoothing Revisited", Journal of the Operational Research Society, Vol. 32 pp. 775-782,1982.

[9] F.R.Johnston, "Exponentially Weighted Moving *Average* (EWMA) with Irregular Updating Periods", Journal of the Operational Research Society, Vol.44, No.7 pp. 711-716,1993.

[10] Spyros Makridakis and Robeat L.Winkler, "Averages of Forecasts ; Some Empirical Results", Management Science, Vol.29, No.9, pp. 987-996,1983.

[11] Naohiro Ishii et al., "Bilateral Exponential Smoothing of Time Series", Int.J.System Sci., Vol.12, No.8, pp. 997-988,1991.

[12] Kazuhiro Takeyasu and Keiko Nagata, "Estimation of Smoothing Constant of Minimum Variance with Optimal Parameters of Weight", International Journal of Computational Science Vol.4,No.5, pp. 411-425,2010.

[13] Kazuhiro Takeyasu, Keiko Nagata, Yuki Higuchi, "Estimation of Smoothing Constant of Minimum Variance And Its Application to Shipping Data With Trend Removal Method", Industrial Engineering & Management Systems (IEMS),Vol.8,No.4, pp.257-263,2009,

[14] Kazuhiro Takeyasu, Keiko Nagata, Yui Nishisako, "A Hybrid Method to Improve Forecasting Accuracy Utilizing Genetic Algorithm And Its Application to Industrial Data", NCSP'10, Honolulu,Hawaii,USA,2010.

[15] Kazuhiro Takeyasu, Keiko Nagata, Kana Takagi, "Estimation of Smoothing Constant of Minimum Variance with Optimal Parameters of Weight", NCSP'10, Honolulu,Hawaii,USA,2010.

[16] Kazuhiro Takeyasu, Keiko Nagata, Tomoka Kuwahara, "Estimation of Smoothing Constant of Minimum Variance Searching Optimal Parameters of Weight", NCSP'10, Honolulu,Hawaii,USA,2010.

[17] Kazuhiro Takeyasu, Keiko Nagata, Mai Ito, Yuki Higuchi, "A Hybrid Method to Improve Forecasting Accuracy Utilizing Genetic Algorithm", The 11th APEIMS, Melaka, Malaysia,2010.

[18] Kazuhiro Takeyasu, Keiko Nagata, Kaori Matsumura, "Estimation of Smoothing Constant of Minimum Variance and Its Application to Sales Data", JAIMS, Honolulu, Hawaii, USA.2011.

[19] Hiromasa Takeyasu, Yuki Higuchi, Kazuhiro Takeyasu, "A Hybrid Method to Improve Forecasting Accuracy in the Case of Bread", International Journal of Information and Communication Technology Research, Vol.2 , No.11, pp.804～812,2012.

A Comparative Study of Game Tree Searching Methods

Ahmed A. Elnaggar
Computer Science Department
Modern Academy in Maadi
Cairo, Egypt

Mostafa Abdel Aziem
College of Computing and IT
AAST
Cairo, Egypt

Mahmoud Gadallah
Computer Science Department
Modern Academy in Maadi
Cairo, Egypt

Hesham El-Deeb
Computer Science Faculty
MTI University
Cairo, Egypt

***Abstract*—In this paper, a comprehensive survey on gaming tree searching methods that can use to find the best move in two players zero-sum computer games was introduced. The purpose of this paper is to discuss, compares and analyzes various sequential and parallel algorithms of gaming tree, including some enhancement for them. Furthermore, a number of open research areas and suggestions of future work in this field are mentioned.**

Keywords—game tree search; searching; evaluation; parallel; distributed; GPU

I. INTRODUCTION

Game playing; where a human play versus a computer have a long history. In the earlier game playing programs, the computer couldn't win a human because of the weakness of the game tree algorithms for finding the best next-move or the limitation of computer computation and memory space. The advances in this field and the field of computer architecture finally allowed computers to win humans in most complex games, including chess. Many algorithms have already been invented to find the best next-move for the computer, including sequential algorithms such as MiniMax [1], NegaMax [2], Negascout [3], SSS* [4] and B* [5] as well as parallel algorithms such as Parallel Alpha-Beta algorithm [6]. These Parallel algorithms are now modified to run not only on CPUs, but also on GPUs [7] to provide a faster solution.

Almost all game playing programs use game trees to find the next-best move. An example of a game tree for Tic-Tac-Toe game [8] is shown in Fig. 1 The figure shows the following:

- Each node represents a game state.
- The root represents the current game state.
- All the branches for a given node represent all the legal moves for that node.
- The node that doesn't have any successor called a leaf.

Evaluation function is used to determine whatever a leaf represents a win, lose, draw or just a score, in case the algorithm was stopped before any player won, lose or the game ended with a draw. The developers usually do that because in more complex games, there is no practical algorithm that can search in the entire tree in a reasonable amount of time even if it uses the power of parallel processing. An example for this is the checkers and chess game where they need to evaluate about 10^{20} and 10^{40} nodes respectively. W^D is used as an estimation of the number of nodes needs to be visited, where W represents the average number of legal moves for each node, and D represents the game length. Two solutions for this problem are to use a fixed depth "D" or to use a specific time to stop generating the tree.

There are many categorization methods for sequential game tree. However, the most common categorization is based on depth-first and breadth-first, which was used in this paper. Depth-first search "DFS" [9] means the algorithm will start from the root and explores as long as the depth limitation didn't meet along each branch before backtracking. Breadth-first search "BFS" [10] begins from the root and inspects all the children of the root node before it inspects all the children of each root children's node. The parallel algorithms are hard to categorize as depth or breadth first since some parallel algorithms work as follows: each core or each processor inspects a child of the root using DFS or BFS. In the first case, the distribution of the root's children uses a BFS algorithm, but each core or processor uses a DFS. In this case, it is called a hybrid-system.

To make it clear for the reader, the paper was organized as follows:

Section [II], presents a discussion for sequential game tree algorithms categorized into depth-first & breadth-first algorithms. Section [III], presents a discussion for parallel game tree algorithms from the programming point of view and from the hardware point of view. Section, [IV], provides an analysis for depth algorithms and breadth algorithms based on algorithm complexity for both time and memory. Section [V], concludes the paper with some future work that can enhance the current algorithms.

II. SEQUENTIAL GAME TREE ALGORITHMS

As mentioned in the previous section, sequential algorithms were categorized into depth-first search [9] and breadth-first search [10]. Furthermore, the depth-first search algorithms categorized again into brute-force search and selective search [11]. The brute-force search is looking at every variation to a given depth while the selective search is looking at important branches only. Section [A], presents a discussion for the brute-force search in depth-first search. Section [B], presents a discussion for the selective search in depth-first search. Section [C], presents a discussion for the breadth-first search.

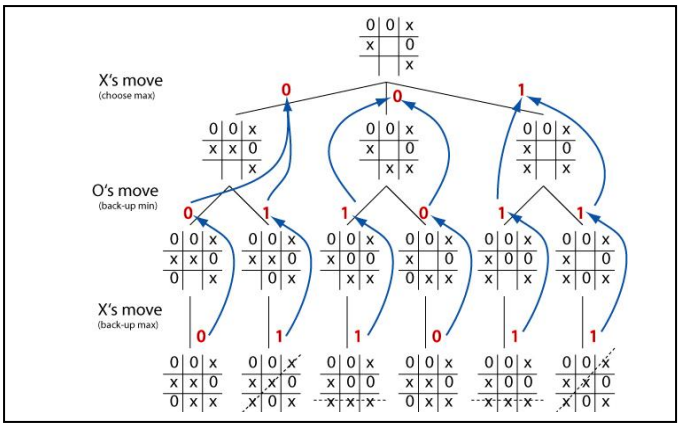


Fig. 1. Game tree for Tic-Tac-Toe game using MiniMax algorithm.

A. Brute-Force algorithms in Depth-First Search

The most famous algorithms in brute-force search are MiniMax [1], NegaMax [2], Alpha-Beta [12], NegaScout [3], and Principle-Variation [13]. Following is a description of each of these algorithms.

MiniMax [1] algorithm is a game tree algorithm that is divided into two logically stages, the first one for the first player which is the computer and the second one for the second player which is the human. The algorithm tries to find the best legal move for the computer even if the human plays the best move for him. Which means, it maximizes the computer score when it chooses the computer move, while minimizing that score by choosing the best legal move for the human when it chooses the human move.

In Fig. 1 there is a simulation of MiniMax search for Tic-Tac-Toe game [8]. Every node has a value calculated by the evaluation function. For a given path, the value of the leaf nodes passed back to its parent. An example for that the value of any O's move will always choose the minimum value for the computer, while the value for any X's move will always choose the maximum value for the computer.

In Fig. 2 a pseudo code for the MiniMax algorithm [1] is presented. The program first calls the function MiniMax which starts the chain of calls for MaxMove and MinMove. Each time MaxMove function or MinMove function is called it automatically calls the other function until the game ended, or it reached the desired depth.

NegaMax [2] algorithm is an identical algorithm for MiniMax with only one slight difference. It uses only the maximization function rather than using both maximization and minimization functions. This can be done by negating the value that is returned from the children from the opponent's point of view rather than searching for the minimum score. This is possible because of the following mathematical relation:

$$\text{Max}(a, b) == -\text{Min}(-a, -b) \quad (1)$$

```
MinMax (GamePosition game) {
 return MaxMove (game);
}

MaxMove (GamePosition game) {
 if (GameEnded(game)) {
  return EvalGameState(game);
 }
 else {
  best_move < - {};
  moves <- GenerateMoves(game);
  ForEach moves {
    move <- MinMove(ApplyMove(game));
    if (Value(move) > Value(best_move)) {
      best_move < - move;
    }
  }
  return best_move;
 }
}

MinMove (GamePosition game) {
 best_move <- {};
 moves <- GenerateMoves(game);
 ForEach moves {
   move <- MaxMove(ApplyMove(game));
   if (Value(move) > Value(best_move)) {
     best_move < - move;
   }
 }

 return best_move;
}
```

Fig. 2. MiniMax Algorithm Pseduo Code

In Fig. 3 there is a pseudo code for NegaMax algorithm. Clearly, (1) was used to simplify the MiniMax algorithm.

Alpha-Beta [12] algorithm is a smart modification that can be applied to MiniMax or NegaMax algorithms. Kunth and Moore proved that many branches could be pruned away of the game tree which reduces the time needed to finish the tree, and it will give the same result as MiniMax or NegaMax. The main idea of the algorithm is cutting the uninteresting branches in the game tree. The following examples illustrating the idea: Max (8, Min (5, X)) and Min (3, Max (7, Y)) The result is always 8 in the first example and 3 in the second example, no matter the values of X or Y. This means the algorithm can cut the node X or Y with its branches. The Alpha-Beta algorithm uses two variables (alpha & beta) to detect these cases, so any value less than alpha or larger than

beta will automatically cutoff without affecting the result of the search tree.

The enhanced version of the NegaMax algorithm from Fig. 3 with Alpha-Beta property is shown in Fig. 4.

```
// Search game tree to given depth, and return evaluation of
// root node.
int NegaMax(gamePosition, depth)
{
        if (depth=0 || game is over)
                // evaluate leaf gamePositionition from
                // current player's standpoint
                return Eval (gamePosition);
         // present return value
        score = - INFINITY;
        // generate successor moves
        moves = Generate(gamePosition);
        // look over all moves
        for i =1 to sizeof(moves) do
        {
                // execute current move
                Make(moves[i]);
                // call other player, and switch sign of
                // returned value
                cur = -NegaMax(gamePosition, depth-1);
                 // compare returned value and score
                // value, update it if necessary
                if (cur > score) score = cur;
                // retract current move
                Undo(moves[i]);
        }
        return score;
}
```

Fig. 3. NegaMax Algorithm Pseudo Code

Several enhancements for the Alpha-Beta algorithm was published [13]; some of them is listed as follows:

- Move Ordering: The speed and the number of cutoffs of the Alpha-Beta algorithm can change dramatically depending on the moving search order. The best move should be examined first, and then the second best move and so on. This will maximize the effectiveness of the algorithm. Many techniques developed to solve this problem, including:
    - Iterative deepening.
    - Transposition tables.
    - Killer Move Heuristic.
    - History Heuristic.
- Minimal Window Search: Alpha-Beta algorithms depend on the values of alpha and beta to cutoff the branches, so by narrowing the search window by changing the values of alpha and beta; it will increase the possibilities of the cutoffs. Many algorithms such as NegaScout [3] and MTD (f) [14] used this property to improve the Alpha-Beta algorithm which discussed below.

```
// Search game tree to given depth, and return evaluation of
// root node.
int AlphaBeta(gamePosition, depth, alpha, beta)
{
        if (depth=0 || game is over)
        // evaluate leaf gamePositionition from
        // current player's standpoint
                return Eval (gamePosition);
         // present return value
        score = - INFINITY;
        // generate successor moves
        moves = Generate(gamePosition);
        // look over all moves
        for i =1 to sizeof(moves) do
        {
                // execute current move
                Make(moves[i]);
                // call other player, and switch sign of
                // returned value
                cur = -AlphaBeta(gamePosition, depth-1,
                                -beta, -alpha);
                // compare returned value and score
                // value, update it if necessary
                if (cur > score) score = cur;
                // adjust the search window
                if (score > alpha) alpha = score;
                // retract current move
                Undo(moves[i]);
                 // cut off
                if (alpha >= beta) return alpha;
        }
        return score;
}
```

Fig. 4. Enhanced NegaMax with Alpha-Beta Property Pseudo Code

The NegScout [3] and Principal Variation Search [13] algorithms were based on the scout algorithm which was an enhanced version of the Alpha-Beta algorithm that can make more cutoffs in the game tree. It contains a new test condition that checks whatever the first node in the siblings is either less than or equal to beta value or greater than or equal to the alpha value. If the result of the condition is true, then the algorithm cuts off the root node for these siblings, and if it is false, then it searches the rest of the siblings to get the new values of alpha and beta.

In Fig. 5 there is a pseudo code for the NegaScout [3]. It looks like the same algorithm in Fig. 4 with the modification of the minimal window search.

B. Selectivity algorithms in Depth-First Search

The main difference between the brute-force algorithms and the selectivity algorithms; it doesn't depend on fixed depth to stop looking in each branch. The most common techniques in this category are Quiescence Search and Forward Pruning.

Below is a discussion of the implementation of the Quiescence Search technique [15] and ProbCut [16] algorithm that is based on the Forward Pruning technique.

```
// Search game tree to given depth, and return evaluation of
// root node.
int NegaScout(gamePosition, depth, alpha, beta)
{
        if (depth=0 || game is over)
        // evaluate leaf gamePositionition from
        // current player's standpoint
                return Eval (gamePosition);
         // present return value
        score = - INFINITY;
        n = beta;
        // generate successor moves
        moves = Generate(gamePosition);
        // look over all moves
        for i =1 to sizeof(moves) do
        {
                // execute current move
                Make(moves[i]);
                // call other player, and switch sign of
                // returned value
                cur = -NegaScout(gamePosition, depth-1,
                                -n, -alpha);
                if (cur > score) {
                        if (n = beta ) OR (d <= 2)
                        // compare returned value and
                        // score value, update it if
                        // necessary
                                score = cur;
                        else
                                score = -NegaScout
                                (gamePosition,depth-1,
                                  -beta, -cur);
                }
                // adjust the search window
                if (score > alpha) alpha = score;
                // retract current move
                Undo(moves[i]);
                 // cut off
                if (alpha >= beta) return alpha;
                n = alpha+1;
        }
        return score;
}
```

Fig. 5. NegaScout Algorithm Pseudo Code Using the Minimal Window Search Principle

Quiescence Search [15] based on the idea of variable depth searching. The algorithm follows the normal fixed depth in most branches. However, in some branches the algorithm takes a deeper look and increases the search depth. An example of that is the chess game where in critical moves like checks or promotions, the algorithms extend the depth to make sure there is no threat exists.

In Fig. 6 there is an abstract pseudo code for the Quiescence Search that extends the depth and checks if there is any capture for pieces after a specific move or not.

```
int Quiesce( int alpha, int beta ) {
   int stand_pat = Evaluate();
   if( stand_pat >= beta )
      return beta;
   if( alpha < stand_pat )
      alpha = stand_pat;

   until( every_capture_has_been_examined )  {
      MakeCapture();
      score = -Quiesce( -beta, -alpha );
      TakeBackMove();

      if( score >= beta )
         return beta;
      if( score > alpha )
        alpha = score;
   }
   return alpha;
}
```

Fig. 6. Abstract Pseudo Code Version for Quiescence Search

Forward Pruning technique completes the idea of the variable depth, where it cuts-off unpromising branches. However, this can lead to errors in the result. Many algorithms implemented the idea of this technique, including N-Best Selective Search, ProbCut and Multi-ProCut [16].

N-Best Selective Search only looks for the best N-best moves at each node. All other siblings for the N-best moves will automatically cutoff.

Both ProbCut Multi-ProCut algorithms use the result of shallow search to determine the possibility that a deeper search will change the value of alpha and beta or not.

ProbCut [16] algorithm uses the statistics' correlation techniques to cutoff branches, because it was discovered that there is a strong correlation between values obtained from different depth. The relation was described by Micheal Buro as follows:

$$V\_D = a * V\_D' + b + e \quad (2)$$

Where V_D means the value of a given depth, a & b are real numbers and e is a normally distributed error with zero mean.

Since the value of $a \approx 1$, $b \approx 0$ and σ^2 is small in most stable evaluation function, the probability of V_D >= beta could be predicted from the following equivalent equation:

$$V\_D' >= ((1/ \Phi(P))*\sigma + beta - b) / a \quad (3)$$

Furthermore, the probability of V_D <= alpha could predicted from the following equivalent equation:

$$V\_D' <= (-(1/\Phi(P)*\sigma + alpha - b) / a \quad (4)$$

In Fig. 7 there is an abstract implementation for the ProbCut algorithm. Remember it's up to you to choose the values of D, D', the cutoff threshold (1/Φ(P)), a, b and σ. The algorithm can provide a faster result than any brute-force algorithm. However, it needs many accurate parameters, which may be difficult to choose and may lead to errors in the results.

```
int alphaBetaProbCut(int α, int β, int depth) {
        const float T(1.5);
        const int DP(4);
        const int D(8);

        if ( depth == 0 ) return evaluate();

        if ( depth == D ) {
                int bound;

                // v >= β with prob. of at least p?
                // yes => cutoff */
                bound = round( ( T * σ + β - b) / a );
                if ( alphaBetaProbCut( bound-1, bound,
                                        DP) >= bound )
                        return β;

                // v <= α with prob. of at least p?
                // yes => cutoff */
                bound = round( (-T * σ + α - b) / a );
                if ( alphaBetaProbCut( bound, bound+1,
                                        DP) <= bound )
                        return α;
        }
        // the remainder of alpha-beta goes here
        ...
}
```

Fig. 7. Abstract Pseudo Code Version for ProbCut Search without the alpha-beta implementation

Multi-ProbCut [16] algorithms generalize the idea of ProbCut by using additional cutoff thresholds and checks, including allowing more regression parameters and cutoff thresholds, using many depth pair and using internal iterative deepening for shallow searches.

C. Breadth-First Search

As mentioned before the BFS begins from the root node then it visits its first child after that it visits all its siblings from the same depth before it moves to the next depth. One of the problems with this technique; it requires huge memory to store node's data. Many algorithms use this technique like NegaC* [17], MTD (f) [14], SSS* [4], B* [5] and Monte-Carlo search [18] algorithms, which discussed below.

NegaC* [17] algorithm uses the minimal-window with fail-soft Alpha-Beta algorithm like NegaScout [3] algorithm, but it parses the tree in Breadth-First way. The fail-soft technique uses two more variables than the Alpha-Beta algorithms to cutoff more branches.

In Fig. 8 there is an abstract pseudo code implementation of the NegaC* algorithm.

```
int negaCStar (int min, int max, int depth) {
        int score = min;
        while (min != max) {
                alpha = (min + max) / 2;
                score = failSoftAlphaBeta (alpha,
                                        alpha + 1, depth);
                if ( score > alpha)
                        min = score;
                else
                max = score;
        }
        return score;
}
```

Fig. 8. An Abstract Pseudo Code Implementation of the NegaC* Algorithm

MTD (f) [14] algorithm, which is an abbreviation for "Memory-enhanced Test Driver" also uses the minimal-window technique like NegaScout [3] algorithm, but it does it efficiently. It was introduced as an enhancement to the Alpha-Beta Algorithm as mentioned before. It also uses two more variables to determine the upper-bound and lower-bound. The normal Alpha-Beta algorithm uses only alpha and beta variables with -∞ & ∞ as a start and the values are updated one time at each call to make the only one returning value lies between the alpha and beta values. However, MTD (f) may search more than one time at each Alpha-Beta [12] call and use the returned bounds to converge toward it using the lower-bound and upper-bound to make faster cutoffs of the tree. Furthermore, the algorithm uses a transposition table to store and retrieve data about portions of the search tree to use it later to reduce the over-head of re-examining same game state. However, it uses memory space to store this data, which required more memory space.

Fig. 9 shows a pseudo code for the MTD (f) algorithm without the implementation of the Alpha-Beta algorithm which was described in Fig. 4.

SSS* [4] is another famous breadth-first search algorithm, which is non-directorial search algorithm. The algorithm expands into multiple paths at the same time to get global-information of the search tree. However, it searches fewer nodes than fixed depth-first algorithms like Alpha-Beta algorithm.

The algorithm stores' information for all active nodes which didn't solve yet in a list in decreasing order depends on their importance. The information consists of three parts:

- N: a unique identifier for each node.
- S: a status of each node whatever it's live or has been solved.
- H: an important value for the node.

```
function MTDF(root, f, d){
        g := f
        upperBound := +∞
        lowerBound := -∞
        while lowerBound < upperBound{
                if g = lowerBound then
                        β := g+1
                else
                        β := g
                g := AlphaBetaWithMemory (root,  β-1,
                                                 β, d)
                if g < β then
                        upperBound := g
                else
                        lowerBound := g
        }
        return g
}
```

Fig. 9. Pseudo Code for the MTD (f) Algorithm Without the Implementation of the Alpha-Beta algorithm Which was Described Previously in Fig. 4

The core of the SSS* [4] algorithm depends on two phases:

- Node Expansion: Top-down expansion of a Min strategy.
- Solution: Bottom-up search for the best Max strategy.

Fig. 10 shows the pseudo code for the SSS* algorithm using three function push, pop and insert to store, remove and update node information.

Monte-Carlo Tree Search "MCTS" [18] algorithm made a breakthrough in game theory and computer science field. The algorithm is based on randomized exploration of the game tree. The algorithm also uses the results of previous examined values for nodes. Every time the algorithm runs it produces a better estimation of values. However, the game tree gradually grows in the memory, which is the main disadvantages of breadth-first algorithms.

The algorithm consists of four phases, which is repeated as long as there is still time for the computer to think:

- Selection phase, it starts from the root node; it traverses the game tree by selecting the most promising move until reaching a leaf node.
- Expansion phase, if the number of visits reaches a pre-determined threshold, the leaf is expanded to build a larger tree.
- Simulation phase, calculates the outcome value of the leaf by performing a play-out at it.
- Back-propagation phase, it traces back along the game tree path from the leaf to the root to update the values changed in the simulation phase.

```
int SSS* (node n; int bound)
{
  push (n, LIVE, bound);
  while ( true ) {
    pop (node);
    switch ( node.status ) {
    case LIVE:
      if (node == LEAF)
        insert (node, SOLVED, min(eval(node),h));
      if (node == MIN_NODE)
        push (node.1, LIVE, h);
      if (node == MAX_NODE)
        for (j=w; j; j--)
          push (node.j, LIVE, h);
      break;
    case SOLVED:
      if (node == ROOT_NODE)
        return (h);
      if (node == MIN_NODE) {
        purge (parent(node));
        push (parent(node), SOLVED, h);
      }
      if (node == MAX_NODE) {
        if (node has an unexamined brother)
          push (brother(node), LIVE, h);
        else
          push (parent(node), SOLVED, h);
      }
      break;
    }
  }
}
```

Fig. 10. Pseudo Code for the SSS* Algorithm

Fig. 11 shows the pseudo-code for MCTS algorithm using the four phases described before.

B* [5] is the final algorithm that will be described in the BFS. It finds the least-cost path from the node to any goal node out of one of more possible goals. The main idea of this algorithm is based on:

- Stop when one path is better than all the others.
- Focus the exploration on paths that will lead to stopping.

The algorithm expands the searching based on prove-best and disprove-rest strategies. In prove-best strategy, the algorithm chooses the node with the highest upper-bound because it has a high probability to raise its lower bound higher than any other nodes' upper-bound when it expands. On the other hand, the disprove-rest strategy chooses the next highest upper-bound node because it has a good probability to reduce the upper-bound to less than the lower-bound of the best child when it expands.

```
Data: root node
Result: best move
while (has time) do
   current node ← root node

   /* The tree is traversed
   while (current node ϵ T ) do
      last node ← current node
      current node ← Select(current node)
   end

   /* A node is added
   last node ← Expand(last node)

   /* A simulated game is played
   R ← P lay simulated game(last node)

   /* The result is backpropagated
   current node ← last node
   while (current node ϵ T ) do
      Backpropagation(current node, R)
      current node ← current node.parent
   end

end

return best move = argmaxN ϵ Nc (root node)
```

Fig. 11. Pseudo-Code for MCTS Algorithm

In the next section, a discussion of most famous parallel game tree search algorithms is presented.

III. Parallelism in Game Tree Search

The technological advances of the computer architecture and the release of multiprocessors and multi-core computers, allows algorithms that can be partitioned into independent segments to be solved faster. Many enhancements were done on the sequential algorithms to make it run in parallel as well as new algorithms are designed for parallel computing. The problem of parallel computing is the trade-off between the overhead of communication and synchronization and the benefits of exploring many nodes in the same time in parallel. This made the speedup is sub-linear rather than linear. Section [A], presents a discussion of various techniques & algorithms made to solve these problems. Section [0], presents a discussion for the parallelism of the game search tree from the hardware point of view.

A. Game Tree Parallelism Techniques & Algorithms

As mentioned before many techniques were designed to solve the overhead problem. One of them is the "Shared Hash Table" technique, which stores information about nodes in the game tree so it could be used by any processor or core in the system. This reduces the communication over-head between processors or cores, especially if the processors are not on the same physical computer.

Many algorithms use the previous technique as well as other techniques, including ABDADA, Parallel Alpha-Beta, Parallel PVS, YBWC and Jamboree and Dynamic Tree Splitting. Next, a description of each algorithm is presented.

ABDADA is a loosely synchronized and distributed search algorithm designed by Jean-Christophe. The algorithm uses the shared hash table technique as well as adding more information for each node like the number of processors searching this node.

Parallel Alpha-Beta [6] is the parallel version of the previously discussed Alpha-Beta algorithm. The basic idea is splitting the search tree into sub-search trees and run each one in specific core or more in case of multi-core and one or more processor in case of multi-processor system. The problem of this algorithm is the complexity of implementing it. However, it can maximize the utilization of the cores or processors. The two methods of splitting the tree are showing in Fig. 12.

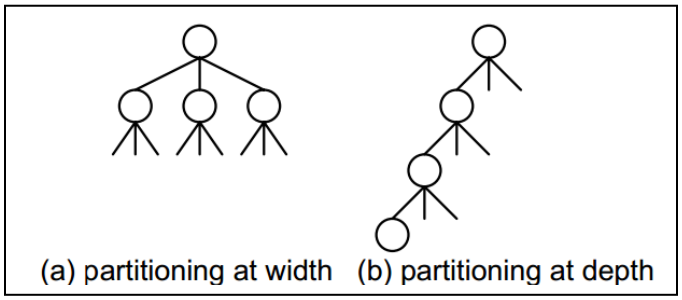


Fig. 12. Partitioning of the Game Tree for Alpha-Beta Search

PVS [13] algorithm expresses each node as a thread which can run in parallel. However, before running them in parallel, the problem of data dependency that exists among threads must be solved. A simple solution is to get the initial required value from the first node among any siblings then run the remaining siblings in parallel. The sequential and parallel tasks for PVS algorithm using two processors is showing in Fig. 13 while a pseudo code for the algorithm is showing in Fig. 14.

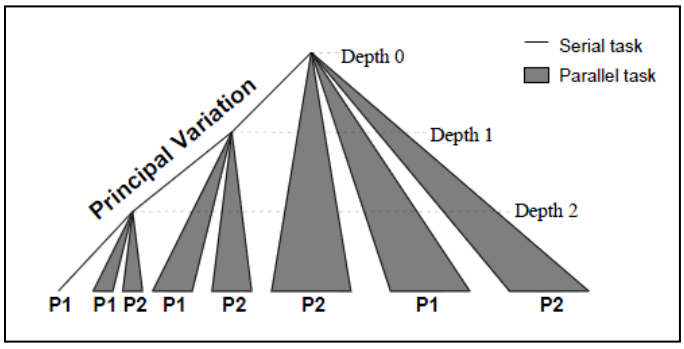


Fig. 13. Sequential and Parallel Tasks by Two Processors using PVS

YBWC and Jamboree algorithm [19] is shown in ,which based on a recursive algorithm that visits the first node in siblings before spawning the remaining sibling's nodes in parallel. It uses this technique because the first node may produce a cutoff so it does not waste the others processors' time by searching in the sibling nodes or will produce better

bounds, then the algorithm search the remaining siblings in parallel. A pseudo code of the algorithm is showing in Fig. 15.

```
PVSplit (Node curnode, int alpha, int beta, int result)
{
        if(cur_node.is_leaf)
                return Evaluate(cur_node);

        succ_node = GetFirstSucc(cur_node);
        score = PVSplit(curnode, alpha, beta);

        if(score > beta)
                return beta;
        if(score > alpha)
                alpha = score;

        //Begin parallel loop
        while(HasMoreSuccessors(curnode))
        {
                succ_node = GetNextSucc(curnode);
                score = AlphaBeta(succnode,alpha,beta);

                if(score > beta)
                        return beta;
                if(score > alpha)
                        alpha = score;
        }//End parallel loop
        return alpha;
}
```

Fig. 14. Pseudo Code for PVS Algorithm Using Alpha-Beta Function

Finally, DTS algorithm [20] is the most complex parallel game tree search algorithm and there are a few implantation of it. However, it gives the best performance in symmetric multi-processors systems. A pseudo-code of the DTS algorithm presented in Fig. 16.

Table I shows the speedup for the three algorithms compared using 1, 2, 4, 8, and 16 processors.

TABLE I. POPULAR GAME TREE PARALLEL ALGORITHMS SPEEDUP

Algorithm	**Number of Processors**				
	1	***2***	***4***	***8***	***16***
PVS	1.0	1.8	3.0	4.1	4.6
YBWC	1.0	1.9	3.4	6.1	10.9
DTS	1.0	2.0	3.7	6.6	11.1

All the previous parallel algorithms need a parallel programming language or library that can handle threads and parallel computing. Many libraries and programming languages were released to support CPU parallelism in general. However, the most famous library that used in parallel game tree algorithms is MPI (Message Passing Interface) [21] which is an extension of C programming language. MPI handles the burden of synchronization, communication and distributed resources management. The latest version of MPI is version 2 which supports C++ and object-oriented programming.

```
jamboree(CNode n, int α, int β, int b)
{
        if (n is leaf)
                return static_eval(n);

        c[] = the childen of n;
        b = -jamboree(c[0], -β, -α);
        if (b >= β) return b;
        if (b >  α) α = b;

        In Parallel: for (i=1; i < |c[]|; i++)
        {
                s = -jamboree(c[i], -α - 1, -α);

                if (s >  b)
                        b = s;
                if (s >= β
                        abort_and_return s;
                if (s >  α)
                {
                //Wait for the completion of previous
                iterations of the parallel loop
                        s = -jamboree(c[i], -β, -α);

                        if (s >= β)
                                abort_and_return s;
                        if (s >  α)
                                α = s;
                        if (s >  b)
                                b = s;
                }
        }
  return b;
}
```

Fig. 15. Pseudo Code for Jamboree Algorithm

```
DTS(root)
{
        while (Stopping_criterion() == false)
        {
                //One processor search to ply = N
                SearchRoot(root);

                //Detect free processors, and begin tree split
                Split(node v);

                //Initialize new threads.
                ThreadInit();

                //Copy a "split block" to begin a new search
                CopytoSMP(node v);
                SearchSMP(node v);
        }
        ThreadStop();
}
```

Fig. 16. Pseudo Code for DTS Algorithm

Another trending library for artificial intelligence algorithms that runs on CPU is Microsoft Task Parallel Library (TPL) [22]. This library uses the concept of finite CPU-bound computation based on task notation and the concept of replicating task using work-stealing technique. It is more effective to develop parallel algorithms such as DTS and YBWC.

Other programming libraries were designed to run parallel algorithms in GPU rather than CPU, including CUDA [23]. However, few algorithms were implemented using these libraries because of the complexity of programming search trees using these libraries. On the other hand, the implemented algorithms that tested on GPU showed a better speedup than the CPU.

B. CPU & GPU for Parallel Game Tree Search

In the previous fifteen years, all researches focused on designing parallel algorithms that can run in parallel on multi-cores or multi-processors. However, the new trend of search trees field is to design and implement algorithms that can run on parallel on the GPU. Early the GPU was built just for graphic computing.

However, in the last 10 years the GPU became a platform for general parallel processing computing. The idea of GPU is to have hundreds or thousands of simple cores that can run threads in parallel with higher GFLOPS than the CPU. On the other hand, the CPU contains few powerful cores or few powerful multi-processors that can run more instructions and have a faster clock speed than the GPU. As mentioned before few algorithms were modified to support GPU. However, in the next five years many AI algorithms will designed to use the power of GPUs. Fig. 17 shows the CPU and GPU architecture.

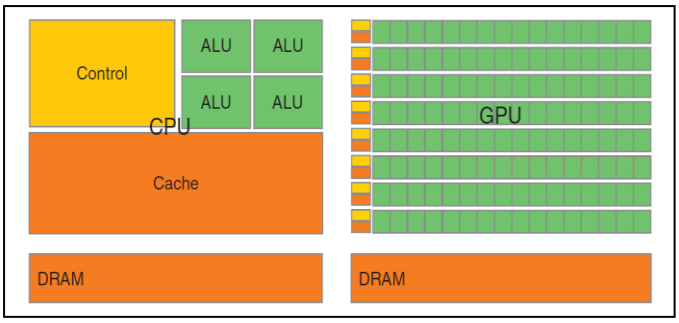


Fig. 17. Difference Between CPU and GPU Architecture

In the next section, an analysis of the previous algorithms based on several criteria is presented.

IV. GAME TREE ALGORITHMS ANALYSIS

As the game tree was categorized according to specific criteria, the evaluation of the previous algorithms is categorized according to specific criteria ; which are:

- Completeness: whatever if the algorithm finds a solution if one exists.
- Time Complexity: the number of nodes generated.
- Space Complexity: the maximum number of nodes in memory during the search.
- Optimality: whatever if the algorithm always finds the least-cost or the best solution.

The following terms were used to measure the time and space complexity:

- B: maximum branches factor.
- D: depth of the best solution.
- M: maximum depth of the state space.
- L: depth cut-off

Table II compares the various sequential algorithm categories based on the previous criteria.

TABLE II. SEQUENTIAL ALGORITHMS ANALYSIS

Criterion	*Breadth-First*	*Depth-First*	*Depth-Limited*	*Iterative Depending*
Completeness	Yes	No	Yes, if l >= d	Yes
Time	b^d	b^m	b^l	b^d
Space	b^d	bm	bl	bd
Optimality	Maybe	No	No	Maybe

All algorithms that fall into any of the categories should match the mentioned table criteria value. However, this is the worst-case scenario. In most cases, a better space and time values may be found. An example of that is the Alpha-Beta algorithm where the average time is equal $b^{3m/4}$.

The complexity analysis of parallel game tree searching algorithms is more difficult than the sequential algorithms. Usually, the parallel game tree searching algorithms could be analyzed in terms of:

- Time complexity T (n): How many times steps are needed?
- Processor complexity P (n): How many processors are used?
- Work complexity W (n): What is the total work done by all the processors?

The sequential Minimax algorithm with alpha-beta modification has a time complexity of $O(b^m)$ in the worst case, $O(b^{3m/4})$ in the average case and $O(b^{m/2})$ in the best case; where b is the branching factor and m is the maximum depth of the search tree. The total work done by the parallel alpha-beta algorithm equals the total work done by the sequential algorithm. Hence, table III summarizes the time complexity of the parallel alpha-beta algorithm in the best, average and worst case.By increasing the number of processors or cores, the algorithm could provide a better speed-up. However, it will never reach the ideal speedup, because of the communication overhead between nodes for sharing alpha and beta values as well as the synchronization overhead.

TABLE III. PARALLEL ALGORITHMS ANALYSIS

Criterion	*Case*	*Complexity*
Time Complexity	worst	$T(n) = O\left(\frac{b^m}{p}\right)$
	average	$T(n) = O\left(\frac{b^{3m/4}}{p}\right)$
	best	$T(n) = O\left(\frac{b^{m/2}}{p}\right)$
Processor Compleixty	worst	$P(n) = O(n)$
	average	
	best	
Work Compleixty	worst	$W(n) = O(b^m)$
	average	$W(n) = O(b^{3m/4})$
	best	$W(n) = O(b^{m/2})$

Besides the normal criteria that will make you choice any game searching algorithms; which includes the amount of memory you have, if you need an optimal solution or not, and the time you need to solve, etc.. An important criterion is the nature of the problem or the nature of the game and the nature of the hardware architecture you have.

V. CONCLUSIONS & FUTURE WORK

A discussion of various game tree-searching algorithms was presented in this paper, including sequential and parallel algorithms. The popular sequential algorithms were covered in details the common algorithms in both depth-first and breadth-first. Furthermore, an overview of common parallel algorithms as well as the hardware architecture for parallel game tree searching was presented. In the end, an analysis the algorithms based on four criteria was discussed.

The use of service-oriented approach to expand the searching trees into a distributed system will solve many distributed issues, while using tasks technology rather than threads to implement the parallel algorithms to maximize the utilization of multi-core computers. Another suggestion is to implement the search algorithms using the OpenCL library which allows the code to run on both GPU and CPU, or CUDA library to produce a massive parallel game tree searching algorithm.

The new feature of dynamic parallelism in CUDA v5.5 allows recursion based algorithms to run faster on the GPU, by eliminating the CPU initialization time of each kernel. Furthermore, the unified memory in CUDA 6.0 creates a pool of management memory that is shared between the CPU and GPU, which make the development of complex games easier. Both the dynamic parallelism and unified memory features can improve the speedup of current AI game tree searching algorithms.

REFERENCES

[1] S. Russell and P. Norvig, Artificial intelligence: a modern approach, 3rd ed. Prentice Hall Press, 2009, p. 1152.

[2] G. T. Heineman, G. Pollice, and S. Selkow, “Path Finding in AI,” in Algorithms in a Nutshell, 1st ed., O’Reilly Media, 2008, pp. 213–217.

[3] H.-J. Chang, M.-T. Tsai, and T. Hsu, “Game Tree Search with Adaptive Resolution,” in Advances in Computer Games SE - 26, vol. 7168, H. J. Herik and A. Plaat, Eds. Springer Berlin Heidelberg, 2012, pp. 306–319.

[4] [4] U. Lorenz and T. Tscheuschner, “Player Modeling, Search Algorithms and Strategies in Multi-player Games,” in Advances in Computer Games SE - 16, vol. 4250, H. J. Herik, S.-C. Hsu, T. Hsu, and H. H. L. M. (Jeroen. Donkers, Eds. Springer Berlin Heidelberg, 2006, pp. 210–224.

[5] Y. Tsuruoka, D. Yokoyama, and T. Chikayama, “Game-Tree Search Algorithm Based on Realization,” ICGA J., vol. 25, no. 3, pp. 145–152, 2002.

[6] V. Manohararajah, “Parallel Alpha-Beta Search on Shared Memory Multiprocessor,” Computer Engineering University of Toronto, 2001.

[7] D. Strnad and N. Guid, “Parallel alpha-beta algorithm on the GPU,” CIT, vol. 19, no. 4, pp. 269–274, 2011.

[8] J. Habgood and M. Overmars, “Clever Computers: Playing Tic-Tac-Toe,” in The Game Maker’s Apprentice SE - 13, Apress, 2006, pp. 245–257.

[9] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, “Depth-first search,” in Introduction to Algorithms, 3rd ed., MIT Press, 2009, pp. 603–612.

[10] W. Ertel, “Search, Games and Problem Solving,” in Introduction to Artificial Intelligence SE - 6, Springer London, 2011, pp. 83–111.

[11] M. Schadd, “Selective Search in Games of Different Complexity,” Maastricht University, 2011.

[12] D. Knuth, Selected Papers on Analysis of Algorithms. California: Center for the Study of Language and Information, 2000.

[13] M. H. M. Winands, H. J. van den Herik, J. W. H. M. Uiterwijk, and E. C. D. van der Werf, “Enhanced forward pruning,” Inf. Sci. (Ny)., vol. 175, no. 4, pp. 315–329, 2005.

[14] K. Shibahara, N. Inui, and Y. Kotani, “Adaptive Strategies of MTD-f for Actual Games,” in CIG, 2005.

[15] M. Schadd and M. Winands, “Quiescence Search for Stratego,” in BNAIC, 2009, pp. 225–232.

[16] A. X. Jiang and M. Buro, “First Experimental Results of ProbCut Applied to Chess,” Adv. Comput. Games, vol. 10, 2003.

[17] D. Rutko, “Fuzzified Tree Search in Real Domain Games,” in Advances in Artificial Intelligence SE - 13, vol. 7094, I. Batyrshin and G. Sidorov, Eds. Springer Berlin Heidelberg, 2011, pp. 149–161.

[18] J. Hashimoto, A. Kishimoto, K. Yoshizoe, and K. Ikeda, “Accelerated UCT and Its Application to Two-Player Games,” Adv. Comput. Games, 2011.

[19] [19] J. Steenhuisen, “Transposition-Driven Scheduling in Parallel Two-Player State-Space Search,” Delft University of Technology, 2005.

[20] T. A. N. Ying, L. U. O. Ke-lu, C. Yu-rong, and Z. Yi-min, “Performance Characterization of Parallel Game-tree Search Application Crafty,” vol. 4, no. 2, pp. 2–7, 2006.

[21] D. Jakimovska, G. Jakimovski, A. Tentov, and D. Bojchev, “Performance estimation of parallel processing techniques on various platforms,” in Telecommunications Forum (TELFOR), 2012 20th, 2012, pp. 1409–1412.

[22] D. Leijen, W. Schulte, and S. Burckhardt, “The design of a task parallel library,” Acm Sigplan Not., vol. 44, no. 10, pp. 227–242, Oct. 2009.

[23] D. B. Kirk and W. M. W. Hwu, Programming Massively Parallel Processors: A Hands-on Approach. Morgan Kaufmann, 2012, p. 496.

The Effect of Diversity Implementation on Precision in Multicriteria Collaborative Filtering

Wiranto
Informatics Department
Sebelas Maret University
Surakarta, Indonesia

Edi Winarko
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Sri Hartati
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Retantyo Wardoyo
Department of Computer Science and Electronics
Gadjah Mada University
Yogyakarta, Indonesia

Abstract **—This research was triggered by the criticism on the emergence of homogeneity in recommendation within the collaborative filtering based recommender systems that put similarity as the main principle in the algorithm. To overcome the problem of homogeneity, this study proposes a novelty, i.e. the diversity of recommendations applied to the multicriteria collaborative filtering-based document recommender systems. Development of the diversity recommendation was made by the two techniques, the first is to compare the similarity of content and the second is to use a variation of the criteria. The application of diversity, both content and criteria-based, was proven to provide a sufficiently significant influence on the increase of recommendation precision.**

Keywords—Algorithms; multicriteria; content; collaborative; filtering; systems; similarity; diversity; precision

I. INTRODUCTION

The development of collaborative filtering based recommender systems always puts the aspect of similarity as the main reference in the algorithm, and the main parameter used to assess the performance is the accuracy of prediction. Therefore, most studies on the recommender systems are focused on improving the accuracy of predictions, including when developing a multicriteria collaborative filtering model [1] [2] [3]. The implication of similarity implementation is the resultant recommendation that is homogeneous in nature. It is the advantage of collaborative filtering approach, but on the other hand it can also be disadvantage. The homogeneity of recommendation is due to process in collaborative filtering algorithm that does not involve the description or the content of recommendation object, so that the system does not accommodate the existence of new items [4]. The adverse effect is the case where many objects whose content is very interesting for the users, but it were never promoted to be a part of a list of recommendations.

Based on this fact, it is necessary to conduct a study with a focus on the development of recommendation diversity, but it remains in the corridor of multicriteria collaborative filtering.

Recommendation diversity is very important to be taken into account because it is closely related to the level of user satisfaction. In fact, it is not only important in recommender systems, but is also very important in developing a model of information retrieval systems and social media with very rapid development. The ideas of diversity developed in this study are of two kinds, content-based diversity and criteria-based diversity, which were done on multicriteria collaborative filtering model.

Besides prediction accuracy, there is other parameter used to measure the performance of recommender systems. The parameter is recommendation precision, which is defined as a number that indicates the percentage of items that were given a high predictive value by recommender systems, as well as by users. In this study, how much the influence of the implementation of recommendation diversity on the increase of precision in the multicriteria collaborative filtering applied to construct a scientific document recommender systems will be measured.

The writing of paper is organized as follows. Section 2 describes the content-based diversity. Section 3 describes the criteria-based diversity. The testing of recommendation precision is presented in section 4, while the discussion of the results of test is written in section 5. The writing of the paper is concluded by section 6.

II. CONTENT BASED DIVERSITY

The objects of recommendation in this study are scientific documents with text format, making it possible to do an analysis of its content. The results of the analysis of a document can be compared with other document contents. The results of the comparison of the document contents generate similarity values that are then used as the basis of determining which documents that need to be recommended to the users, with the guideline that the higher the content similarity, the lower the document diversity [6][7]. The scenario for determining the diversity based on the document content can be explained as follows:

a) Documents whose contents are analyzed are those already included into the list of Top-N produced by a multicriteria collaborative filtering engine.

b) Content analysis is sufficiently done on document abstract.

c) Content analytic process is meant to find out or measure the similarity.

d) One of the documents with high enough similarity values are chosen to be included into the list of the recommendation.

The analytic process of document content is done in two steps, i.e. indexing process and similarity measurement.

B. Document Indexing

Document index is a set of terms representing the content. Each document is represented with bag-of-words. The process is started by transforming the document into a bag containing independent words. Each word is stored into a database that is arranged as an inverted index. The arrangement of inverted index required the involvement of linguistic processing with aim of extracting important terms by deleting stop-words and stemming. The definition of stop-words is 'words that have no relevance with main subject, although the words often appear in many documents'. The example of stop-words include a, an, all, also, after, although, because, beside, every, the, this, it, these, those, his, her, my, our, their, your, few, many, several, some, for, and, nor, bit, or, yet, so, if, unless, on, off, over, of, during and etc. Meanwhile, stemming is an operation to gain a form of the roots of word by deleting the prefix or suffix. By the technique, a group of suitable words, where words in the group are variants, will be gained. As an example, the words *write, written, writer, writing* are interchangeably used in term with the general stem of *write*.

Forming the inverted index requires five steps, i.e.:

a) The deletion of format and document markup with many tags and formats such as HTML document.

b) Tokenization. The words in sentence, paragraph or pages are separated into token or pieces of a single word or stemmed word. Being included into the step is to delete certain characters such as punctuation mark and to change all the tokens into lower case.

c) Filtering, i.e. to determine which terms will be used to represent document in order to describe the document content and distinguish them with other documents. The terms with the very high level of frequency in appearance cannot be used for the purpose because they are unable to be discriminator inter-documents or often called by a term of poor discriminator. Moreover, terms often appearing in many documents do not also reflect the definition of the topic of sub-topic of documents. Therefore, the terms often used can be considered as stop-word and must be deleted. In order that the process of stop-word deletion goes fast, a book of stop-word or the stop-list of term that will be deleted must be arranged.

d) The retrieval of term into a form of root. Document can be expanded by searching synonymous for certain term within it. Synonym is words that have similar meaning but morphologically seem different. The step is similar with stemming process, but what want to find is a group of relevant words. The main difference is that synonym does not share in use of term, but found based on thesaurus.

e) The weighting of term. To do the weighting, it can be selected local or global weighting model or the combination of both. The model often used in several applications is a combined weighting by the multiplication of the local weight of term frequency and the global inverse document frequency, written by tf.idf. [8]

C. The Measurement of Content Similarity

To measure the similarity of text formatted document, the bag-of-words need to be converted first into the vector space model with each document represented as a multidimensional vector with dimensions in accordance with the chopped term in the database. Figure 1 shows an example of visualization of three-dimensional vector space models with the terms of T_1, T_2, and T_3 as well as two documents of D_1 and D_2.

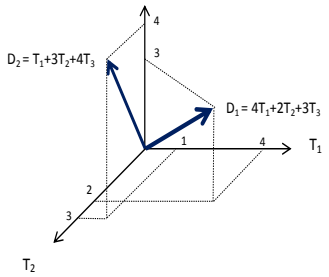


Fig. 1. An Example of Three-Dimensional Vector Space Model

The database of documents is represented by the matrix of term-document or term-frequency where each cell match with the weight given. The value of zero shows that the term does not appear in the document. Figure 2 is an example of term-document matrix for the database containing n document and t term.

	T_1	T_2	.	.	.	T_t
D_1	w_{11}	w_{21}	.	.	.	w_{t1}
D_2	w_{12}	w_{22}	.	.	.	w_{t2}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
D_n	w_{1n}	w_{2n}	.	.	.	w_{tn}

Fig. 2. The Example of Term-Document Matrix

Based on the term-document matrix formed and the weighting of tf-idf, the numeric value of the document can be known, thus the inter-document nearness can be calculated. The nearer the two vectors are, the more similar the two dua documents. The similarity of text content document can be calculated by using a cosine similarity formula. For example, two vectors representing documents d_j and d_k were given, so

the content similarity between both documents was defined as [9].

$$sim(d_j, d_k) = \frac{\sum_{i=1}^{t}(w_{ij}.w_{ik})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 . \sum_{i=1}^{t} w_{ik}^2}}$$

To make understanding easier, the example of 3 documents was given by vector representation as follows :

$D_1 = 2T_1 + 5T_2 + 6T_3$
$D_2 = 5T_1 + 3T_2 + 4T_3$
$D_3 = 4T_1 + 5T_2 + 5T_3$

Thus, the inter document similarity value can be calculated as follows:

(1) similarity D_1 and D_2 :

$$sim(D_1, D_2) = \frac{(2*5+5*3+6*4)}{\sqrt{(4+25+36).(25+9+16)}} = \frac{49}{\sqrt{(65)*(50)}} = 0.86$$

(2) similarity D_1 and D_3 :

$$sim(D_1, D_3) = \frac{(2*4+5*5+6*5)}{\sqrt{(4+25+36)*(16+25+25)}} = \frac{63}{\sqrt{(65)*(66)}} = 0.96$$

(3) similarity D_2 and D_3 :

$$sim(D_2, D_3) = \frac{(5*4+3*5+4*5)}{\sqrt{(25+9+16)*(16+25+25)}} = \frac{55}{\sqrt{(50).(66)}} = 0.95$$

From the three values of document similarity above, it can be known that document D_3 had similarity with the other two documents. The smallest similarity value was gained between document D_1 and document D_2, so the document prioritized to recommend was D_1.

III. Criteria Based Diversity

Referring to a construction of document recommendation system by using multicriteria collaborative filtering, actually a space is available to engineer at the step of recommendation generation to make sure the presence of diversity [2] [10]. The new concept of diversity sufficiently bases at four individual criteria that was determined and used since earlier, different from the concept of document content-based diversity whose process was long enough and need the step of indexing. The construction of document recommendation system by using multicriteria collaborative filtering whose recommendation generation takes criteria-based diversity into account is shown in Figure 3.

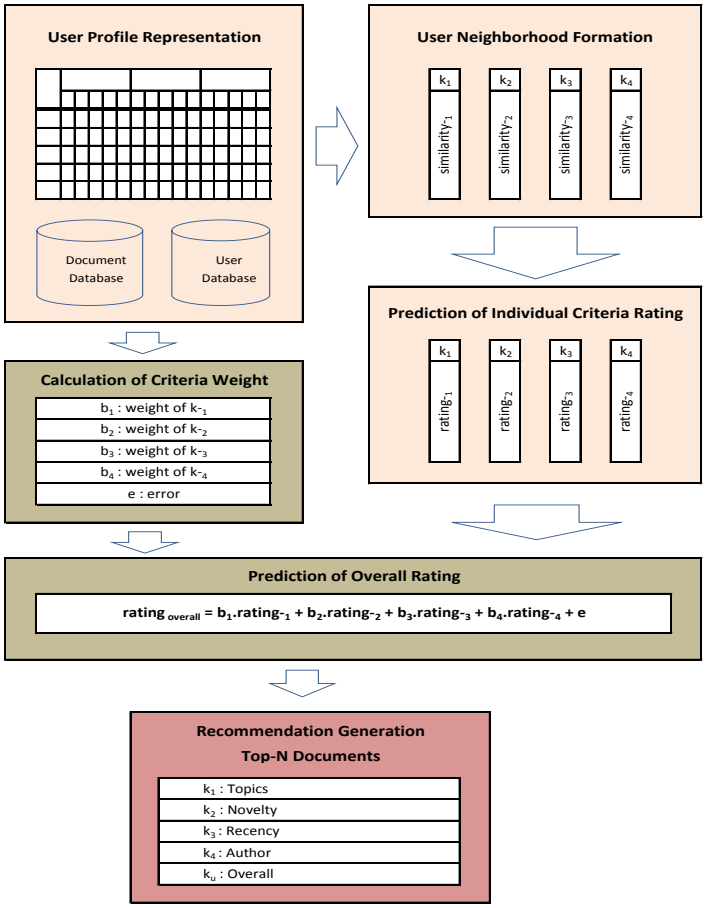


Fig. 3. The Construction of Multicriteria Collaborative Filtering (MCF) Recommender Systems Model Applying the Criteria-Based Diversity.

Figure 3 shows that the scenario run was still in collaborative filtering paradigm with four individual document criteria and one overall criterion. Only the modification of document selection procedures and its generation process are required, so that the recommended documents are more various.

In the Multicriteria Collaborative Filtering, prediction process of ratings is done for each criterion [11]. So, by using the four individual criteria and one overall criterion, actually five values of prediction results were generated and each can be used to generate recommendation. For each criteria (topic, novelty, recency, author and overall), a number of document Top-N with the highest predictive value was generated. After the step, one document was taken respectively and put into a list of document recommended to the users. Thus, there will be five document variations recommended based on different criteria, although it is still possible for the emergence of the same document.

IV. EXPERIMENT

Initially the precision is only used in information retrieval systems and has always been associated with another metric called as recall that is defined as follows: [5]

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}|}{|\{\text{documents retrieved}\}|}$$

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}|}{|\{\text{relevant documents}\}|}$$

For measuring the precision in this study, the term is modified to be "Precision in Top-N" that is defined as a percentage of documents with the high production value, becoming the most relevant N document for the users.

In the testing, rating value will be categorized as high if the value was larger or equal to 4.0. The measuring of precision was done when the number of the users and documents reached 200x400, while the rating value used was the value for the overall criteria. The variation of the testing was based on the neighborhood size by determining a number of the users with the highest similarity value. In the testing, three neighborhood sizes were selected, including 5 users, 10 users, and 50 users. At the last option, the large neighborhood size led the meaning of nearest-neighborhood was bias and the load of computation become large also, but the measuring under the condition still need to do for the performance of system. Meanwhile, the Top-N values used were 5, 10 and 15.

The testing was done under two different conditions. The first was when the system did not apply the recommendation diversity yet, and the second was after the system applied the diversity. There were three variations in recommender systems run in the testing, namely : classic collaborative filtering (CF), multicriteria collaborative filtering using cosinus-based similarity (MCF Cosinus) and multicriteria collaborative filtering using multidimensional distance-based similarity (MCF MD Distance). Result of the testing of precision before the implementation of diversity was presented in Table 1.

From Table 1, it can be indicated that the larger the neighborhood size used, the higher the precision given by all of the recommendation systems models.

TABLE I. PRECISION WITHOUT DIVERSITY IMPLEMENTATION

NO	NEIGHBORHOOD SIZE	MODEL	PRECISION (%)		
			Top-5	Top-10	Top-15
1	5 Users	Collaborative Filtering	65.0	65.0	67.4
		MCF Cosinus	68.2	68.4	68.9
		MCF MD Distance	70.1	71.8	73.3
2	10 Users	Collaborative Filtering	66.9	66.7	68.5
		MCF Cosinus	68.8	68.9	69.3
		MCF MD Distance	70.8	71.6	74.4
3	50 Users	Collaborative Filtering	70.8	69.8	70.1
		MCF Cosinus	71.8	69.9	70.2
		MCF MD Distance	71.9	75.1	76.4

In computation perspective, it can be concluded that the more the members accommodated in collaborative process, the more relevant and appropriate the resultant recommendations for meeting the need of the users. Observation on the size of Top-N document determined also resulted in the same information, where the larger the size of Top-N, the larger the precision value had by all the models. The highest value of recommendation precision was reached by MCF MD Distance at the neighborhood size of 50 users and with Top-N of 15 documents, i.e. 76.4%.

Results of the measuring of recommendation precision after the content-based diversity applied were presented in Table 2. Actually, the precision value increased significantly. It can be clearly seen from comparison between them with results of the measurement of recommendation precision when the diversity was not applied as presented in Table 1.

TABLE II. PRECISION WITH CONTENT BASED DIVERSITY

NO	NEIGHBORHOOD SIZE	MODEL	PRECISION (%)		
			Top-5	Top-10	Top-15
1	5 Users	Collaborative Filtering	67.0	67.8	69.5
		MCF Cosinus	69.2	69.9	70.4
		MCF MD Distance	72.1	74.8	75.3
2	10 Users	Collaborative Filtering	68.0	68.8	70.5
		MCF Cosinus	70.1	71.9	72.7
		MCF MD Distance	72.1	74.8	75.3
3	50 Users	Collaborative Filtering	71.8	72.8	73.1
		MCF Cosinus	72.2	73.9	74.2
		MCF MD Distance	76.9	77.1	77.8

The concept of content-based diversity can be applied in all of the recommendation system models, while the concept of criteria-based diversity can only be applied in multicriteria collaborative filtering model with a process scheme simply illustrated by Figure 3. The results of the testing on the effect of criteria-based diversity concept application on the increase

of recommendation precision also give sufficiently positive information. It can be seen at the results of recommendation precision measurement as presented in Table 3.

The results of the testing increasingly affirm that the larger the neighborhood size used, the higher the precision given by all recommendation system models. Moreover, the larger the Top-N values selected, the larger the precision value had by all models. The highest value of recommendation precision was reached by MCF MD Distance at the neighborhood size of 50 users and the Top-N of 15 documents, i.e. 77.5 %.

TABLE III. PRECISION WITH CRITERIA BASE DIVERSITY

NO	NEIGHBORHOOD SIZE	MODEL	PRECISION (%)		
			Top-5	Top-10	Top-15
1	5 Users	MCF Cosinus	68.2	70.9	70.8
		MCF MD Distance	72.0	73.8	75.4
2	10 Users	MCF Cosinus	70.2	72.5	72.6
		MCF MD Distance	72.6	74.6	76.2
3	50 Users	MCF Cosinus	72.4	74.1	74.4
		MCF MD Distance	76.8	77.3	77.5

V. DISCUSSIONS

The idea of recommendation diversity was generated with the aim to provide the added value, making it possible for the users to get documents that are more relevant to their needs. It can be expected that after getting the relevant documents, the users will be satisfied and give the high value of rating on the documents. This was consistent with the theory of consumer behavior, explaining that when a person feels satisfied and so happy with the service, it will provide a high and sustained appreciation. The more the documents are given the high value of rating by the users, the more the increase of recommendation precision. For the reason, in this testing, the measurement of recommendation precision as recommendations are generated involved content- and criteria-based document diversity.

The two concepts of diversity give a special feature in the process of generating recommendations, so that there is diversity within uniformity. The higher the level of the content similarity, the lower the level of document diversity. The main implication of the application of content diversity was that among the documents with high rating, some documents with relatively different contents are selected. For the criteria-based diversity, it is sufficiently determined based criteria variation. It means that among the documents with high rating, several documents with different criteria are selected.

VI. CONCLUSIONS

Based on the results of the measuring of recommendation precision, it can be concluded that the application of diversity in multicriteria collaborative filtering-based recommendation document system had a positive effect, namely, to increase the recommendation precision. It can be interpreted that basically the users want various recommendations, although generated by a system built on the collaborative filtering concept based on the principle of similarity. The results of the study indicate that each effort to develop the recommender systems should accommodate the idea of diversity in order to produce a kind of recommendation that is more relevant and able to meet the subjective needs of the users. Thus, the principle of similarity in the collaborative filtering can be enriched by the feature of diversity.

REFERENCES

[1] A. Umyarov and A. Tuzhilin, "Improving Rating Estimation in Recommender Systems Using Aggregation- and Variance-based Hierarchical Models," in *Proceedings of the Third ACM Conference on Recommender Systems*, New York, NY, USA, 2009, pp. 37–44.

[2] Wiranto, E. Winarko, S. Hartati, and R. Wardoyo, "Improving The Prediction Accuracy of Multicriteria Collaborative Filtering by Combination Algorithms," *International Journal of Advanced. Computer Science and Application.*, vol. 5, no. 4, May 2014.

[3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, New York, NY, USA, 2001, pp. 285–295.

[4] M. Zhang and N. Hurley, "Avoiding Monotony: Improving the Diversity of Recommendation Lists," in *Proceedings of the 2008 ACM Conference on Recommender Systems*, New York, NY, USA, 2008, pp. 123–130.

[5] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin, "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Trans Inf Syst*, vol. 23, no. 1, pp. 103–145, Jan. 2005.

[6] H. Shimodaira, "Similarity and Recommender Systems." http://www.inf.ed.ac.uk/teaching/courses/inf2b-learn-note02-2up.pdf, 21-Jan-2014.

[7] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, New York, NY, USA, 2010, pp. 257–260.

[8] N. Polettini, *The Vector Space Model in Information Retrieval- Term Weighting Problem*, Department of Information and Communication Technology University of Trento, Italy, 2004.

[9] B. M. Kim, Q. Li, C. S. Park, S. G. Kim, and J. Y. Kim, "A New Approach for Combining Content-based and Collaborative Filters," *J Intell Inf Syst*, vol. 27, no. 1, pp. 79–91, Jul. 2006.

[10] K. Chapphannarungsri and S. Maneeroj, "Combining Multiple Criteria and Multidimension for Movie Recommender System," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2009, vol. I.

[11] L. Liu, N. Mehandjiev, and D.-L. Xu, "Multi-criteria Service Recommendation Based on User Criteria Preferences," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, New York, NY, USA, 2011, pp. 77–84.

AUTHORS PROFILE

Wiranto is a lecturer at the Bachelor Informatics Program, Faculty of Mathematics and Natural Sciences, Sebelas Maret University in Surakarta. He received both Bachelor and Master of Computer Science from Gadjah Mada University in Yogyakarta, Indonesia. He is currently taking his Doctoral Program at Department of Computer Science and Electronics, Gadjah Mada University.

Edi Winarko is an Associate Profesor and Head of Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia. He received his bachelor degree in Statistics from Gadjah Mada University, M.Sc in Computer Sciences from Queen's University, Canada and Ph.D in Computer Sciences from Flinders University, Australia. His research interests are data mining, data warehousing and information retrieval.

Sri Hartati is an Associate Profesor and Chair of Computer Science Graduate Program, Gadjah Mada University in Yogyakarta, Indonesia. She received her bachelor degree in Electronics and Instrumentation from Gadjah Mada University, both M.Sc and Ph.D in Computer Sciences from New Brunswick, Canada. Her research interests are intelligent systems, decision support systems, medical computing and computational intelligence.

Retantyo Wardoyo is an Associate Profesor and Former Head of Department of Computer Science and Electronics, Gadjah Mada University in Yogyakarta, Indonesia. He received his bachelor degree in Mathematics from Gadjah Mada University, M.Sc and Ph.D in Computer Sciences from University of Manchester, United Kingdom. His research interests are fuzzy systems and expert systems.

Ameliorate Threshold Distributed Energy Efficient Clustering Algorithm for Heterogeneous Wireless Sensor Networks

MOSTAFA BAGHOURI
Department of Physics
Faculty of Sciences
University of Abdelmalek Essaâdi
Tetouan, Morocco

SAAD CHAKKOR
Department of Physics
Faculty of Sciences
University of Abdelmalek Essaâdi
Tetouan, Morocco

ABDERRAHMANE HAJRAOUI
Department of Physics
Faculty of Sciences
University of Abdelmalek Essaâdi
Tetouan, Morocco

***Abstract*—Ameliorating the lifetime in heterogeneous wireless sensor network is an important task because the sensor nodes are limited in the resource energy. The best way to improve a WSN lifetime is the clustering based algorithms in which each cluster is managed by a leader called Cluster Head. Each other node must communicate with this CH to send the data sensing. The nearest base station nodes must also send their data to their leaders, this causes a loss of energy. In this paper, we propose a new approach to ameliorate a threshold distributed energy efficient clustering protocol for heterogeneous wireless sensor networks by excluding closest nodes to the base station in the clustering process. We show by simulation in MATLAB that the proposed approach increases obviously the number of the received packet messages and prolongs the lifetime of the network compared to TDEEC protocol.**

Keywords*—Heterogeneous wireless sensor networks; Clustering based algorithm; Energy-efficiency; TDEEC Protocol; Network lifetime

I. INTRODUCTION

Wireless sensor network is the set of sensor nodes, deployed in the hostile environment, in the goal to sense the events detection, such temperature, pressure or vibration and send their measurements toward a processing center called sink [1], [2]. These tiny nodes are limited in their battery capacity which its replacement is impossible. Furthermore, an important part of energy is consumed in the communication circuit which must be minimized. Because of those limitations, the major wireless sensor networks' challenging issues is the energy consumption.

A number of research techniques about energy-efficient have been proposed to solve these problems. In order to support data aggregation through efficient network organization, nodes can be partitioned into a number of small groups called clusters. Each cluster has a cluster head, and a number of member nodes [3]. Among WSN heterogeneous protocols there is DEEC (Design of a distributed energy-efficient clustering algorithm) [4]. This protocol is based on the election of cluster head by the balance of the remaining energy probabilities for each node. It uses the average energy of the network as the energy reference. The cluster-heads are elected by a probability based on the ratio between the residual energy of each node and the average energy of the network. DEEC has improved by a Stochastic approach SDEEC [5], which reduces the intra-cluster transmission. In this protocol the non-CH are going in to sleep mode to conserve more energy. Another version of improved DEEC is DDEEC which define a new residual energy threshold to elect CH [6]. On the other hand TDEEC enhance the network lifetime by introducing a new threshold based on the residual energy to become CH [7]. The last version of TDEEC is ETDEEC which prolong the lifetime by modifying the probabilities of CH election based on the distance average between the CHs and BS [8].

Otherwise, in order to improve the lifetime of the network, ATDEEC employs a new technique which excludes closest nodes to the base station from the clustering process. The remainder of the paper is organized as follows. In section II the main related works are summarized. Section III and IV introduced the problem formulation and proposed approach. Sections V and VI explains the network and the energy models. Therefore, theoretical analysis are presented and discussed in Section VII, whereas section VIII describes performance analysis of the proposed method. Finally, Section IX concludes our work, and discusses some future directions.

II. RELATED WORK

Currently, clustered routing protocols have gained actually increasing attention from researchers because it's potential in extending WSN lifetime. Heinzelman et al. designed and implemented the first distributed and clustered routing protocol with low energy consumption LEACH [9]. Moreover, the heterogeneous protocols are more energy efficient than the homogeneous ones. Q. Li et al. have proposed Distributed Energy Efficient Clustering Protocol (DEEC) [4]. This protocol is based on multi-level and two level energy heterogeneous schemes. The cluster heads are selected using the probability utilizing the ratio between residual energy of each node and the average energy of the network. The epochs of being cluster-heads for nodes are different according to their initial and residual energy. A particular algorithm is used to estimate the network lifetime. Afterward, the network can avoid the need of assistance by routing protocol [4]. TDEEC [7] uses the same process of CH selection and estimation of

average energy as in DEEC. At start of each round, the nodes decide whether or not to become a CH by selecting a random number within 0 and 1. If this selected number is lower than a threshold, then the node becomes a CH for this round. Simulation results show that in terms of network lifetime, both EDEEC and TDEEC protocols are better than DEEC. TDEEC provide best results compared to the three versions over DEEC. Otherwise, Suniti Dutt et al [6], has proposed ETDEEC protocol to enhance the network lifetime by introducing a distance factor in CH probability. However, this approach present a limitation lies in the fact that the network instability observed after the death of the first node is caused mainly by the bad energy distribution. It means that all nodes not die approximately at the same time.

III. Problem Formulation

In this paragraph we formulate the problem that we'll solve in the next sections. We consider a network with N nodes, which are uniformly distributed in a M × M network field as shown in Figure1.

Each node has a mission to send every time the data to the base station which is located at the center of network. This network divide in the cluster regions, and the cluster-heads receive the data from the member nodes to transmitting toward the base station. According to this model, it was found that the member nodes that are closer to the base station must go through a long path to route a data.

Contrariwise, they have the possibility to send the packet messages directly to the base station (Figure 1). In this case, these nodes should not go through the CH election process. Consequently we can conserve the lost energy during this step and we can prolong the network lifetime. To simulate this problem, we present in the next section the model of the studied network.

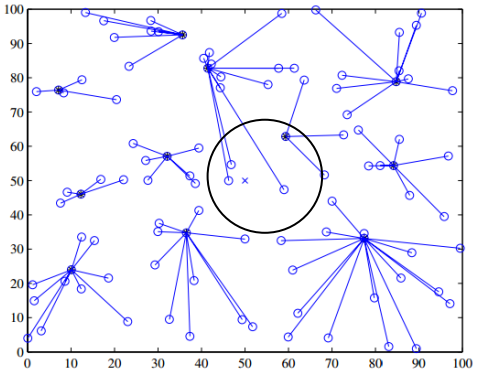


Fig. 1 Through the clustering process, all nodes must form clusters even those who are closest to the base station

IV. Proposed Method

This paper proposes a new approach called Ameliorate Threshold Distributed Energy Efficient Clustering (ATDEEC) algorithm whose main objective is to increase the lifetime of the network and to enhance the ability to deliver more packet messages in the heterogeneous WSN by minimizing the number of the nodes elected to become CH.

V. Energy Model

This study assumes a simple model for the radio hardware where the transmitter dissipates energy for running the radio electronics to transmit and amplify the signals, and the receiver runs the radio electronics for reception of signals [7]. Multipath fading model (d^4 power loss) for large distance transmissions and the free space model (d^2 power loss) for proximal transmissions are considered. Thus to transmit an $l-$ bits message over a distance d, the radio expends:

$$E_{Tx}(l,d) = E_{Tx-elec}(l) + E_{Tx-amp}(l,d) \quad (1)$$

$$E_{Tx-elec}(l) = lE_{elec} \quad (2)$$

$$E_{Tx-amp}(l,d) = \begin{cases} l\epsilon_{fs}d^2, when\ d < d_0 \\ l\epsilon_{mp}d^4, when\ d \geq d_0 \end{cases} \quad (3)$$

Where d_o is the distance threshold for swapping amplification models, which can be calculated as $d_o = \sqrt{\frac{\epsilon_{fs}}{\epsilon_{mp}}}$

To receive an $l\ bits$ message the receiver expends:

$$E_{Rx}(l) = lE_{elec} \quad (4)$$

To aggregate n data signals of length l − bits, the energy consumption was calculated as:

$$E_{DA-expend}(l) = lnE_{DA} \quad (5)$$

VI. Network Model

This section describes the network model and other basic assumptions:

1) N sensors are uniformly distributed within a square field of area $A = M \times M$. The Base Station is positioned at the center of the square region. The number of sensor nodes N to be deployed depends specifically on the application.

2) All nodes are deployed randomly and can fall in the one of two types of regions which can be defined by the threshold distance R from the base station.

3) In this case we define two types of nodes, Excluded and not Excluded nodes. The Excluded are the nodes that not enter in the clustering process because there are closed to the base station and the other are far.

4) All sensors are heterogeneous, i.e., they not have the same capacities.

5) All the sensor nodes have a particular identifier (ID) allocated to them. Each cluster head coordinates the MAC and routing of packets within their clusters. (see Fig. 2)

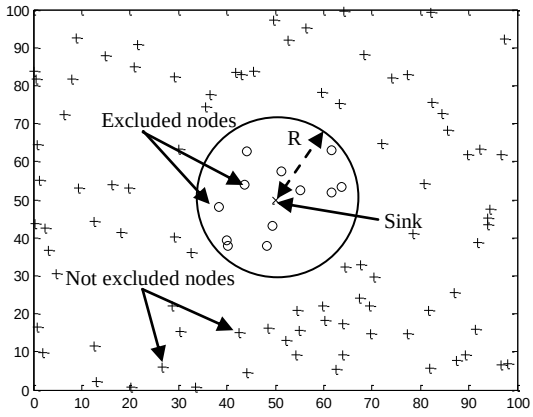


Fig. 2 Wireless Sensor Network model

VII. THEORETICAL ANALYSIS

Let E[d_{toBS_ex}] be the Expected distance of Exclude node from the base station. Assuming that the nodes are uniformly distributed, so it is calculated as follows:

$$E\left[d_{toBS\_ex}^2\right] = \int_0^{x_{max}} \int_0^{y_{max}} (x^2 + y^2)\, \rho(x,y) dx dy \quad (6)$$

$$E\left[d_{toBS\_ex}^2\right] = \int_0^{R} \int_0^{2\pi} r^2\, \rho(r,\theta) r dr d\theta \quad (7)$$

If the density of sensor nodes is uniform throughout the area then becomes independent of r and θ. It is equal to $\rho = \frac{1}{\pi R^2}$ then:

$$E\left[d_{toBS\_ex}^2\right] = \frac{R^2}{2} \quad (8)$$

According to the energy model proposed in section 5, the energy consumed by each Excluded nodes is:

$$E_{Exclu} = lE_{elec} + lE_{DA} + l\epsilon_{fs} d_{toBS\_ex}^2 \quad (9)$$

By combining the equations (8) and (9) the energy consumed by each Excluded nodes is:

$$E_{Exclu} = lE_{elec} + lE_{DA} + l\epsilon_{fs} \frac{R^2}{2} \quad (10)$$

The energy consumed by the Not Excluded nodes is:

$$E_{not\_Exclu} = cE_{cluster} = cE_{CH} + (N-s)E_{not\_CH} \quad (11)$$

Where E_{CH} and E_{not_CH} are the energy consumed by each cluster head and member node respectively and can be calculated by:

$$E_{CH} = lE_{elec}\left(\frac{N-s}{c} - 1\right) + lE_{DA}\left(\frac{N-s}{c}\right) + lE_{elec} + l\epsilon_{mp} d_{toBS\_not\_ex}^4 \quad (12)$$

$$E_{not\_CH} = lE_{elec} + l\epsilon_{fs} d_{toCH}^2 \quad (13)$$

Where $d_{toBS\_not\_ex}$ is the average distance of not Excluded node from the base station and d_{toCH} is the average distance between cluster members to CH.

Now $d_{toBS\_not\_ex}$ and d_{toCH} can be calculated as:

$$d_{toBS\_not\_ex}^2 = \int_0^{\sqrt{\frac{M^2-\pi R^2}{\pi}}} \int_0^{2\pi} r^2\, \rho(r,\theta) r dr d\theta = \frac{M^2 - \pi R^2}{2\pi} \quad (14)$$

$$d_{toCH}^2 = \int_0^{\sqrt{\frac{M^2-\pi R^2}{c\pi}}} \int_0^{2\pi} r^2\, \rho(r,\theta) r dr d\theta = \frac{M^2 - \pi R^2}{2\pi c} \quad (15)$$

Where c denoting the number of the clusters. The energy total dissipated in a network is:

$$E_{total} = sE_{Exclu} + (N-s)E_{non\_Exclu} \quad (16)$$

Where *s* is the number of the excluded nodes.

Using the Eq.11 to Eq. 16 the expected value of the energy dissipated in the network is calculated as follows:

$$E_{total} = ls\left[E_{elec} + \epsilon_{fs}\frac{R^2}{2}\right] + l(N-s)\left[NE_{elec} + (N-s)E_{DA} + c\epsilon_{mp}\left(\frac{M^2-\pi R^2}{2\pi}\right)^2\right] + l(N-s)^2\left[E_{elec} + \epsilon_{fs}\frac{M^2-\pi R^2}{2\pi c}\right] \quad (17)$$

The optimal number of clusters can be found by letting $\frac{\partial E_{total}}{\partial c} = 0$

$$C_{opt} = d_0\sqrt{\frac{2\pi(N-s)}{M^2 - \pi R^2}} \quad (18)$$

Where $d_0 = \sqrt{\frac{\epsilon_{fs}}{\epsilon_{mp}}}$ is the distance threshold for swapping amplification models and R must be less the threshold R_o, where $R_o < \frac{M}{\sqrt{\pi}}$.

The different forms of the E_{total} calculation will lead to different optimal c_{opt} settings depending on the values of, R and s. The optimal probability for becoming a cluster-head can also be computed as $P_{opt} = \frac{c_{opt}}{N-s}$

In Figure 3, we show the average energy consumption by each sensor node against varying numbers of clusters for different values of number of excluded nodes *s* and threshold distance *R* from base station.

While the number of cluster increases, the total energy starts to decrease and reaches a minimum for clusters number comprised between 10 and 18 depending on the value of *s* and *R*. However, it is clearly shown that when s increases, the energy consumption decreases and turns between 4.069 J and 1.473 J. These results are coincided with our conception and our goals. In the next section we have evaluate these results by computer simulation the network in Matlab.

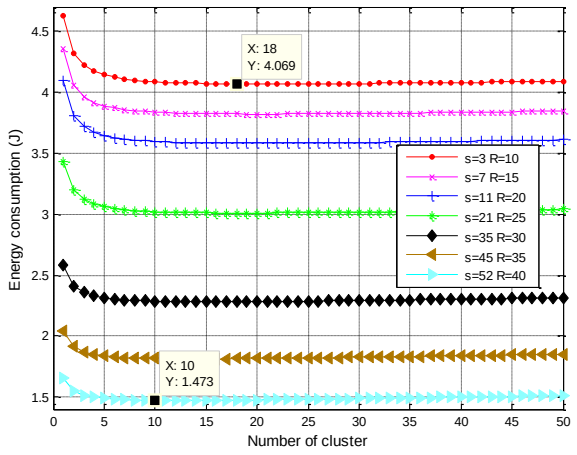


Fig. 3 Variation of energy consumption for different values of *R* and *s* depending on clusters number *c*.

VIII. Simulation Results

In this section, we simulate the performance of ATDEEC protocol under different scenarios using MATLAB. We consider a model illustrate in the figure 2 with N = 100 nodes randomly distributed in a 100m × 100m field. To compare the performance of ATDEEC with TDEEC protocol, we ignore the effect caused by signal collision and interference in the wireless channel. The radio parameters used in our simulations are shown in Table1.

TABLE I. Energy Model Parameters

Parameter	Value
Initial Node Energy	0.5J
N	100
E_{elec}	50 nJ/bit
E_{DA}	5 pJ/bit
ϵ_{fs}	10 pJ/bit/m^2
ϵ_{mp}	0.0013 J/bit/m^4
d_0	87 m
L	4000 Bytes
Rounds	8000

We define two performance metrics to evaluate our protocol as: First Node Dies (FND), or stability period and Last Node Dies (LND), or instability period.

First, we present an empirical result for the optimal number of cluster-head C_{opt} and optimal threshold distance to the base station for our ATDEEC protocol shown in Figure 4. The number of cluster-heads decreases from 10 to 45 meters. This figure reveals that although the cluster-heads decreases from 5 to 17, the FND improves significantly and has a maximum value at 20 meters. Beyond this value, the curve starts descending. The optimality of C_{opt} lies around 17 cluster-heads for our setup. This result can be interpreted by when the threshold distance R start to increase, the closer nodes to the base station consume less energy, because they send data directly to it. However, when this distance increases the nodes become farther away and consume more energy.

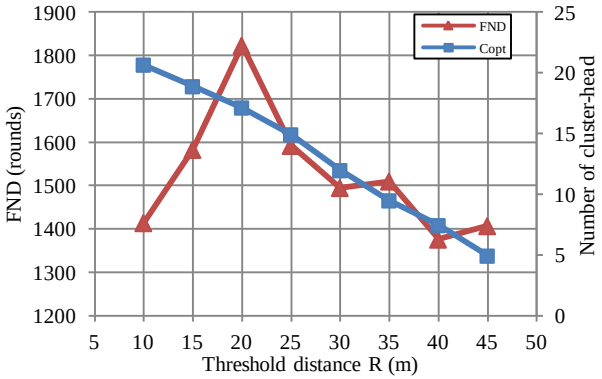


Fig. 4 FND and C_{opt} vs Threshold distance R

On other hand, we study three other performance metrics such as, the number of live nodes per round, energy residual and number of message packets for both ATDEEC and TDEEC protocols. The simulation results are discussed below.

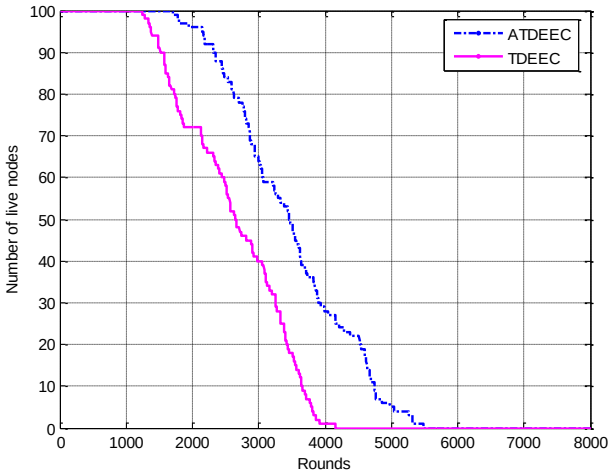


Fig. 5 Life time ATDEEC and TDEEC comparison

Figure 5 shows the network lifetime of ATDEEC and TDEEC for threshold distance equal to 20m. Since the TDEEC protocol is designed to be robust with respect to a heterogeneous network, we test the performance of ATDEEC against these criteria. Based on our experimental results, we conclude that ATDEEC has a superior stability period life time performance compared with TDEEC by an increase with 25% as shown in this same figure.

In the Figure 6, we emphasis our discussion on how each node consumes its own residual energy in the network. This energy is calculated during the network operation, by observing the variation of energy levels between the nodes at each round. The total initial energy of the network is 90 J which decreases linearly up to 3000 rounds and after that there is a difference from the round where first node dies in respect to them. Energy residual per round for ATDEEC is more as compared to TDEEC.

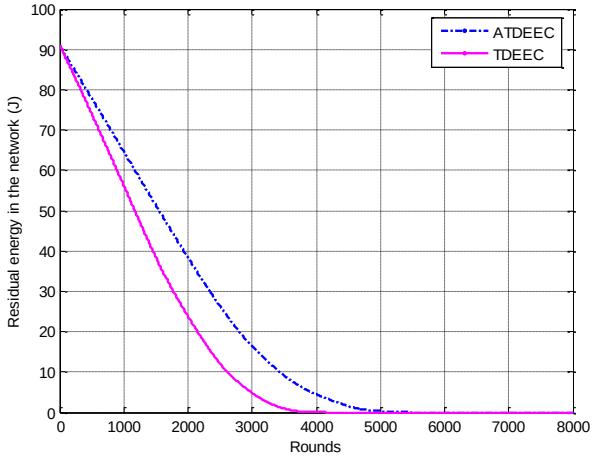


Fig. 6 Total residual energy over rounds TDEEC and TDEEC

Referred to figure 7, it show clearly that proposed approach provide a better throughput compared to TDEEC protocol, this increase is justified by the life time enhancement which give the improved ATDEEC protocol.

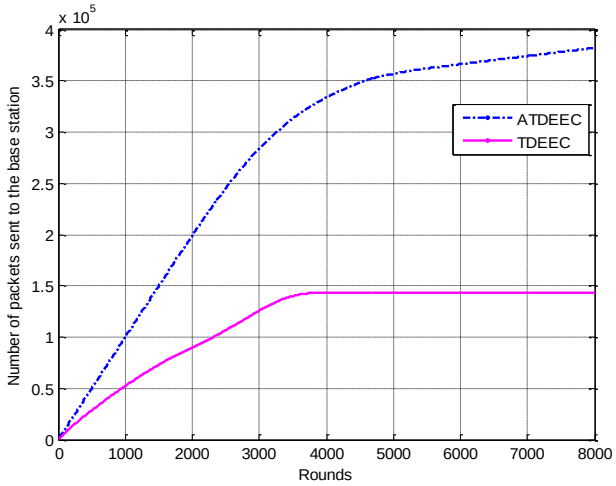


Fig. 7 Performance of the protocols

Generally, we can illustrate the increase of the proposed protocol in the figure 8. It's noted that the throughput increases twice as much than TDEEC due to its energy efficiency. Whereas, ATDEEC outperforms the FND of TDEEC by 25% and by 46% for LND.

IX. Conclusion and Future Work

In this paper, an energy efficient protocol ATDEEC has been proposed to solve the problem of the closest nodes to the base station which were consumed more energy in data traffics.

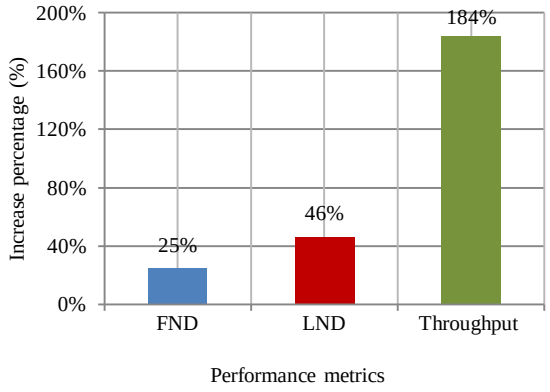


Fig. 8 Performance metrics of the ATDEEC protocol

The simulation result by Matlab, demonstrate the ability of developed algorithm to prolong the network lifetime significantly and increase the number of packet messages received by the base station. In the future work we'll evaluate this approach by the real-time performances and simulate it by adequate simulator software.

References

[1] Kay Romer and Friedemann Mattern. "The Design Space of Wireless Sensor Networks". IEEE Wireless Communications, 11(6):54–61, December 2004.

[2] Kay Romer and Friedemann Mattern. "The Design Space of Wireless Sensor Networks". IEEE Wireless Communications, 11(6):54–61, December 2004.

[3] V. Raghunathan, C. Schurgers, Park. S, and M. B. Srivastava, "Energy aware wireless micro-sensor networks". IEEE Signal Processing vol. 19 no. 2, pp. 40 –50, 2002,

[4] Li Qing, Qingxin Zhu, Mingwen Wang, "DEEC: Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor networks", Computer Communications 29 (2006) 2230–2237.

[5] Elbhiri Brahim,Saadane Rachid, Driss Aboutajdine, "Stochastic Distributed Energy-Efficient Clustering (SDEEC) for heterogeneous wireless sensor networks", ICGST-CNIR Journal, vol. 9, no. 2, pp. 11-17, Dec. 2009.

[6] Elbhiri Brahim,Saadane Rachid, Sanaa El fkihi, Driss Aboutajdine, "Developed Distributed Energy-Efficient Clustering (DDEEC) for heterogeneous wireless sensor networks", IEEE Communications and Mobile Network (ISVC), 5th International Symposium on, oct. 2010.

[7] Parul Saini, Ajay.K.Sharma, "Energy Efficient Scheme for Clustering Protocol Prolonging the Lifetime of Heterogeneous Wireless Sensor Networks", International Journal of Computer Applications (0975 8887), vol. 6, no.2, September 2010.

[8] Suniti Dutt, O. S. Khanna, "An Enhanced Energy Efficient Clustering Scheme for Prolonging the Lifetime of Heterogeneous Wireless Sensor Networks", International Journal of Computer Applications (0975 – 8887) vol. 76, no.8, August 2013

[9] Wendi R. Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan, "Energy efficient communication protocol for wireless microsensor networks", IEEE International Conference on System Sciences, pp 1-10, 2000.

Green Technology, Cloud Computing and Data Centers: the Need for Integrated Energy Efficiency Framework and Effective Metric

Nader Nada
Fatih University
Istanbul, Turkey

Abusfian Elgelany
Sudan University
Khartoum, Sudan

***Abstract*—Energy efficiency (EE), energy consumption cost and environmental impact are vibrant challenges to cloud computing and data centers. Reducing energy consumption and emissions of carbon dioxide (CO_2) in data centers represent open areas and driving force for future research work on green data centers. Our Literature review reveals that there are currently several energy efficiency frameworks for data centers which combine a green IT architecture with specific activities and procedures that led to decrease the impact on environment and less CO_2 emissions. The current available frameworks have some pros and cons that is the reason why there is an urgent need for an integrated criterion for selecting and adopting energy efficiency framework for data centers. The required energy efficiency framework criteria should also consider the social network applications as a vital related factor in elevating energy consumption, as well as high potential for better energy efficiency in data centers. Additionally, in this paper, we highlighted the importance of the identification of efficient and effective energy efficiency metric that can be used for the measurement and determination of the value of data centers efficiency and their performance combined with sound and empirically validated integrated EE framework.**

Keywords—Cloud Computing; green Cloud; Datacenter; Energy efficiency

I. Introduction To Green Technology In Cloud Computing

Cloud computing is a promising area in distributed computing. Data centers are the main component of cloud computing. Data centers energy consumption cost and environmental effect are dynamic challenge to cloud computing. Additionally, the growing usage of social applications and the expansion of e-business require an increase in the number of data centers. However, the combination of global warming and inconsistent climate make the cost of energy a major challenge for the sustainability of e-business [1]. It is a corner stone of the infrastructure of cloud computing approach by which a variety of information technology (IT) services were built. They extended the ability of centralized repository for computing, hosting, storage, management, monitoring, networking and deployment of data.

With the rapid increase in the capacity and size of data centers, there is a continuous increase in the demand of energy consumption [2]. Data centers, beside their ongoing high energy consumption, also produce carbon dioxide that riddled with IT inefficiencies. International Data Corporation (IDC) annual report found that cloud computing reached $42bn in 2012 and the revenue of cloud in 2013 was $150bn [3].

Environmental impact of Information Technology (IT) under the banner of "Green IT" was been discussed by academia, media and government Since (2007), when the Environmental Protection Agency (EPA) submitted a report to the US Congress [5] about the expected energy consumption of data centers. Since then Green IT has been receiving growing attention. The overall objective of Green IT is to increase energy efficiency and reduce CO_2 emissions [6], figure 1 represents the effect of good practice of green data centers to gas emission. There are two ways to make data center greener: First, improve energy efficiency of data center, second, use clean energy supply. Cloud computing has different techniques to solve energy-efficient problem by minimizing the impact of cloud computing on the environment. These techniques deal with energy efficiency consumption like virtualization, hardware base, operating systems base and data centers. Some new features arise like energy performance, and time wise. However, the concerns should be to swap problem between energy consumption and performance.

II. Literature Review On Energy Efficiency Frameworks For Cloud Computing

In our literature review below is based on previous studies of investigated energy efficiency on cloud computing and focused on data center technology.

Asghar Sabbaghiet al.[9], investigated previous researches and introduced energy efficiency framework on information technology that enabled Green supply chain management. They proposed a unique conceptual taxonomy of information technology for sustainability. They also identified the relationship between Green supply chain management information flow, IT governance and Green infrastructure components.

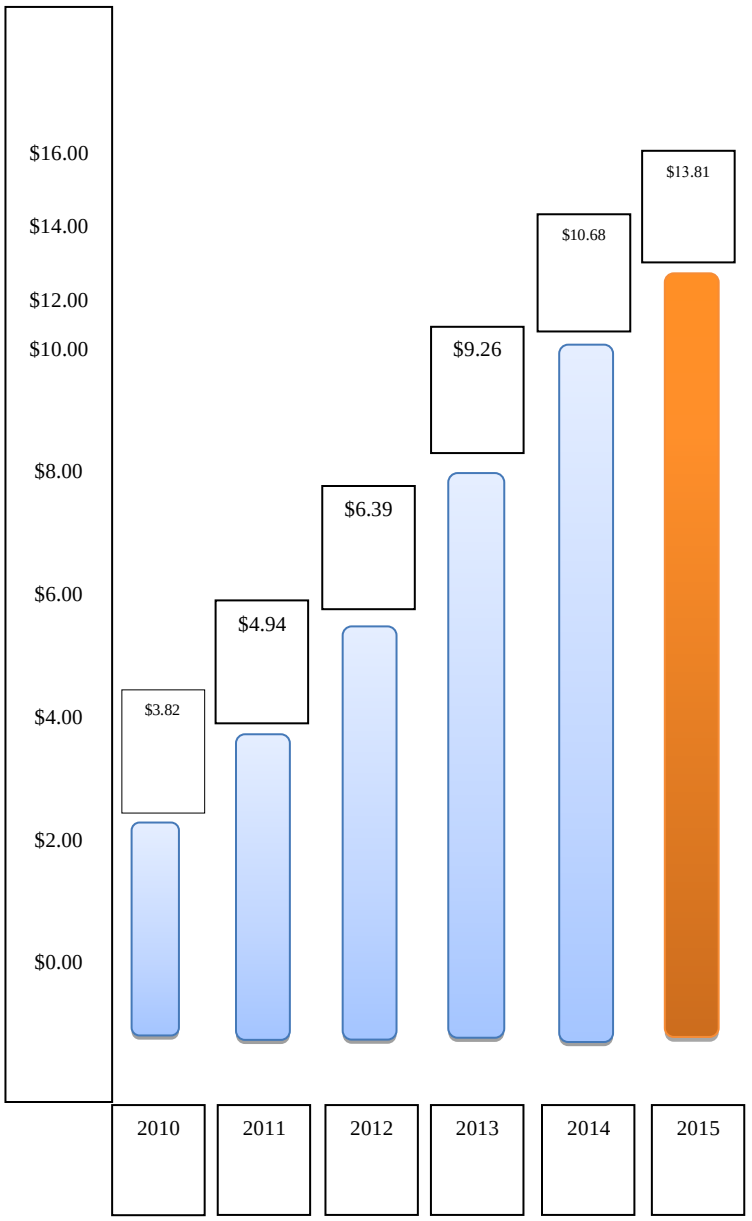


Fig. 1. Green Data Center Market Value [7]

Zhiming Wang et al.[10], proposed mechanism to support maximizing resource utilization by using active and idle energy consumption by finish time minimization. This mechanism reduces the power consumption by allowing spare servers to be in idle state. This mechanism put into account QoS of cloud datacenter.

RajkumarBuyya et al.[11], proposed a novel mechanism in three ways: (a) architectural principles for energy-efficient management of Clouds; (b) energy-efficient resource allocation policies and scheduling algorithms considering QoS, and devices power usage characteristics; and (c) a novel software technology for energy-efficient management of Clouds.

Anton Beloglazov et al.[12], developed a unique mechanism which supports dynamic consolidation of VMs based on adaptive utilization thresholds, which put into account Service Level Agreements (SLA).

Nguyen Quang Hung et al.[13], proposed unique server selection policy, and four algorithms solving the lease scheduling problem. This approach reduces 7.24% and 7.42% energy consumption than the existing greedy mapping algorithm.

Uddin et al.[14]and his team introduced a unique framework to improve the performance and energy efficiency of data centers. They developed a classification mechanism for data center components depending on different resource pools and different parameters like energy consumption, resource utilization, workload, etc. The framework highlights the importance of implementing green metrics like Power Usage Effectiveness (PUE) to measure the efficiency of data center in terms of energy utilization and carbon dioxide (CO_2) emissions. The framework is based on virtualization and cloud computing to increase the resource utilization of already installed servers from 10% to more than 50%.

Meenakshi Sharma et al.[15],developed a new mechanism with two steps: firstly they developed an analysis of different Virtual Machine(VM) load balancing algorithms, second introduced a new VM load balancing algorithm that has been developed and implemented in Virtual Machine environment of cloud computing in order to achieve better response time and cost.

In S. Kontogiannis et al.[16], the research team developed a unique mechanism called Adaptive Workload Balancing algorithm (AWLB) for cloud datacenter based web systems which deals with agents into two dimensions the web datacenter and web servers. AWLB algorithm also supports protocol specification for signaling purposes among web switch and datacenter nodes and also utilizes other protocols such as SNMP and ICMP for its balancing process. Performance gains are shown from tests of AWLB against known balancing Least Connections (LC) and Least Loaded (LL) algorithms. Table 1 represents the summary of our literature review on cloud computing energy efficiency frameworks and techniques.

III. Urgent Need For Energy Efficiency Integrated Framework For Cloud Computing and Data Centers

Reducing energy consumption and emissions of carbon dioxide (CO_2) in data centers represent open challenges and driving the future research work for green data centers. Our Literature review reveals that there is an urgent need for integrated energy efficiency framework for data centers which combines a green IT architecture with specific activities and procedures that led to minimal impact on environment and less CO_2 emissions. The required energy efficiency framework should also consider the social network applications as a vital related factor in elevating energy consumption, as well as high potential for energy efficiency.

TABLE I. DATA CENTERS ENERGY EFFICIENCY TECHNIQUES

No	Author	Approach	Strengths	Limitation
1	Asghar Sabbaghi	Conceptual taxonomy of information technology	Supply Management	Focus on infrastructure only
2	Zhiming Wang	Maximizing resource utilization	Put into account QoS	Much job performance take amount of time Sleep-in-Waking up-ready.
3	RajkumarBuyya	Resource allocation and scheduling	Quality-of-service	No parameter to indicate CO_2 emission
4	Anton Beloglazov	Adaptive utilization	Meeting the Service Level Agreements (SLA)	No parameter to show the energy efficiency level
5	Meenakshi Sharma	Load balancing algorithms	Good in reduce energy, pricing and time	Much calculation need more time to take decision
6	Mueen Uddin	Virtualization	Increase the utilization ratio	High utilization leads to introduce CO_2
7	S. Kontogiannis	workload balancing algorithm	can balance the workload in multidimensional resources	Increase the Web traffic

IV. GREEN METRICS TO MEASURE AND ASSESS ENERGY EFFICIENCY OF DATA CENTER

Globally, the energy consumption of data centers is continuously on the increase [17]. The energy operations cost will continue to double every five years between 2005 and 2025 [18]. This increase led to higher emission of CO_2that reflects negatively on global warming and environmental health.

Measuring energy consumption of data centers has become a significant concern of all datacenters stakeholders to meet end-user agreement [19].Energy efficiency metric is a tool used to measure energy efficiency in data centers [20]. The most important challenge in the data centers industry is the limitation of effective standard energy efficiency metrics, which supports improving energy efficiency [21,22].

For an effective energy efficiency assessment on data centers and its components, we need to assess the effectiveness of the used metrics and to measure the energy efficiency of data centers. To determine whether these metrics are effective or not we need to assess these metrics against its intended goals and under a range of common used cases to determine the values of its effectiveness in terms of reporting, targets, education, analysis and decision support [23].

Our literature review on common energy efficiency metrics that are currently in use by data centers reveals that none of these metrics are meeting the prior mentioned criteria. Therefore, our research is not only introducing a comparative review of the most common used metrics and their features (criteria) but also attempting to recommend better metric to be used in the assessment of data centers energy efficiency.

In last few years operators have adopted PUE metrics as the measure of energy efficiency for the mechanical and electrical infrastructure of the data center. The process of assessment has submitted a focus and comparable measure of performance, which has enabled data centers operators to make substantial improvements. However, until now no consensus about IT or software energy efficiency and most energy efficiency measurements stopped at the IT power cord. In this paper we are proposing the Fixed to Variable Energy Ratio (FVER) metric which could be used to measure the data centers energy efficiency instead of PUE. The reason behind our choice in favor of the FVER metric is that it combines and meets all the needed criteria for better energy efficiency assessment in data centers, listed in table 2, including the usage of IT and software applications in data centers [24]. Figure 1 depicts the difference between FVER and PUR and Table 2 represents the different Goals of energy efficiency metrics including PUE, DCiE, FVER, and DCeP where:

Power Usage Effectiveness (PUE) =

Total Facility Power/ IT Equipment Power (1)

Data Center Infrastructure Effectiveness (DCiE) = 1 / PUE (2)

Fixed to Variable Energy Ratio (FVER) =1+Fixed Energy/Variable Energy (3)

Data Center Energy Productivity (DCeP) = Useful Work Produced / Total Data Center Energy Consumed over time (4)

V. CONCLUSION AND CONTRIBUTION

The first contribution of this paper is our literature review on current energy efficiency frameworks. The study reveals that there are currently several energy efficiency frameworks for data centers which combine a green IT architecture with specific activities and procedures that will lead to decrease the impact on environment and the diminution of CO_2 emissions. The current available frameworks have some pros and cons (see Table 1) that is why there is an urgent need for an integrated energy efficiency framework for data centers and cloud computing. The framework should consider a common and integrated set of criteria. The selection and adoption of such framework should be in accordance with the data center area of application and its surrounding environment.

The second contribution was the literature review on energy efficiency metrics that are currently used for the assessment of energy efficiency in data centers (depicted in Table 2 and Figure 2). This part of our study developed a comparative study of the most commonly used metrics and their features (criteria), additionally we recommended the use of FVER instead of PUE as a better metric for the assessment of data centers energy efficiency which was based on certain required criteria including the usage of IT and software

applications in data centers. Our future work will focus on the development and empirical validation of an integrated energy efficiency framework for cloud computing and data centers.

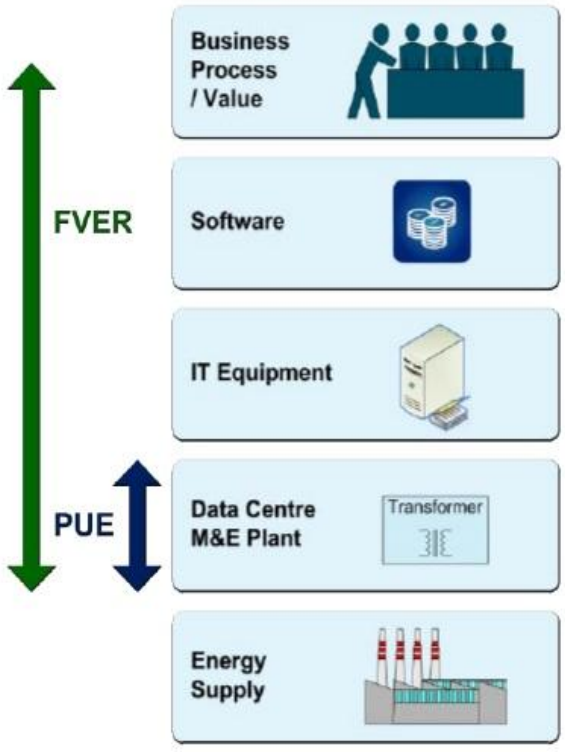


Fig. 2. FVER Vs PUE [25]

TABLE II. GOALS OF ENERGY EFFICIENCY METRICS [24]

No	Goal	PUE	DCiE	FVER	DCeP
1	Provide a clear, preferably intuitive understanding of the measure	Y	Y	Y	
2	Provide a clear, preferably intuitive direction of improvement		Y	Y	Y
3	Describe a clearly defined part of the energy to useful work function of the IT services		Y	Y	Y
4	Be persistent, i.e. the metrics should be designed to be stable and extensible as the scope of efficiency measurement increases, rather than confusing the market with rapid replacement	Y	Y	Y	
5	Demonstrate the improvements available in a modern design of facility		Y	Y	
6	Demonstrate the improvements available through upgrade of existing facilities using more efficient M&E systems		Y	Y	
7	Provide a clear, preferably intuitive understanding of the impacts of changes			Y	Y
8	Be reversible, i.e. it should be possible to determine the energy use at the electrical input to the data center for any specified device or group of devices within the data center	Y		Y	Y
9	Be capable of supporting 'what if' analysis for IT and data center operators in determining the energy improvement and ROI for improvements and changes to either the facility or the IT equipment it houses	Y		Y	

REFERENCES

[1] Mell, P. and T. Grance. The NIST Definition of Cloud Computing, 2009.

[2] Mueen Uddin, framework for energy efficient data centers using virtualization, 2012.

[3] IDC - Press Release, 2013.

[4] Ian Foster, Cloud Computing and Grid Computing 360-Degree Compared, 2008.

[5] James W. Smith, Green Cloud A literature review of Energy-Aware Computing, 2011.

[6] Asghar Sabbaghi, Green Information Technology and Sustainability, 2012.

[7] Ariel Schwartz, Green Data Center Marketto More than TripleOver Next Five Years, 2010.

[8] Eric Woods, Data Center Electricity Consumption 2005-2010: The Good and Bad News, 2011.

[9] Asghar Sabbaghi, Green information technology and sustainability: A Conceptual taxonomy, volume 13, Issue 2, pp. 26-32, 2012.

[10] Zhiming Wang, Energy-aware and revenue-enhancing Combinatorial Scheduling in Virtualized of Cloud Datacenter, Volume7, Number1, January 2012.

[11] Rajkumar Buyya, Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges, 2010.

[12] Anton Beloglazov, aSurvey on Power Management Solutions for Individual Systems and Cloud, 2010.

[13] Nguyen Quang Hung, Performance constraint and power-aware allocation for user requests in virtual computing, 2011.

[14] Mueen Uddin, Green Information Technology (IT) framework for energy efficient data centers using virtualization, 2012.

[15] Meenakshi Sharma, Performance Evaluation of Adaptive Virtual Machine Load Balancing Algorithm, 2012.

[16] S. Kontogiannis, A probing algorithm with Adaptive workload load balancing capabilities for heterogeneous clusters, journal of computing, volume 3, issue 7, July2011.

[17] Lacity, Mary C and Khan, Shaji A and Willcocks, Leslie P, A review of the IT outsourcing literature: Insights for practice, The Journal of Strategic Information Systems, Elsevier, 18, 130-146 (2009).

[18] Laura Sisó, Ramon B. Fornós, Assunta Napolitano &Jaume, Energy- and Heat-aware metrics for computing modules, 2012.

[19] Tung, Teresa, Data Center Energy Forecast, Silicon Valley Leadership Group, San Jose, CA, (2008).

[20] Wang, Lizhe and Khan, Samee U, Review of performance metrics for green data centers: a taxonomy study, The Journal of Supercomputing, Springer, 1-18 (2013).

[21] Belady, Christian L and Malone, Christopher G, Metrics and an infrastructure model to evaluate data center efficiency, Proceedings of the Pacific Rim/ASME InternationalElectronic Packaging Technical Conference and Exhibition(IPACK), ASME, (2007).

[22] Rivoire, Suzanne and Shah, Mehul A and Ranganathan, Parthasarathy and Kozyrakis, Christos, JouleSort: a balancedenergy-efficiency benchmark, Proceedings of the 2007 ACMSIGMOD international conference on Management of data,ACM, 365-376 (2007).

[23] Liam Newcombe, Data center energy efficiency metrics existing and proposed metrics to provide effective understanding and reporting of data center energy, 2013.

[24] Liam Newcombe, Data center Fixed to VariableEnergy Ratio metricDC-FVER, 2012

[25] Peter Hopton, Move Over PUE, 2012.

[26] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[27] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[28] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[29] K. Elissa, "Title of paper if known," unpublished.

[30] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[31] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[32] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Computation of Single Beam Echo Sounder Signal for Underwater Objects Detection and Quantification

Henry M. Manik
Department of Marine Science and Technology Faculty of Fisheries and Marine Sciences
Bogor Agricultural University (IPB)
Kampus IPB Darmaga Bogor
Indonesia

Asep Mamun
Department of Marine Science and Technology Faculty of Fisheries and Marine Sciences
Bogor Agricultural University (IPB)
Kampus IPB Darmaga Bogor
Indonesia

Totok Hestirianoto
Department of Marine Science and Technology Faculty of Fisheries and Marine Sciences
Bogor Agricultural University (IPB)
Kampus IPB Darmaga Bogor
Indonesia

***Abstract*—Underwater Acoustic methods have been extensively used to locate and identify marine objects. These applications include locating underwater vehicles, finding shipwrecks, imaging sediments and imaging bubble fields. Ocean is fairly transparent to sound and opaque to all other sources of radiation. Acoustics technology is the most effective tool for monitoring this environment because of the sound's ability to propagate long distance in water. We used single beam echo sounder to discriminate underwater objects. Development of the algorithm and applied it to detect and quantify underwater object such as fish, sea grass, and seabed. We found the detected target has different backscatter value.**

***Keywords*—single beam; echo sounder; backscattering; algorithm**

I. INTRODUCTION

Ocean acoustics is the use of sound to measure the distribution and abundance of fish and other aquatic organisms. The physics of sound propagation in seawater is well understood [1; 2] and an appreciation thereof is helpful to interpret acoustics data correctly. Fortunately, analysis software performs most calculations; however, fish are complex sound scatterers, and theory developed for simple bodies such as spheres is only partly applicable [3;4;5;6]. Therefore, practical experience with fisheries surveys is also important. A scientific echo sounder consists of a transceiver (which includes transmitting and receiving electronics), a transducer, and a recording device, which is usually a computer. The computer controls operation of the echo sounder [7;8;9;10;11]. The transceiver sends a short electric signal to the transducer, which transforms this electric energy to a sound pulse (also called a ping). Ease of use the time until the next transmission, the transducer "listens" for any returning echoes and back transforms them to electric voltages that are digitized by the transceiver and recorded, typically on the computer hard drive [12;13;14;15].

The main considerations when selecting an echo sounder are frequency, beam width, and type of transducer. Transducer configurations can be single-beam, dual-beam, split-beam or multi beam[16;17;18;19;20]. A single-beam system provides no information on target location, thereby precluding direct estimation of target strength (TS). Target strength distributions must be estimated statistically when using single-beam systems. To address this problem, we develop the algorithm and applied it to detect and quantify the receiving signal using single beam echo sounder.

II. METHODOLOGY

A. Development of Sonar Equation for Single Beam

The sonar equation deals with working relationship that tie together the effects of the medium, the target, and the equipment. For single beam echo sounder, the received signal are computed using

$$TS = 20 \log (counts) - SL - RS + PS + TL + TVG \quad (1)$$

where TS is target strength, SL is source level, RS is receiving sensitivity, PS is power setting, TL is transmission loss due to absorption and geometrical spreading of acoustic wave, and TVG is time varied gain.

$$Counts = DN / 255 \quad (2)$$

DN is digital number of signal with 8 bit sampling.
To calculate the beam pattern, the equation 1 is developed by

$$TS = 20 \log (Count) - SL - RS + PS + TL + TVG + Ce + C \quad (3)$$

$$Ce = 10 \log (c \, \tau \, \psi / 2) \quad (4)$$

c is sound speed, τ is pulse width, ψ is equivalent beam angle for volume backscattering and C is correction factor for acoustics instruments.

To calculate the fish abundance, the volume backscattering (SV) is calculated by

$$SV = 20 \log (Count) - SL - RS + PS + TL + TVG + Ce + C + V \quad (5)$$

where V is volume sampling of acoustic beam

$$V = c \times \tau/2 \times \psi \times R^2 \quad (6)$$

R is range.

Algorithm design for single beam acoustic processing is shown in Figure 1.

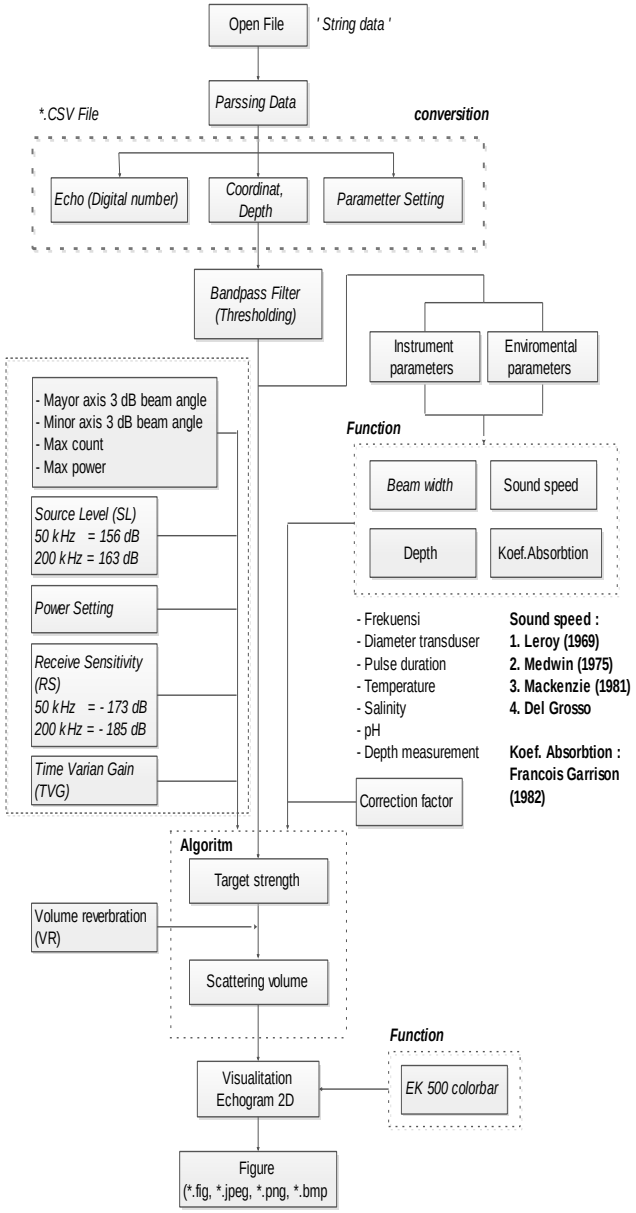


Fig. 1. Algorithm design for single beam acoustic processing

III. Results and Discussion

A. Acoustics Calibration

Scientific echo sounders must be calibrated at regular intervals to ensure consistently high data quality. Calibration involves measuring volume backscattering (*SV*) and TS of standardized (known TS) copper or tungsten carbide spheres located on the main axis and in the far field of the transducer (Figure 2). Calibration should be done under the conditions and field settings of the survey to provide whole-system calibration that combines source-level and receiver sensitivity into one correction factor.

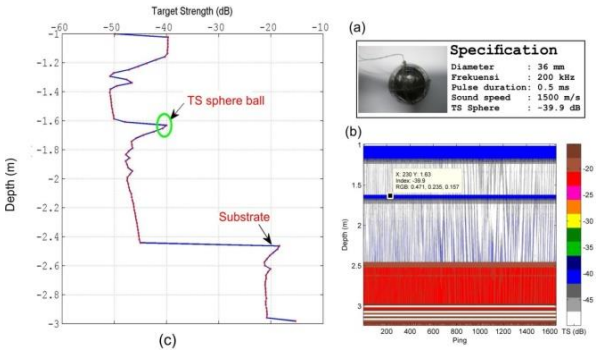


Fig. 2. Acoustic calibration using sphere ball

TABLE I. Comparison Using Measurement and Theoretical Value

		Measurement	*Theory*
N	valid	1654	1654
	missing	0	0
Mean		-39.9	-39.9
Median		-39.9	-39.9
Modus		-39.9	-39.9
Std. Deviation		0.13	0.00
Variance		0.02	0.00
Minimum		-40.2	-39.9
Maximum		-39.7	-39.9

B. Application of Acoustic Algorithm

The application of algorithm using single beam sounder are compared using another system of split beam method for underwater vegetation, fish and seabed. Figure 3 show the backscatter intensity (Sv) of seagrass. Table 2 shows data comparison using this system.

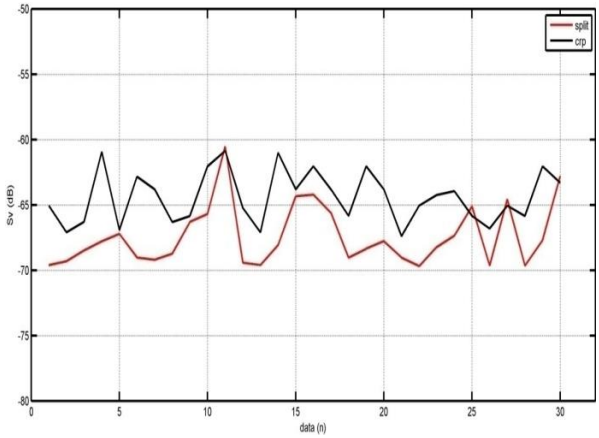


Fig. 3. Backscattter intensity of seagrass (rea line for single beam, black line for split beam).

Table 2 shows the backscattering intensity for single beam ranges from -60,5 dB to -69.7, for split beam ranged from -60,8 to -67,40 dB with the average are -67,4 dB and -66,4dB, respectively. Figure 4 shows the backscatter intensity from coral reef and Table 3 is the statistical vale of the results. Backscatter intensity for coral reef using single beam ranged from -31 dB to -33 dB, while using split beam ranged from -28 dB to -35 dB.

TABLE II. COMPARISON SINGLE BEAM AND SPLIT BEAM FOR SEAGRASS DETECTION

		Single beam	Split beam
N	valid	30	30
	Missing	0	0
Mean		-67.4	-64.4
Median		-68.2	-64.6
Std. Deviation		2.33	2.01
Variance		5.44	4.03
Minimum		-69.7	-67.4
Maximum		-60.5	-60.8

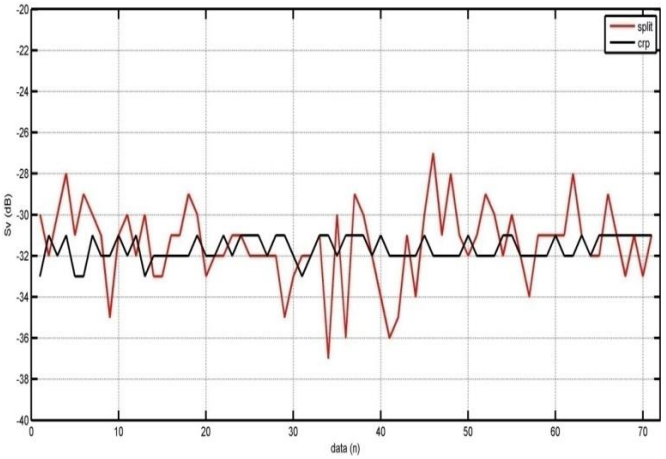


Fig. 4. Backscatter intensity of coral reef using single and split beam

TABLE III. COMPARISON OF SINGLE BEAM AND SPLIT BEAM FOR CORAL REEF DETECTION

	Single beam	Split beam
N valid	70	70
Missing	0	0
Mean	-31.7	-31.4
Median	-32.0	-31.0
Mode	-32.0	-32.0
Std. Deviation	0.69	1.61
Minimum	-33.0	-35.0
Maximum	-31.0	-28.0

Figure 5 shows backscatter intensity from mud bottom using single beam and split beam with the statistical value in Table 4.

Table 4 shows backscatter intensity for mud bottom using single beam ranged from -30.0 dB to -28.4 dB and using split beam ranged from -32.3 to -25.1 dB.

Figure 6 shows the backscatter intensity from sand using single and split beam. Range of intensity from -19.6 dB to -19.8 dB using single beam and -25.1 dB to -16.7 dB (Table 5).

Figure 7 and Table 6 shows the comparison of backscatter intensity using single and split beam for fish. The backscatter intensity of fish range from -58.3 dB to -49.0 dB for single beam and for split beam range from -58.3 to -45.8 dB.

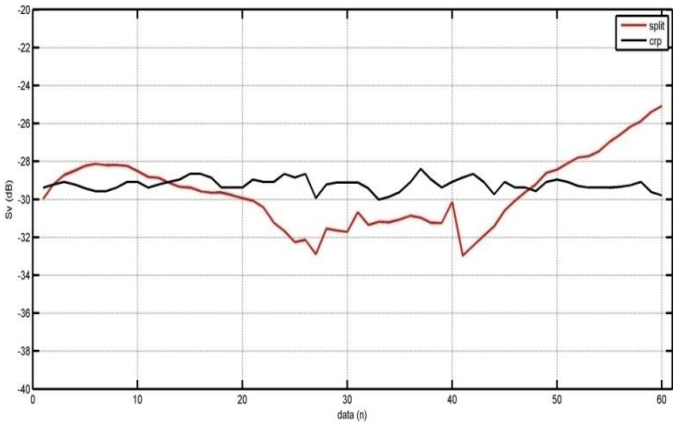


Fig. 5. Backscatter intensity of mud bottom using single and split beam

TABLE IV. COMPARISON SINGLE BEAM AND SPLIT BEAM FOR MUD BOTTOM

		Single beam	Split beam
N	Valid	60	60
	Missing	0	0
Mean		-29.2	-29.6
Median		-29.2	-29.6
Std. Deviation		0.33	1.86
Variance		0.1	3.5
Minimum		-30.0	-32.3
Maximum		-28.4	-25.1

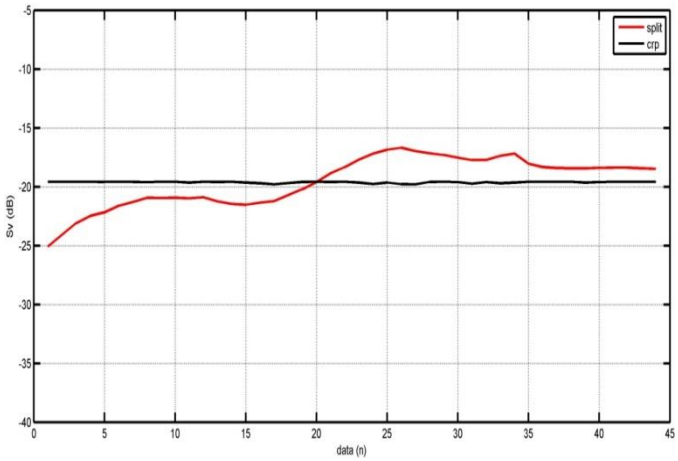


Fig. 6. Backscatter intensity from sand bottom

TABLE V. COMPARISON OF SINGLE AND SPLIT BEAM FOR SAND BOTTOM

		Single beam	Split beam
N	Valid	44	44
	Missing	0	0
Mean		-19.6	-19.5
Median		-19.6	-18.5
Std. Deviation		0.06	2.12
Variance		.004	2.498
Minimum		-19.8	-25.1
Maximum		-19.6	-16.7

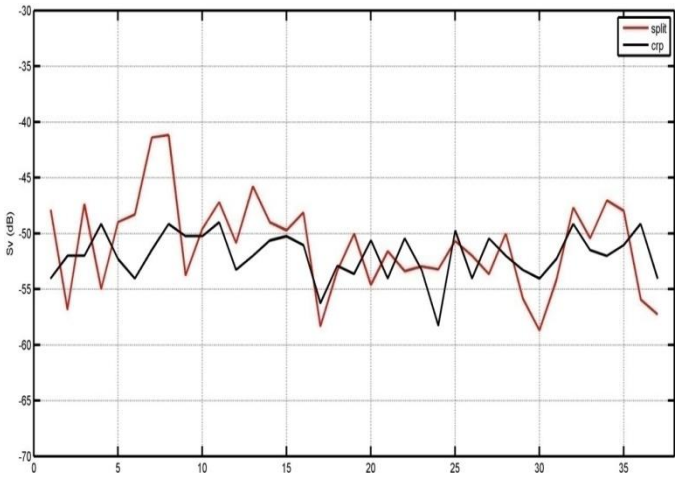


Fig. 7. Backscatter intensity from fish using single and split beam

TABLE VI. COMPARISON OF SINGLE BEAM AND SPLIT BEAM FOR FISH

		Single beam	Split beam
N	Valid	37	37
	Missing	0	0
Mean		-51.9	-51.4
Median		-52.0	-51.5
Std. Deviation		2.1	2.9
Variance		4.4	8.5
Minimum		-58.3	-58.3
Maximum		-49.0	-45.8

IV. CONCLUSION

We had developed algorithm of processing signal using single beam echo sounder. The application of this algorithm had applied to detect underwater objects such as coral reef, fish, seagrass, mud and sand bottom. We also compare the single beam echosounder result with established split beam acoustics method. From the result of comparison we conclude that backscatter intensity measured using developed algorithm using single beam is nearby the established acoustic system using split beam. Future work is to examine this algorithm in the real condition in ocean field, simultaneously with split or multi beam method. For quantitative purpose, single beam acoustic is easy to operate and classification of detected underwater target is possible.

ACKNOWLEDGMENT

The authors would like to thank Ocean Acoustics Laboratory of Department of Marine Science and Technology Faculty of Fisheries and Marine Sciences of Bogor Agricultural University (IPB) research members and students for data collection.

REFERENCES

[1] Au, W. W. L., and Benoit-Bird, K. J. (2008). "Broadband backscatter from individual Hawaiian mesopelagic boundary community animals with implications for spinner dolphin foraging," J. Acoust. Soc. Am. 123,
2884–2894.

[2] Benoit-Bird, K. J., Gilly, W. F., Au, W. W. L., and Mate, B. (2008).

[3] "Controlled and in situ target strengths of the jumbo squid Dosidicus gigas and identification of potential acoustic scattering sources," J. Acoust. Soc. Am. 123, 1318–1328.
Iida, K., Takahashi, R., Tang, Y., Mukai, T., and Sato, M. (2006).
"Observation of marine animals using underwater acoustic camera," Jpn.
J. Appl. Phys., Part 1 45, 4875–4881.

[4] Jones, B. A., Lavery, A. C., and Stanton, T. K. (2009). "Use of the distorted wave Born approximation to predict scattering by inhomogeneous objects: Application to squid." J. Acoust. Soc. Am. 125, 73–88.

[5] Kang, D., Iida, K., Mukai, T., and Kim, J. (2006). "Density and sound speed contrasts of the japanese common squid Todarodes pacificus and their influence on acoustic target strength," Fish. Sci. 72, 728–736.

[6] Kang, D., Mukai, T., Iida, K., Hwang, D., and Myoung, J. (2005). "The influence of tilt angle on the acoustic target strength of the Japanese common squid (Todarodes pacificus)," ICES J. Mar. Sci. 62, 779–789.

[7] Kaipio, J. & Somersalo, E. 2005 Statistical and Computational Inverse Problems, Springer Science+Business Media, New York.

[8] Lavery, A. C., Chu, D., and Moum, J. N. (2010). "Measurements of acoustic scattering from zooplankton and oceanic microstructure using a broadband echosounder," ICES J. Mar. Sci. 67, 379–394.

[9] Lavery, A. C., and Ross, T. (2007). "Acoustic scattering from doublediffusive microstructure," J. Acoust. Soc. Am. 122, 1449–1462.

[10] Lavery, A. C., Stanton, T. K., McGehee, D. E., and Chu, D. (2002). "Threedimensional modeling of acoustic backscattering from fluid-like zooplankton," J. Acoust. Soc. Am. 111, 1197–1210.

[11] Lawson, G. L., Wiebe, P. H., Ashjian, C. J., Chu, D., and Stanton, T. K. (2006). "Improved parameterization of antarctic krill target strength models," J. Acoust. Soc. Am. 119, 232–242.

[12] Lee, W. J., Lavery, A. C., and Stanton, T. K. (2010). "Interpretation of the compressed pulse output for broadband acoustic scattering from inhomogeneous weakly scattering objects," J. Acoust. Soc. Am. 128, 2460.

[13] Medwin, H., and Clay, C. S. (1998). Fundamentals of Acoustical Oceanography (Academic Press, San Diego, CA), pp. 348–401.

[14] Mooney, T. A., Lee, W.-J., and Hanlon, R. T. (2010). "Long-duration anesthetization of squid (Doryteuthis pealeii)," Mar. Freshwater Behav. Physiol. 43, 297–303.

[15] Urick, Robert J., Principles of Underwater Sound 3rd Ed. (McGraw - Hill, New York), (1983).

[16] Simmonds, J., and D. MacLennan. 2005. Fisheries acoustics: theory and practice. Blackwell, Oxford, UK.

[17] Stanton, T. K. 2009. Broadband acoustic sensing of the ocean. Journal of the Marine Acoustical Society ofJapan 36:95–107.

[18] Stanton, T. K., D. Chu, J. M. Jech, and J. D. Irish. 2010. New broadband methods for resonance classification and high-resolution imagery of fish with swim bladders using a modified commercial broadband echosounder. ICES Journal of Marine Science 67:365–378.

[19] Stanton, T. K. & Chu, D. 2008 'Calibration of broadband active acoustic systems using a single standard spherical object', Journal of the Acoustical Society of America, vol. 124 (1), July 2008.

[20] Quintino, V., Freitas, R., Mamede, R., Ricardo, F., Rodrigues, A. M., Mota, J., Pe´rez-Ruzafa,A´ ., and Marcos, C. 2010. Remote sensing of underwater vegetation using single-beam acoustics. – ICES Journal of Marine Science, 67: 594–605.

Inter-organizational Workflow for Intelligent Audit of Information Technologies in terms of Entreprise Business Processes

Meriyem Chergui
ENSEM, Hassan II University
LISER-EAS
Casablanca, Morocco

Hicham Medromi & Adil Sayouti
ENSEM, Hassan II University
LISER-EAS
Casablanca, Morocco

***Abstract*—IT governance is critical to the success of Enterprise governance by providing effective, efficient and measurable improvements in business processes by ensuring that information technologies are in line with business objectives. Consequently, this paper provides an intelligent solution to audit Information System Business processes using the IT Governance Framework COBIT. The particularity of this solution is the use of Inter-organization Workflows (IOW), Multi-agent System and semantic web. In fact Inter-Organizational Workflow is used to cooperate autonomous, heterogeneous and distributed organizations processes to reach a common goal. In this paper case the goal is the dynamic alignment of every Business Process with the convenient Information System component and this through a permanent interaction with different stockholders. Multi-agent Systems (MAS) are known as the natural solution for IOW modeling since they provide dynamic modification and execution of adaptive processes. In addition, MAS have the ability to describe distribution and coordination of IOW organizations in micro and macro level, with high level communication protocols. As for the semantic web, the proposed IT Governance IOW based on COBIT, has the principal role to match Enterprise real Business Goals with COBIT Business goals , so the use of the semantic web is a way to share business terminology and avoid semantic conflict for a correct and efficient Audit operation.**

Keywords*—*Inter-organizational Workflow; COBIT; Audit; Information System; IT Governance; Business Processes; Multi-Agent System; Semantic Web; Ontology

I. INTRODUCTION

Highlight the competitiveness and cost-effectiveness ratio implies increased confidence in information technologies which are becoming an essential component of business strategy. The automation of business functions dictates the incorporation of most powerful control mechanisms not only in computers and networks but in Business Processes, Human Resources and Services as well.

Indeed, many successful enterprises recognize the potential benefits of the Information technologies and understand in the same time how to manage the risks associated with their implementation by the use of IT Governance Frameworks.

The idea of this work is to propose a solution to control and govern IT and Business Processes (BP) in a parallel intelligent and interactive way, taking the benefit of COBIT, the referential framework of Information Systems Governance. The solution also avoids the high cost of Audit missions and this by interfacing it to any kind of Information System (any technology, any dimension, any architecture…). potential users evaluate permanently their Information technologies in terms of Business Processes [1].

To implement such solution, the proposition was the use of Inter-Organizational Workflow able to cooperate many organizations (Information System components) to achieve a comment goal: Audit Operation in COBIT way, which consists on :

- IT alignment with business
- Responsible use of OT Resources
- Appropriate IT risk management

In fact, IOW is a technical model helping heterogeneous and autonomous Enterprises/organizations to put in common their respective BP and skills in order to produce a global cooperative service [2]. IOW have three additional aspects from the classical Workflow:

- The distribution process organizations.
- The autonomy of organizations: each individual organization takes decisions regarding the conditions of cooperation,
- The heterogeneity of organizations to cooperate: this relates to the differences in terms of models and systems.

It's the reason why an IOW context was chosen to deal with simultaneous audit of different components of an Information system.

Ontologies are the key of the semantic web, and they are used in many fields of Computer Sciences for automatic processing, interaction and interoperability of machines. There are many definition of ontology, the most common is that ontology includes or implies a certain view of the world with respect to a given domain; this view is often designed as a set of concepts [3]. IOW can also get the benefit of ontologies, in this solution, its use is necessary for the understanding of the

common terminology to avoid semantic conflicts and to ensure the right matching of Business objectives.

At least, the use of Multi-agent System to implement the IOW is justified by the theoretical background this technology propose to deal with heterogeneity, autonomy and distribution constraints of IOW. MAS also support ontologies through communication protocol.

This article is organized as follows. Section 2 presents IOW and its specificities and justify it use in the Audit context. Section 3 gives an overview of Multi-agent System and Web Semantic. Section 4 talks about COBIT as IT Governance Framework and define the relations between its different components. Section 5 presents the organizational model and the global architecture. Section 6 is devoted to the mediation layer architecture. Section 7 shows the used ontology and extracts as example the case of COBIT processes DS5done in Protégé 4.3 platform. Section 8 presents a MadKit simulator of the IT Governance IOW. Section 9 concludes the paper.

II. Why An Inter-Organizational Workflow?

A. Workflow and Inter-organizational Workflow

A workflow in general is the total or partial automation of business process execution, execution during which documents, information tasks from one participant to another to perform specific activities according to predefined rules**.**

There are many kinds of workflows namely:

- **Administration Workflow**: [4] devoted to manage administrative procedures whose rules of conducts are established and known by everyone in the company.
- **Production Workflow**: [5] devoted to manage the production process in the company.
- **Collaboration Workflow**: [6] devoted to manage awareness and group collaboration in a project of creative work
- **Ad-hoc Workflow**: [7] is a class of workflows for specific situations where the flow logic to be followed is set during execution. It forms a hybrid solution collecting characteristics administration, production, and collaboration

The interested on these kinds of Workflow will find in the references more details about them the advantages and drawbacks of every one.

Inter-organizational Workflow: is an extension of the classical Workflow aiming at cooperating between heterogeneous and autonomous organizations. The reason why it was chosen as a workflow model for this Audit solution

B. Interoperability in Inter-Organizational Workflows

There are many forms of interoperability in an IOW:

- **Capacity sharing** (static context): structural cooperation among organizations with a well-established infrastructure among pre-defined partners in conception phase. Involved organizations, in this case are engaged in a long-term cooperation and their workflows (business processes) are interdependent [8].

- **Chained execution**: modeling a global workflow into several disjoint workflows executed sequentially. Each partner is responsible for a part of the workflow. Once this part is executed, the partner transfers the stream to the next partner. Not in a parallel way.

- **Subcontracting**: allowing to a main partner to delegate the implementation and coordination of part of its workflow to other partners. Workflow control is hierarchical; the partner sees the subcontracted workflows as atomic while they may have complex structures at the running level.

-**Loose IOW** (dynamic context): occasional and opportunist cooperation, without structural constraints, where the partners involved and their number are not pre-defined. Workflows must be increased by a structure of interactions to allow communication between the different partners and the correct execution instances. Interaction is achieved through asynchronous communication and is based on the flow of messages between local partner's workflows.

C. IT Governance Inter-Organizational Workflow

This article is about Information system Audit context which consists on evaluating the adequacy of every Business Process in the Company in terms of existing parts of the IS. In fact, nowadays IS are more and more complex and open to World Wide Web and new network technologies constraints. So , for this problematic the most adaptive interoperability form for the IOW is the Loose scenario, since sub-IS should not obligatory be known in advance and be interconnected and every part has his own objectives and participate in the same time to the global goal achievement.

III. Sementic Web and Multi-Agent Systems

A. IOW modeling with MAS and Semantic Web

The combination of Multi-agent system and semantic Web are widely used for modeling system coordination [9]. It seems to be appropriate to describe the coordination of IOW as a dynamic system aiming at finding "supply service for a demand service" and adopting the negotiation between partners. In fact, agent technology is a custom frame for IOW abstraction: it resolves its constraint of distribution, heterogeneity, autonomy and flexibility:

-Autonomy: every organization of the IOW can be encapsulated in an Agent as autonomous entity having its intentions goals and resources and able to be executed alone or in an environment, depending on the context.

- Distribution: IOW is a distributed context and MAS includes specific architecture, communication protocols and languages to support this constraint.

-Heterogeneity: Agent technology allows communication and interaction between heterogonous agents through Agent-Communication-Languages (ACL). It also provides synchronous and asynchronous ways of communication depending on the agent localization and constraints.

MAS offer many Meta-Models to cover the organizational aspect of Workflow. It also covers the scalability and security worries in loose IOW context.

As for the semantic Web which is the collaborative movement of W3C providing a model that allows data to be shared and reused across applications, enterprises and groups of users [10]. It helps to represent shared business terminology of the IOW in a formal way to solve semantic conflicts in the one hand and to define properly services (supply and demand) in the other hand.

The best representation of semantic web on MAS context is the use of ontology recognized in communication protocol of agents.

B. *Ontlogies conceptualisation*

As defined before, ontology includes or implies a certain view of the world with respect to a given domain; this view is often designed as a set of concepts such as entities, attributes, processes…etc.

It can take different forms but it necessarily includes a vocabulary of terms and specification of their meaning.

To define ontology four points are essential[11]:

- Ontology type : there five types of ontologies namely :
  - Domain ontology
  - Generic ontology
  - Problem resolution ontology
  - Application ontology
  - Representation ontology
- Properties definition : it's the definition and classification of concepts and their properties (simple or complex)
- Relation "is a": it's called "**subsomption**" which define a generalization relationship.
- Author relations: it concerned conceptions relations other than "is a" such as "part of", "primitive of"..etc.

C. *Ontologies Editors*

There are many ontologies editors namely:

Protégé [12]: graphical environment for ontologies development based on hierarchical knowledge model (classes →attributes → properties). It's one of the most used editors regrouping a wide community of users , it has compatibility with OWL reference , Knowledge base management , ontologies visualizations, alimentation and fusion.

OILED [13]: it's also based on classes' hierarchy, it provides roles specialization, properties test but it's limited to the construction of OIL ontologies example.

OntoEDit[14] : it's an owner solution based on hierarchical concepts , able to express axioms but it's not reliable since it's limited to a lexical comparison of terms .

The most adaptable ontology to the proposed IT Governance IOW is domain ontology to match BP Demand and BP supply in IT Governance Domain. In this article, Protégé 4.3 is used for modeling this ontology in OWL-S supported by FIPA-ACL as this solution Agent communication languages

IV. COBIT: IT GOVERNANCE FRAMEWORK

IT Governance is a structure of relationships and processes to control the enterprise to achieve its objectives by generating value while finding the right balance between risk and benefits of IT and processes. It could not be efficient without a referential framework giving best practice. this article is based on COBIT 4.1 (Control Objectives for Information and related Technology Business).

A. *What is COBIT? how to grasp it?*

COBIT [8] is an IT Governance framework developed in 1994 (published in 1996) by ISACA (The Information System Audit and Control Association). It is designed for the control objectives of information technology.

COBIT proposes best practices through a framework by domain and by process. It presents activities in a manageable and logical structure. Its practices focused more on control, less on execution. To optimize IT-enabled investments, ensure service delivery and provide a good measure to face potential risks

For COBIT, as shown in the figure below, every information system can be decomposed into 34 processes, which are divided into four functional areas:

- ✓ Planning and Organization) (10 processes).
- ✓ Acquire and Implement) (7 processes).
- ✓ Deliver and Support) (13 processes).
- ✓ Monitor (4 processes).

These four areas can cover 318 goals with different criterias

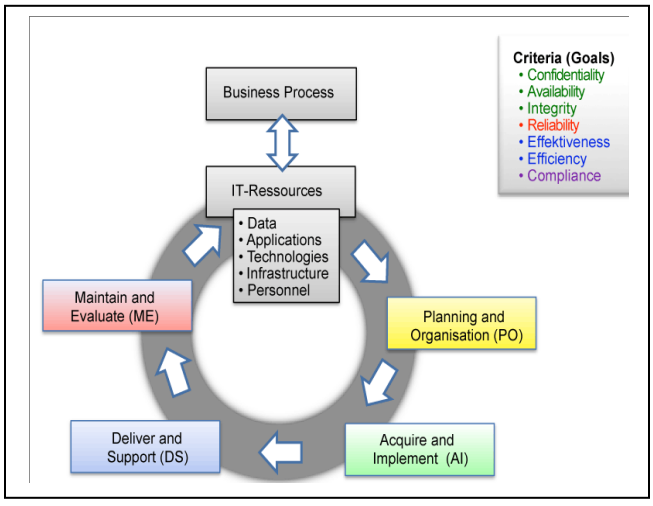


Fig. 1. COBIT Process Model (IT Processes and Doamains).

B. *Mapping Business Goals, IT Goals and COBIT Processes*

COBIT offers variety of components interconnected to guide Audit mission and/or IT Governance procedure.

In fact, COBIT proposes three essential kinds of components namely: Business Goals, IT Goals, and IT Processes. These components cover mainly the totality of possible Goals and processes for an Information System.

-Business Goals [15]: COBIT V4.1 (the used version) proposes 20 Business Goals distributed according to the four pillars of balanced square i.e. customer perspective; financial perspective; Information System Direction (ISD) internal perspective and future or anticipation perspective.

-IT Goals: the 20 Business goals refer to 28 IT goals, themselves related to COBIT process. The same IT Goal can be associated with one or more COBIT process (one of the 34 processes presented before.)

Consequently, COBIT offers to every Business Goals, IT goals, IT processes, Key activities, Controls, Metrics, RACI Chart, etc. These outputs represents recommendations and measures ISD and Top management should consider for better IT governance.

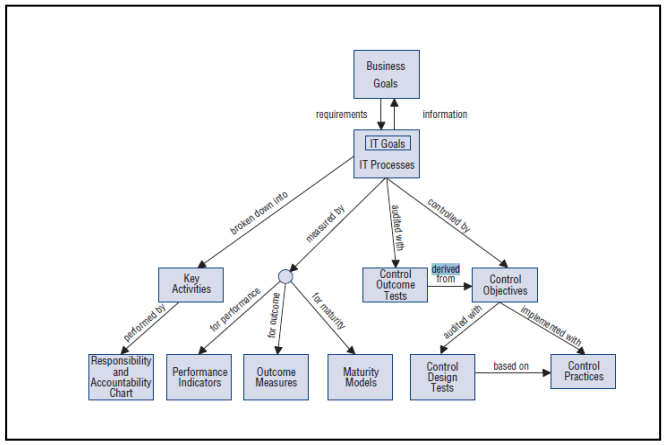


Fig. 2. COBIT Components relationship

the proposed loose IOW architecture is based on the process oriented aspect of COBIT and the "agentification" of its components detailed before. In fact, COBIT provide hierarchy able to be divided between Actors who can take the responsibility of giving a full image of IS business Objective (BO). The added value of this work is the intelligent matching between real Enterprise Business Goals (expressed by users and managers about IT worries and standard (Business Goals of the BSC) proposed in the framework. This matching is the first mission of IOW Agents, and then an Audit operation will be launched as shown in Figure 2.

C. Case Study : DS5 process Goals and Metrics

In this paper, to illustrate the flow of the proposed IT Governance loose IOW, the case study is as following : an IS user evokes an IS business objective about information reliability for top decisions.

This IS BO will be matched with the 9th COBIT Business Objective: "Obtain reliable and useful information for strategic decision making". This 9th BO calls many IT Objectives (see fig3).

To simplify and well clarify the case study for next sections only the example of 9th Business Objective calling the 20th IT Objective will be illustrated:" Ensure that automated business transactions and information exchanges can be trusted."

This 20th IT Objectivecall three IT processes (see fig4) . For the same reason, let's take the exemple of DS5 COBIT process witch concerns "system security insurance". It belongs to "Deliver and Support" Domain includes establishing and maintaining IT security roles and responsibilities, policies, standards, and procedures.

TABLE I. LINKING 9TH BUSINESS OBJECRIVE TO IT OBJECTIVES [9]
TABLE TYPE STYLES

BO N°	List of the correspondents ITO				
9	2	4	12	20	26

a.

TABLE II. LINKING 20TH IT OBJECTIVE WITH COBIT IT PROCESSES [9]

ITO N°	List of the correspondents COBIT IT Processes		
20	PO6	A17	DS5

V. PROPOSED ARCHITECTURE OF IT GOVERNANCE INTER-ORGANIZATIONAL WORKFLOW

A. Organizational Model

Agent-Group-Role (AGR) is a Multi-Agent System Meta model where an agent as an intelligent and communicating entity can play one or more roles through membership in a group or groups without any constraints on its architecture [1].

Based on AGR, the proposed organizational model is organized around the following components:

- Five types of groups represented by an eclipse (Audit, Finding Audit, Finding Auditor, Audited and Auditor)
- Ten roles represented by a circle as every agent has double role in every group (Mediator, SI Connection Server, COBIT Connection Server, IS Workflow Agent, COBIT Agent)
- Communication between agents is represented by arrows.

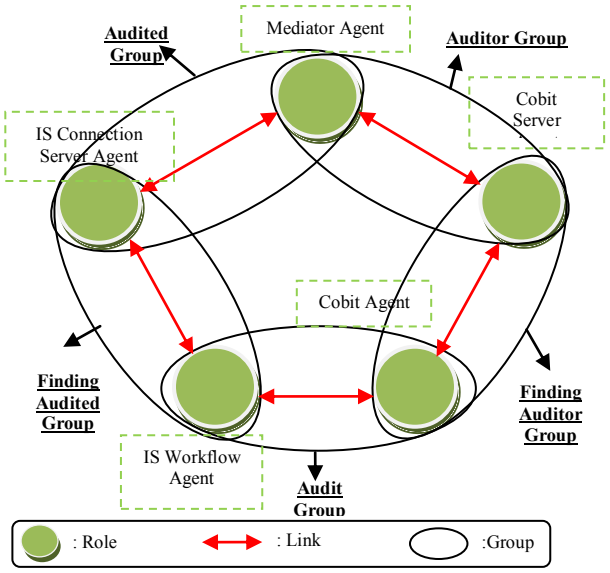


Fig. 3. Organizational Mode

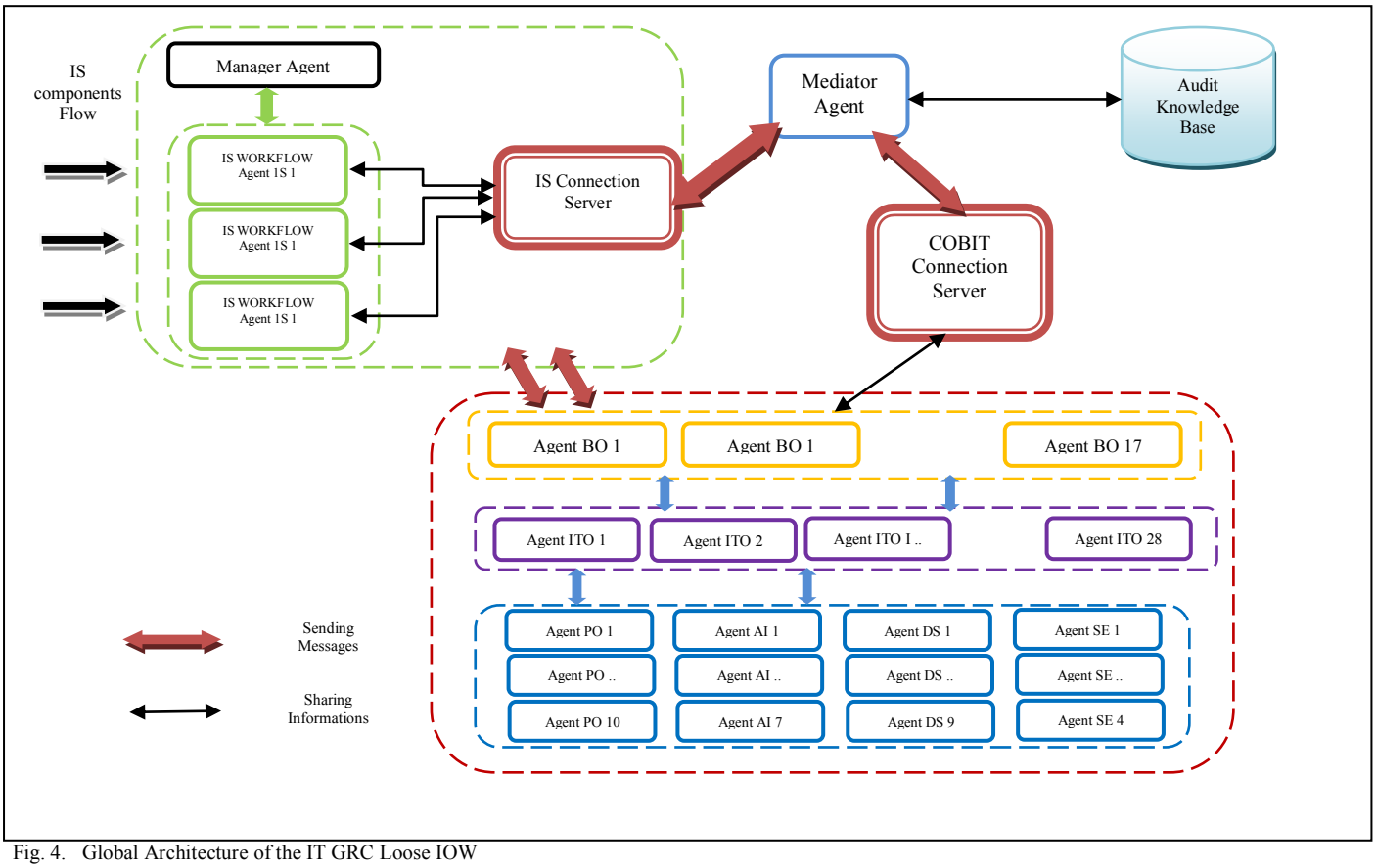


Fig. 4. Global Architecture of the IT GRC Loose IOW

Every part of Information system (application, ERP module, etc..) is encapsulated on an Agent having IS Workflow Agent role. In the same way every COBIT Business Goal or Business Objective is encapsulated in an Agent having COBIT Agent.

Connection server submits an audit request about a business objective; it allows the mediator agent to return the identity of the appropriate COBIT agent in Audited Group.

IS Workflow Agent and COBIT Agent, after getting each other identities from connection servers negotiate the more priority COBIT process to implement; the RACI matrix, the key metrics and the maturity model to follow in Audit Group.

IS Workflow Agent or COBIT Agent interact with a connection server (COBIT or IS) from which they get requested partner identity in Finding Audited Group and Finding Auditor Group.

Connection server via a mediator Agent (recording COBIT Agents capabilities), release the appropriate COBIT process (offered by COBIT Agent) in Auditor Group.

B. Global Architecture

This architecture is essentially based, on:

- Loose WIO literature model [19]
- Workflow reference Architecture [20]
- Agentifacation of COBIT 4.1 components relationship.

It contains the following Agents:

IS Workflow Agent: Agent encapsulating a part of the IS and launched by stakeholders requests about the audit of one or many business processes of the system.

Manager Agent monitors and controls the running of IS Workflow Agents.

COBIT Agent is the auditor agent who broadcasts services throw the COBIT Connection Server. Once into contact with an IS Workflow Agent, COBIT Agent calls other agents: IT Objective Agents and COBIT IT Process Agents to audit the IS BO in COBIT framework way.

Connection Server Agent is responsible for publishing Workflow IS Agents requests and getting convenient COBIT Agents from **Mediator Agent**.

Mediator Agent: it's a yellow pages Agent which publishes COBIT Agents offered services and requests made by the IS Workflow agents. the next section will be devoted to it.

VI. Mediation Layer : Business Objectives Matchmaking

A. Medaition Agent proposed architecture

There are principally three kind of mediator Agent: Matchmaker, Broker [21] and Facilitator [22]. The difference between a Matchmaker and a facilitator is that the second one intermediate transaction and the first one links provider with

requester by identities exchanging, then they communicate directly. As for a Broker, he gets delegated services with preferences from the requester, asks the provider for results and sends directly this result to the requester.

In this paper case a Matchmaker agent is necessary so as to link between IS Workflow Agent and COBIT agent and let them exchange audit information directly in Audit Group without interfering. This to simulate a real Audit operation consisting on interviewing IS user to propose convenient recommendations.

The role of the Matchmaker in the WIO is to find convenient partner (COBIT BO) for every IS BO instance.

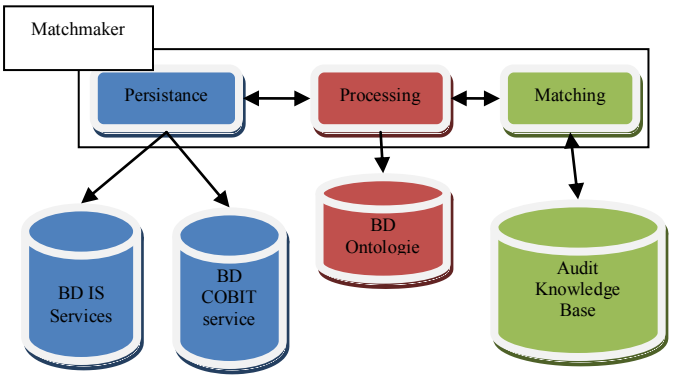


Fig. 5. Mediator Agent Architecture

There are three parts In the Matchmaker Agent:

-Persistence: it's a dynamic layer responsible for COBIT Agents offered services saving and optionally IS Workflow Agents demands saving. This layer communicates with COBIT Services Data base and IS Objectives Database (optional).

- **Processing**: it's a dynamic layer where Audit ontology en OWL-s format is created and saved. In fact it's the hierarchical description of demand services and supply ones. This layer communicates with an ontology Data-base, in this paper, Protégé save ontologies by default in a web localization; so data-base could be replaced with an XML file containing ontologies URL.

Matching: it's the comparison and link between a demand and convenient offers; it's a return of convenient COBIT Agent Addresses to IS Workflow Agent. The comparison is based on the Audit Ontology defined in Processing layer and need an algorithm to filter offers (not yet done). This is the intelligent layer of the Matchmaker agent and it's linked to Knowledge Base of Audit operation.

B. AUML Sequence Diagram for BP Matchmaking

To illustrate the intelligent matching of IS Business Objectives and COBIT Business Objectives by mediation entities, the following AUML Sequence Diagram is proposed (see Fig.6.):

IS Workflow Agent sends the IS BO (demanded service) to IS Connection Server. IS Connection Server confirms the demand reception by an acknowledgement to **IS Workflow Agent and Send**. Then, it sends demand service to **Matchmaker Agent**.

At the same time, **COBIT Agent** send throw its own **Connection Server** COBIT BO (supplied services).

Matchmaker Agent saves the service coming from **IS Workflow Agent** and all supplied services (Persistence layer of Matchmaker Agent see Fig5), processing every service via existing ontologies and compares them (matching layer). In the next section these two layers roles will be detailed.

Once the Matchmaker Agent find convenient supplied service: COBIT BO for the demanded service IS BI; it **sends COBIT Agent** Address to IS Workflow **Agent.**

To conclude, the Mediation entity in this paper is a Matchmaker Agent able to save "Supply and Demand" Business Objectives, define them throw Audit ontology and match IS Business Demand with the corresponding COBIT Business Objective.

This matching simulates the Audit activity first step: indentifying the problematic IS Business Objectives and its measure in COBIT Framework. From this step, IT Objectives and IT Processes of this IS demand can be defined to get as a result recommendations about:

- Activities : list of activities to achieve IT Objective
- Metrics: measures able to quantify IT Processes performance.

Responsibilities chart: repartition of activities among

- Maturity model : degree of IT Process implementation (0: nonexistent -5 optimized)
- IS stakeholders with the following values(**R**esponsible, **A**ccountable, **C**onsulted, **I**nformed)

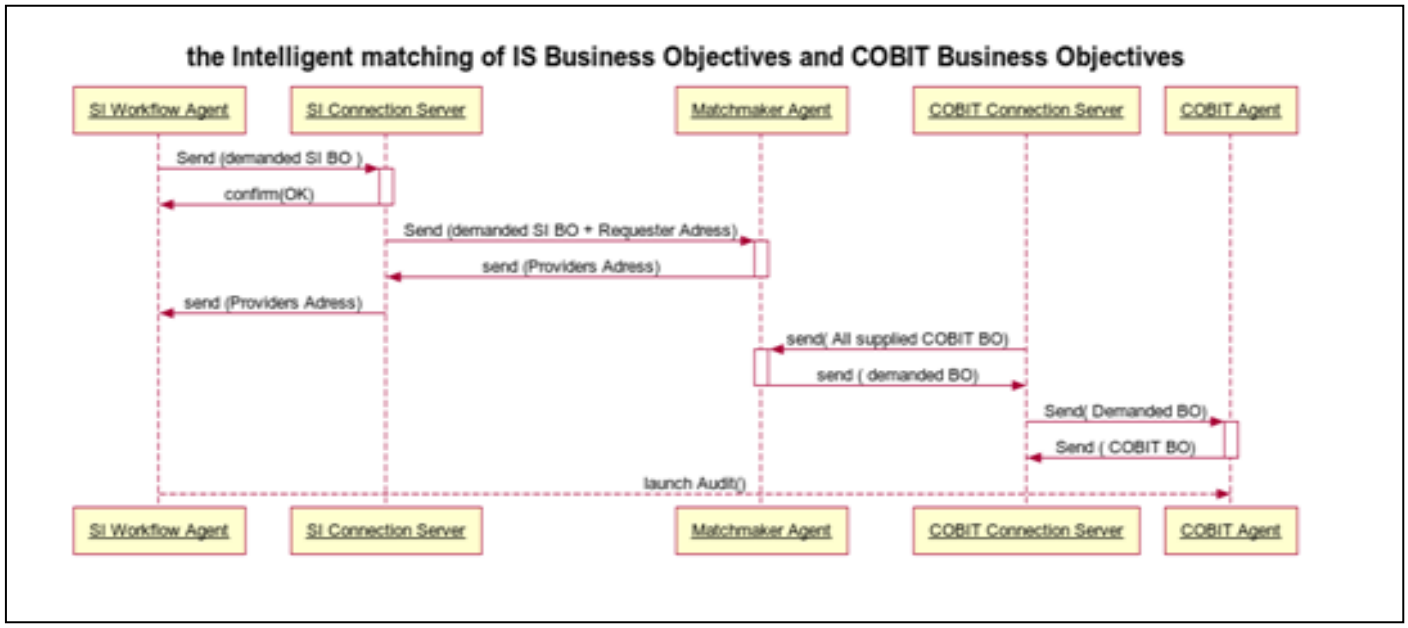


Fig. 6. AUML Sequence Diagram for the Mediation Layer

VII. Audit Domain Ontology : Case Study DS5

A. Audit ontology in OWL

As said before, domain ontology is used in the proposed IT Governance Inter-Organizational Workflow to define IS Business Objectives as demanded services and COBIT Business Objectives as supplied services. The role of this ontology is to understand the common vocabulary of IOW organizations and to allow the Matchmaker to compare and match "demand with supplies".

As a result of ontologies state of art, the "Audit Ontology" of this solution is implemented with Protégé 4.3 in OWL (with Resource Description Framework (RDF) format).

OWL is a widely used web semantic language; it provides many advantages through its hierarchical structure [17], namely:

- Service definition through a process model.
- Attributes detailed description (Inputs, outputs, constraints)
- Support of different structures of service (atomic, simple or complex)
- Set operators default use.
  - Maturity model : degree of IT Process implementation (0: nonexistent -5 optimized)
  - IS stakeholders with the following values(**R**esponsible, **A**ccountable, **C**onsulted, **I**nformed)

This technical choice is in line with the fact that the IT Governance IOW simulator is developed in MadKit 5 platform with FIPA-ACL as Agent Communication language, more details will be given about this point in the next section.

At this stage, this same language is kept as Agent Capability description Language since it supports performatives, and ontologies and offers development flexibility.

Coming back to "Audit Ontology" : the Matchmaker Agent is connected to an ontology Data base. Once it gets the COBIT BO service and /or IS BO service, it calls the ontology, extracts entities and properties and defines the class of each concept of the proposed service, eventually equality, inclusion and difference.

In future works the states of arts of ontologies concepts comparison will be presented and the algorithm to compare concepts in terms of "Audit Ontology" will be implemented.

Other reason to implement OWL ontology is the interoperability of defined services: they could be eventually manipulated as web service for a better reusability and without any environmental or architectural integration constraint.

B. Case study and Ontlogy exemple

In this article Protégé 4.3 is used as ontology editor and "Audit ontology" is based on COBIT 4.1 Business Goals definition.

In facts COBIT Business Objectives are divided into 4 categories of perspectives (Financial Perspective; Customer Perspective; Internal Perspective and Learning and Growth Perspective).

Every perspective contains many Business Objectives; Audit ontology concepts and properties are defined around these BO related to the four Perspectives as shown in the figure bellow.

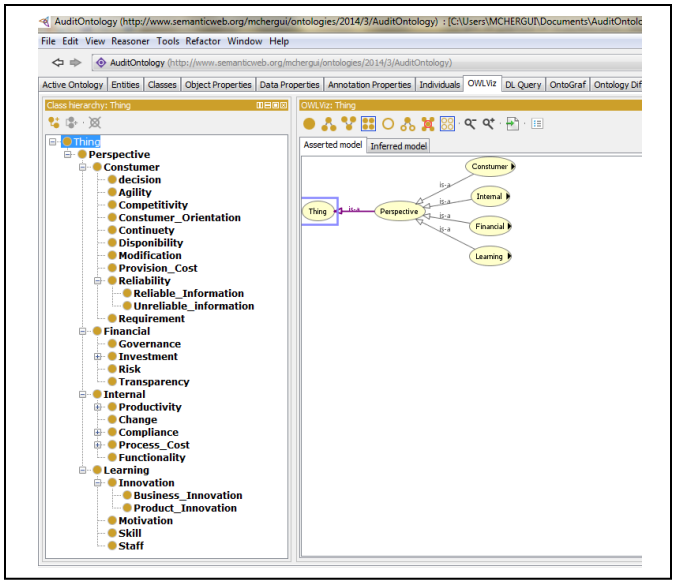


Fig. 7. Audit Ontology overview- Protégé4.3

To illustrate that, let's show the part of Audit ontology about the case study (see Section IV) IS BO = "IS information's reliability for top management decisions".

The key concepts are: reliability, information, and decision.

On Audit ontology, **"Reliability"** and **"Decision"** are sub-classes of **"Consumer"** which is a sub-class of "**Perspective".**

"Reliable information" is a sub-class of "Reliability".

As for object properties: "Reliable information" **is useful for** "Decision" (see Figure 8).

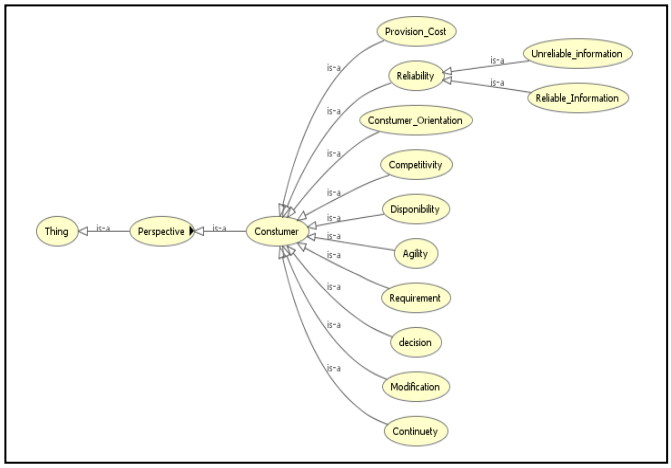


Fig. 8. OWL Viz Asserted model of a part of Audit Ontology (Constumor Perspective concepts)

The generated OWL file arround this part of "Audit Ontology" is as bellow:

```
<rdf:RDF xmlns="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#"
    xml:base="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <owl:Ontology rdf:about="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology"/>
<owl:ObjectProperty
//Object Properties
rdf:about="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#usful_for"/>
// Classes
<owl:Class rdf:about="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#Reliability">
      <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#Consumer"/>
      <rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty
rdf:resource="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#usful_for"/>
          <owl:someValuesFrom
rdf:resource="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#decision"/>
        </owl:Restriction>
      </rdfs:subClassOf>
  </owl:Class>
…..
<owl:Class
rdf:about="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#Reliable_Information">
      <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#Reliability"/>
  </owl:Class>
<owl:Class
rdf:about="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#Unreliable_information">
      <rdfs:subClassOf
rdf:resource="http://www.semanticweb.org/mchergui/ontologies/2014/3/AuditOntology#Reliability"/>
  </owl:Class>
…..
</rdf:RDF>
```

Fig. 9. Portion of Generated Audit ontology OWL file

VIII. Implementation and Discussion

As implementation of the proposed IT Governance IOW, a multi-agent simulator of Audit operation is developed. Bellow technical specificities of this simulator are presented:

A. *MAS Plateforme choice*

As the choice of Multi-Agent platform has a great influence on the design and implementation of MAS, FIPA has produced standards that describe how an agent platform should be. These standards exist to ensure uniform design agents regardless of the platform.

The platform choice is based on the above comparative table [18],

TABLE III. Mult-Agent Platforms Comparaison

Platforms	MAS Types	Agent Model	Methodology	Language
ZEUS	Economic system/ planning scheduling /	Collaborative Agent	Agent , Goal , Task : Zeus agent architecture	Java
MADKIT	Any kind	AGR + adaptive to other models	AALAADIN	Java/
JADE	Simple systems / pedagogic illustration	None / Hybrid agent (JADEX)	None	Java
Agent-Builder	Any kind	BDI	OMT	Java

MADKIT platform was chosen since it is a generic MAS platform supporting AGR model in which the IOW conception

is based. In addition to that, MADKIT can build complex systems, control Agent life cycle and provide a complete layer of Agent communication (asynchrony message / broadcast message, etc).

B. Implementation

The solution is developed with JAVA to ensure system portability and to benefit from APIs for Agent and ontology implementation.

Eclipse IDE is used for java development with MADKIT 5 API.

The following roles of the IOW was developed namely:

- IS Workflow Agent Role
- IS Connection Server Role
- Mediator Role
- COBIT Server Role
- COBIT Agent.

The screenshot below present different agents' execution:

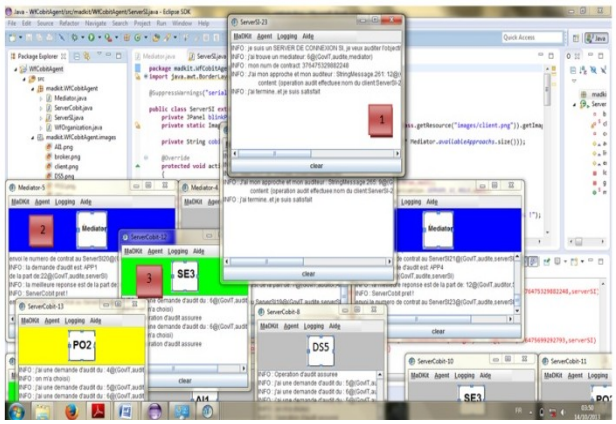


Fig.10. Overview of IT Governance IOW

In this level, a graphic interface was implemented for each Agent and a main class java to simulate the audit operation. a user interface will be proposed later to launch IS Workflow Agent and IS BO input.

The request now is imitated to find the convenient COBIT connection server publishing one of COBIT processes (the choice is based on the mediator matching)

We are working at Mediation entity implementation to integrate Audit ontology, and preparing mediation algorithm for its use.

The screenshot presents:

1: IS connection Server Agent presenting a static business objective, it asks the mediator to find its supply service and wait for a request.

3: COBIT Connection Server publishing its service through the mediator, and waits to be chosen as an auditor.

2: Mediator Agent matches the audited agent with the convenient auditor agent. It sends service title + COBIT Agent address to IS Connection server.

IX. Perspective and Conclusion

The purpose of this paper is to deploy an agent based Inter-organization Workflow to provide permanent and interactive Audit operation of Information systems.

Many literature issues were invoked namely:

- Inter-Organizational Workflows
- Multi-agent System and artificial intelligence
- Mediation entities
- Semantic Web and ontologies

The choice of every issue has an added value for this solution; in fact, Inter-organization Workflows provide the orchestration of heterogeneous components of an IS in an autonomic way.

Multi-agent system insures the intelligent dimension of the solution with high level communication protocol and modeling architecture.

Mediation in MAS gives a theoretical model of matching services among intelligent entities.

Ontologies offer the semantic alignment of stakeholders with COBIT framework vocabulary like experts Audit operation context.

This paper opens many perspectives of this research work namely:

- Audit negotiation operation between IS Workflow Agent and COBIT Agent and detailed architecture of each of them,
- BO Services better description with SOA,
- IOW Intelligent user interface modeling and implementation.
- Simulator amelioration in parallel with proposed architectures.

In fact, the IT Governance IOW role is not only to find the convenient COBIT Business Objectives for IS goals but to negotiate COBIT recommendation and measure the reality of IS alignment with Enterprise Business, so the next main step of this research is to implement the negotiation infrastructure of the Inter organizational Workflow. The Second important point to develop is the Web Service representation of Processes to benefit from semantic web power and to ensure more portability of our platform. Of course, this should be in parallel of modeling and developing ergonomic platform with

References

[1] Chergui, M., Sayouti, A., & Medromi, H. IT Governance through an Inter-Organizational Workflow based on Multi-Agent System

[2] Picard, W., Paszkiewicz, Z., Strykowski, S., Wojciechowski, R., & Cellary, W. (2014). Application of the Service-Oriented Architecture at the Inter-Organizational Level. In *Advanced SOA Tools and Applications* (pp. 125-201). Springer Berlin Heidelberg.

[3] Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.

[4] Lambrinoudakis, C. Kokolakis, S. Karyda, M. Tsoumas, V. Gritzalis, D. and Katsikas, S. (2003): Electronic voting systems: security implications of the administrative workflow. In Proceedings of the 14th International Workshop on Database and Expert Systems Applications.

[5] Muehlberger, R. Orlowska, M.E. and Kiepuszewski, B. (1999) : Backward step: The right direction for production workflow systems. In Proceedings of the Australian Database Conference.

[6] Jiang, P. Mair, Q. and Newman, J. (2003) : Using uml to design distributed collaborative workflows: from uml to xpdl. In Proceedings of the Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises.

[7] Huth, C. Erdmann, I. and Nastansky, L. (2001) : Groupprocess: using process knowledge from the participative design and practical operation of ad hoc processes for the design of structured workflows. In Proceedings of the 34th Annual Hawaii International Conference on System Sciences.

[8] W. M. P. van der Aalst. (1999) : Process-Oriented Architectures for Electronic Commerce and Interorganizational Workflow. Information Systems, 24(9): pp.639-671.

[9] Gandon, F. (2002). Distributed Artificial Intelligence and Knowledge Management: ontologies and multi-agent systems for a corporate semantic web (Doctoral dissertation, Nice).

[10] W3C Semantic Web Activity » [archive], World Wide Web Consortium (W3C), 7 novembre 2011.

[11] Charlet, J., Bachimont, B., & Troncy, R. (2004). Ontologies pour le Web sémantique. Revue Information, Interaction, Intelligence I3.

[12] Noy, N. F., Crubézy, M., Fergerson, R. W., Knublauch, H., Tu, S. W., Vendetti, J., & Musen, M. A. (2003). Protege-2000: an open-source

[13] ontology-development and knowledge-acquisition environment. In AMIA Annu Symp Proc (Vol. 953, p. 953).

[14] S. Bechhofer, I. Horrocks, C. Goble, and R. Stevens. OilEd: a reason-able ontology editor for the semantic web. In Proc. of the Joint German/Austrian Conf. on Artificial Intelligence (KI 2001), number 2174 in LNAI, pages 396–408.Springer-Verlag, 2001

[15] Sure, Y., Erdmann, M., Angele, J., Staab, S., Studer, R., & Wenke, D. (2002). *OntoEdit: Collaborative ontology development for the semantic web* (pp. 221-235). Springer Berlin Heidelberg.

[16] CobiT Steering Committee. (2007). COBIT 4.1. *Rolling Meadow: IT Governance Institute*.

[17] Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, *5*(2), 51-53.

[18] Garneau, T., & Delisle, S. (2002, October). Programmation orientée-agent: évaluation comparative d'outils et environnements. In *JFSMA* (pp. 111-124).

[19] Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Wooldridge, M. J., & Sierra, C. (2001). Automated negotiation: prospects, methods and challenges. *Group Decision and Negotiation*, *10*(2), 199-215.

[20] Aubert, B. A., & Dussart, A. (2002). *Systèmes d'information inter-organisationnels*. CIRANO.

[21] Weiss, G. (Ed.). (1999). Multiagent systems: a modern approach to distributed artificial intelligence. The MIT press

[22] Genesereth, M. R. (1997). An agent-based framework for interoperability. *Software agents*, 317-345.

Performance Evaluation of Private Clouds Eucalyptus versus CloudStack

Mumtaz M.Ali AL-Mukhtar
Departement of Internet Engineering
AL-Nahrain University
Baghdad, Iraq

Asraa Abdulrazak Ali Mardan
Departement of Networks Engineering
AL-Nahrain University
Baghdad, Iraq

***Abstract*—the number of open source cloud management platforms is increasing day-by-day. The features of these software vary significantly and this creates a difficulty for the cloud consumers to choose the software based on their business and scientific requirements. This paper evaluates *Eucalyptus* and *CloudStack*, the two most popular open source platforms used to build private *Infrastructure as a service* (IaaS) clouds. The performance of virtual machines (VMs) initiated and managed by Eucalyptus and CloudStack are evaluated in terms of CPU utilization, memory bandwidth, disk I/O access speed, and network performance using suitable benchmarks. Different VM management operations such as add, delete and live migration are also assessed to determine which cloud solution is more suitable than other to be adopted as a private cloud solution. As a further performance testing, a simple web application has been implemented on the both clouds to evaluate their suitability in web application hosting.**

Keywords—Cloud Computing; CloudStack; Eucalyptus; IaaS; Virtual Machine; Performance Evaluation

I. INTRODUCTION

Cloud computing as a new Internet service concept has become popular to provide a variety of services to users. It is a combination of technologies that have been developed over the last several decades, which includes virtualization, dynamic provisioning, internet delivery of services, grid computing, cluster computing and utility computing [1][2]. According to NIST (National Institute of Standards and Technology), "Cloud Computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [3].

There are three deployment models by which Cloud computing services are delivered: public, private, and hybrid. Public Cloud is a cloud that is made available as ―pay-as-you-go and accessible to the general public such as Amazon Web Services. Private Cloud refers to a cloud infrastructure that is internal to an organization and is not available to the general public. A private cloud's data centers can be on premise and the physical infrastructure is owned and managed by the organization that owns it [4]. Hybrid cloud is a composition of two or more cloud deployment models that are bound together by a standardized or proprietary technology [4].

There are certain legal, political, socio-organizational reasons that may discourage an organization from using public cloud infrastructure for certain kinds of activities, for example processing and storing citizens' private data. There is also the issue of privacy, security, location and ownership of data [4]. Many companies hesitate to use public cloud in which computing resource are shared with other companies. These companies do not have any knowledge of where their applications are run and their data are stored or control access to them [1]. Hence, private cloud infrastructure is considered an appropriate alternative.

Another big reason to increase the interest in setting up and managing private cloud is the SLA. The public cloud providers nowadays provide guarantees on their service levels and when service failures occur, they only offer to refund their customers regarding the infrastructure outages. However, service providers are not inclined to pay penalties of low performance level that would refund customers for loss of business revenue. Cloud providers are not only required to supply correct services but, also, to meet their expectations in the context of performance [5]. Also some software systems and applications require different performance levels, quality of services, reliability, and security, which are generally not guaranteed by a public cloud. Private cloud is an alternative to companies or researchers that need more control over that data [1] [6].

There are many commercial and open source cloud management platforms that are used to build Infrastructure as a service (IaaS) private cloud solution such as Eucalyptus, OpenNebula, and Vmware cloud. However the open source solutions are gaining a lot of popularity and momentum with their features, rapid developing with low investment cost which present a viable option for academic and scientific worlds [7], and enterprises who want first to test the cloud computing solution suitability to their business environment before purchase the thousands dollars commercial solution.

The number of cloud platforms related to a private IaaS cloud is increasing day-by-day. The features of cloud management software vary significantly and this creates a difficulty for cloud consumers to choose the software based on their business requirements.

An example for this problem is choosing platform much suitable for hosting web applications or running high performance computing (HPC) applications, or meeting specific user usage way like users that demand a few virtual machines (VMs) but want to run them for a long period of time with guarantees on high-availability, or scientists requiring a large number of resources to conduct actual calculations and analyses of data. The advent of several Open Source Cloud platforms guarantees the performance and uptime. It is not easy for non-expert users to choose from the different platforms without comprehending the characteristics and advantages of each of this platform [6].

As a consequence, performance evaluation of cloud computing platforms has been receiving considerable attention by both the users and service providers as a prominent activity for exploring the limitations of the cloud platforms and improving service quality, infrastructure planning, and making a wiser selection of the platforms. In addition cloud management software vendors can develop and include additional features to their software by fixing the platforms bugs and including the missing features.

The rest of this paper is organized as follows: Section II presents related work. Section III describes the test environment and methodology. Section IV covers the performance evaluation of Eucalyptus and CloudStack VMs. Section V assesses VMs startup and release time. Section VI evaluates live migration of VMs. Section VII presents response time of web application in the both clouds. Finally conclusions are drawn in the last section.

II. Related Work

Many studies have been conducted to evaluate performance of open source cloud platforms such as Eucalyptus, Opennebula and Nimbus. However these research papers did not perform a complete performance analysis of the cloud platform, and compare only the architectures and features of the cloud management platforms. Nevertheless a little work has been done yet to evaluate CloudStack due to the fact that it is relatively new.

De Sousa et al. [1] evaluated Eucalyptus VMs considered processing and disk I/O performance only while in [6, 7, 8, 9, 10, 11] authors brought out an overview of architectures of open source platforms and comparison of their general features and Characteristics. Mao and Humphery [12] investigated the performance of VM startup and release time of public clouds. However, D. Steinmetz, et al. [13] evaluated performance and studied VM launch time of Eucalyptus and OpenStack but performance benchmarking was not specific and gave a general view of performance. While Folgar, et al. [14] evaluated performance of CloudStack primary storage disk I/O only.

Differently from previous works, this paper evaluates performance of Eucalyptus and CloudStack clouds VMs covering versatile parameters including performance of cloud management platform considering add, delete and live migration of VMs. Performance of VMs in term of CPU utilization ,memory bandwidth, disk I/O speed and networking performance is rated as key point of our evaluation. Also the performance of VMs are compared with regard to bare-metal or traditional IT infrastructure.

III. Test Environment and Methodology

CloudStack 4.1 cloud with one zone, pod and cluster has been deployed using 3 identical physical servers. One server is used as a management server including primary and secondary storage and the other two servers are used as host machines. Eucalyptus 3.2 cloud with one cluster has been deployed using 3 identical physical servers each. One server is used as a cloud controller (CLC) including cluster controller (CC) and Walrus storage and the other two servers are used as node controllers (NCs). Our servers are Intel^{R} Core^{TM} i5-2410M CPU 2.3GHz, 4GB RAM, 500GB SATA Hard Disk and 100MB Ethernet interface. Centos 6.3 (final) is installed on each server as native OS. CloudStack with NFS storage configuration is deployed while Eucalyptus is deployed with local storage configuration. Each host in both clouds is configured with kernel-based virtual machine (KVM) as a hypervisor.

In order to evaluate and analyze VMs performance of both clouds, we have employed a number of benchmarks each for different evaluation purpose. Table I shows the selected benchmarks.

A customized CloudStack template (image used to establish VM) and Eucalyptus VM image have been created in which all benchmarks are installed and configured to save time and ease of work.

Each benchmark test is repeated five times consequently and the average of results is considered. Different numbers and types of VM are regarded in the performance evaluation. In each cloud the same VM type is used and the same OS is run which it Centos 6.3. Moreover, each cloud is built with similar hardware and uses the same hypervisor (KVM) to achieve a fair comparison between Eucalyptus and CloudStack and eliminate virtualization and hardware differences that may affect evaluation. Table II shows types of VMs that are provided by Eucalyptus and CloudStack.

TABLE I. Benchmarks Deployed For VM Testing

No.	Benchmarks	Testing Resource
1	LINPACK	Processor
2	Bonnie++	Disk I/O
3	STREAM	Memory Bandwidth
4	Iperf	Network
5	Lookbusy	Processor
6	UnixBench	Overall System

TABLE II. Offering VM Types

Type	RAM	CPU core	Disk(GB)
Small	512M	1	10
Medium	1G	1	20
Large	2G	2	40
XLarge	4G	4	60

IV. Performance Evaluation of VMs

VMs of both clouds have been evaluated using selected benchmarks considering different relating metrics. The VM performance has not been compared just between Eucalyptus and CloudStack, but it also has been compared in regard to bare-metal or traditional IT infrastructure.

A. Comparison with Traditional IT infrastructure

The first question that comes in the mind of the cloud users or organization that plan to adopt the cloud computing solution is that "does the cloud virtual machine performance is the same as traditional physical machine? ". To answer this question, the performance of both machine with same hardware and software is tested using the same benchmark that mimic the real workload. UnixBench benchmark has been run with the traditional hardware stack on the host server of both cloud, then is run on both Eucalyptus and CloudStack Cloud on a single VM utilizing the whole host server resources.

As shown in figures 1 and 2, the performance of Eucalyptus cloud VM is nearly the same as physical one while there is a 7% gain in performance of the CloudStack VM. This result suggests that the cloud computing management system exploits or utilizes the computing resources on the same hardware stack better than the bare-metal or traditional IT system.

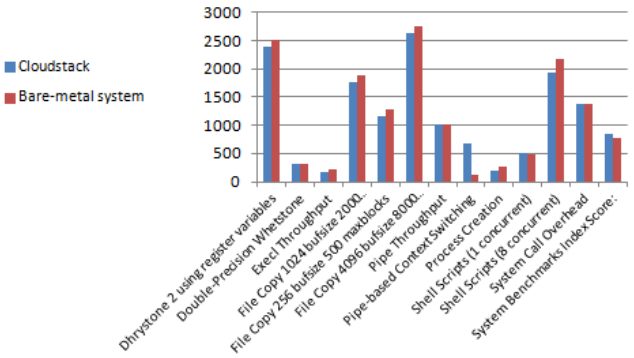


Fig. 1. Performance of CloudStack vs. Bare-metal system

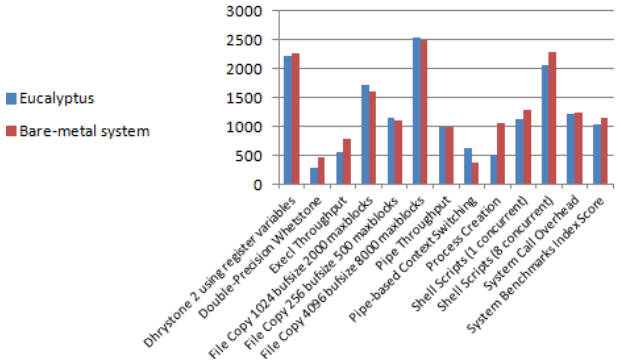


Fig. 2. Eucalyptus VM vs. Bare-metal system

B. Processing Performance

The Eucalyptus and CloudStack VM computing power has been assessed to test its ability in running a HPC (High performance Computing) applications. LINPACK is a benchmark that measures a computer's floating-point rate of execution by solving a dense *n* by *n* system of linear equations in double precision. Gflop/s is the rate of execution; it refers to billions of floating point operations per second.

In this test three scenarios have been applied. First two types of VMs (small and large) are evaluated as VM computing power varies according to its type. The number of linear equations is set to n = 7000 in small VM and n= 10000 in large one. In the Second scenario, performance of VM is evaluated when there are different numbers of VMs are running the LINPACK simultaneously in order to test CPU isolation of VMs and check if there is any interference among them because of resource sharing. In this scenario a medium type VM with n=7000 has been used.

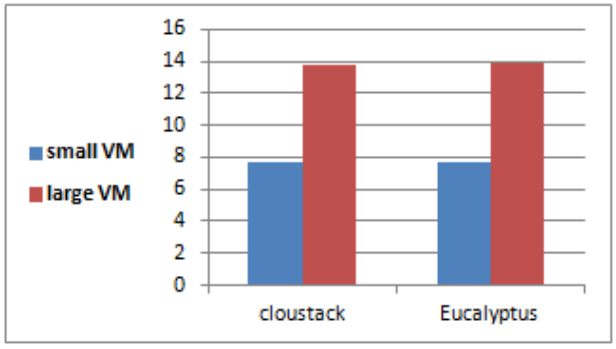


Fig. 3. CPU Performance of VMs

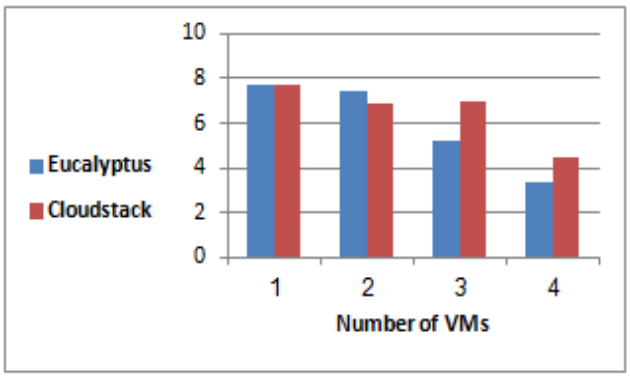


Fig. 4. CPU Isolation

Figure 3 shows the performance of VMs types. Eucalyptus and CloudStack VMs get a similar score. The floating point execution rate is considered very good with 7.7 Gflop/s and 13.8 Gflop/s for small and large VMs respectively in Eucalyptus, and 13.7Gflop/s and 7.6 Gflop/s in CloudStack as compared to values with performance of physical machines with similar hardware specifications as in [15]. Figure 4 represents the performance when the benchmark is running on a multiple VMs. The figure reveals that CloudStack provides a slightly better VMs CPU isolation than Eucalyptus. In this scenario the VMs have been assigned the entire physical cores of host server.

The third scenario tests the performance when CPUs overcommitting is implemented. CPU overcommit is the process of allocating more virtualized CPUs (vCPU) to VM

than actual physical CPUs of system. This requires underlying hardware and hypervisor support, and this is one of reason why KVM has been chosen in the clouds deployment. It allows resource utilization and running fewer CPU cores which saves power and money. After testing and customizing the overcommit ratio in our clouds, it has been set to two times the number of physical CPUs in the system.

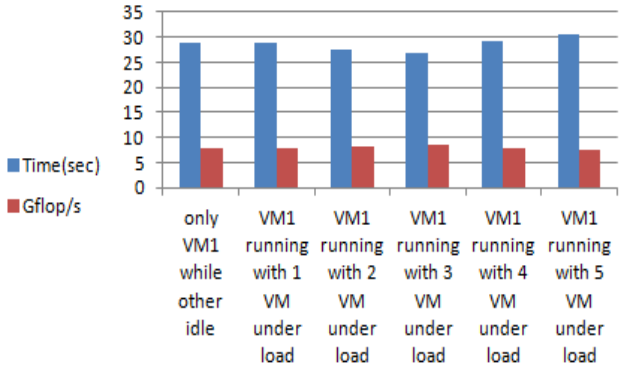


Fig. 5. CloudStack CPU Overcommitted Performance

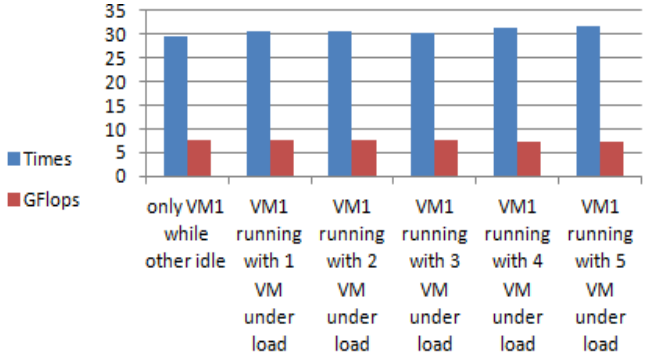


Fig. 6. Eucalyptus CPU Overcommitted Performance

Figures 5 and 6 represent performance of VMs with vCPU. LINPACK is run on medium VM with N= 7000. Then the test was repeated when there are other VMs running with 90% CPU utilization to test the effects of processor interference due to overcommiting. Lookbusy has been used to generate a high CPU utilization in VMs; it is an application for generating synthetic load on a system by generating fixed, predictable loads on CPUs, keeping chosen amounts of memory active, and generating disk traffic.

Figures reveal that assigning VM a vCPUs is appropriate and works as expected, as there is no effects from other VMs on the tested VM that run Linpack. Floating-point rate and time of execution are nearly the same as number of VMs with high utilization increased in each case on the both cloud platforms. This scenario revealed that the cloud vCPU solution is better that using normal CPU core in performance and isolation; this is due to CPU job scheduling and fair sharing techniques implementations of CPU overcommit.

C. Disk I/O Performance

As previously mentioned, Eucalyptus uses host local disk for VM, while CloudStack uses primary storage that access via NFS for VMs disks.

To evaluate and compare the performance of both clouds VM disk I/O, the Bonnie++ benchmark is adopted in this test.

Bonnie++ is a well-known Disk I/O performance benchmark suite that uses a series of tests including data read and write speeds, maximum number of seeks per second, maximum number of file creations, and deletion or gathering of file information per second.

Two scenarios are implemented on both clouds. First, Disk I/O of two types of VMs, small and large are evaluated. Bonnie++ documentation recommend that file size should be double RAM size, therefore files with 1GB and 4GB sizes for small and large VM respectively were considered. Second, performance of VM when there is another VM performing intensive disk I/O operation is inspected.

This is done to test isolation between VMs and check if there is any interference.

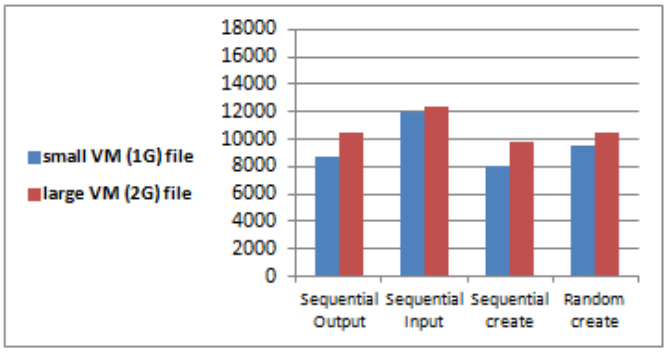


Fig. 7. Disk Access Speed in CloudStack

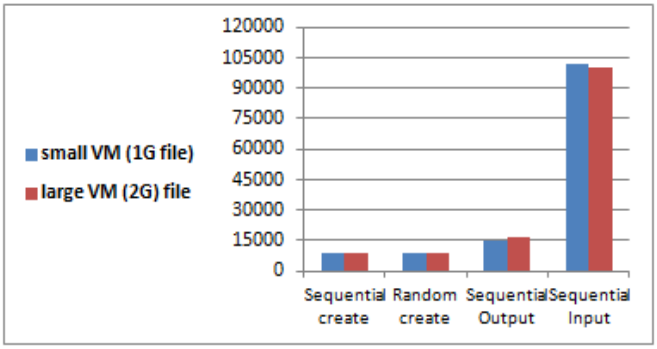


Fig. 8. Disk Access Speed in Eucalyptus

Figures 7 and 8 show the performance of VM types of both clouds. Sequential Output shows the speed in KB/s in which the data has been written. Sequential Input is the speed the data has been read, Sequential and Random create refer to the number of files created per sec.

Eucalyptus has a better overall performance than CloudStack; this is due to using of local disk configuration for VMs in Eucalyptus so VM access the host disk locally, while in CloudStack it accesses shared disk of primary storage over the network via NFS which declines disk I/O speed and performance.

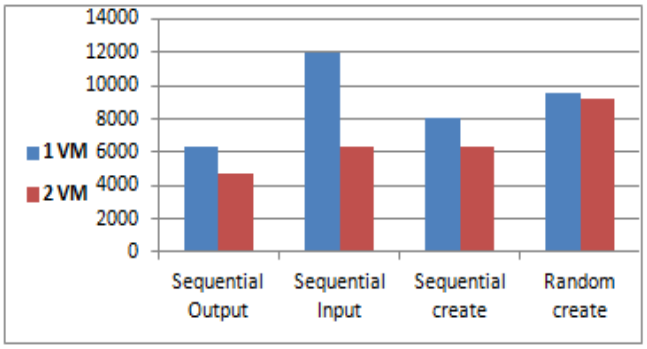


Fig. 9. Disk Isolation in CloudStack

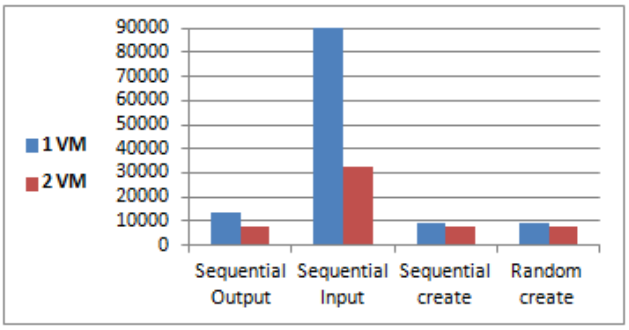


Fig. 10. Disk Isolation in Eucalyptus

Figures 9 and 10 illustrate the performance when two VMs are carrying intensive read write file operations concurrently. In this scenario medium type VM with 1GB file size is dealt with. It reveals that disk I/O performance of VM disk is impacted by the other VM as its performance drops in the both clouds. In CloudStack NFS configuring, this is expected due to primary storage disk sharing and available network bandwidth of VM.

In Eucalyptus, the NC's disk capacity and bandwidth is typically shared between VMs. The capacity is shared in a straightforward way: each virtual machine has a virtual disk image of a determined size that is allocated at the VM starting time. It does not change until the termination of the VM execution. On the other hand, the bandwidth of the disk is shared between all the resident VMs and there is currently no method of dividing this bandwidth or imposing limitations on its consumption by VMs. Therefore, the disk I/O performance of one user would be interfered by another user's VM with intensive disk I/O behavior.

Despite that the interference problem is existed in both cloud platforms; Eucalyptus has a better disk performance than CloudStack. This is due to local storage configuration where the VM disk is accessed locally (within host server) not over the network via NFS.

D. Memory Performance

The memory performance stress test is based upon a bandwidth test, as this is what distinguishes between types of memories. To measure the memory bandwidth the STREAM benchmark is used. It is a synthetic benchmark tool that measures memory bandwidth (in MB/s). It is specifically designed to work with datasets much larger than the available cache on any given system, so that the results are more indicative of the performance of very large, vector style applications.

Figure 11 indicates the results of memory performance of small and large VMs in MB/s of both clouds. The array size applied in the benchmarking is 10,000,000 elements for small VM and 70,000,000 elements for large VM.

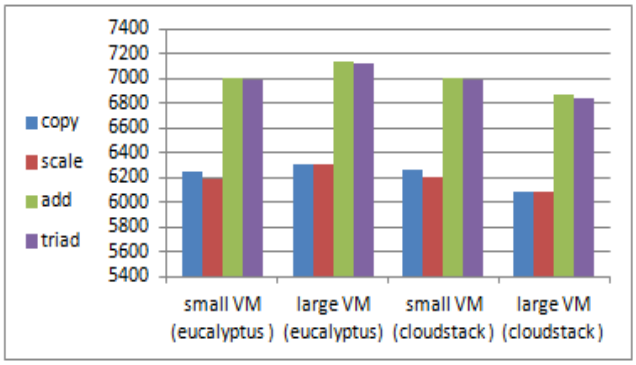


Fig. 11. Memory Performance

Figures 12 and 13 show the memory isolation between VMs, residing on the same host server. In this scenario STREAM benchmark is run on multiple VMs simultaneously. The tests demonstrates that with only one VM provisioned, there are plenty of rooms for further utilization of memory but as the number of VM increase the bandwidth available to each drops. Hence it requires a scheduler to avoid such effects.

Despite that the memory isolation problem is existed in the both cloud platforms; CloudStack shows better memory performance than Eucalyptus.

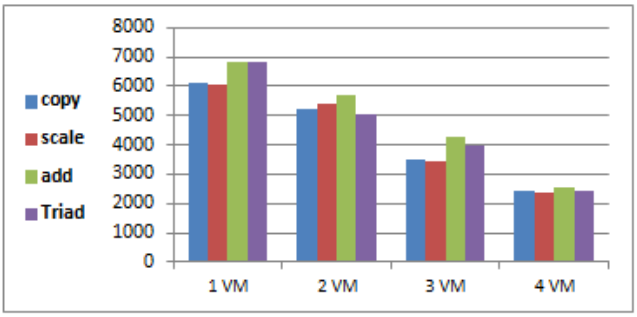


Fig. 12. Memory Isolation in CloudStack

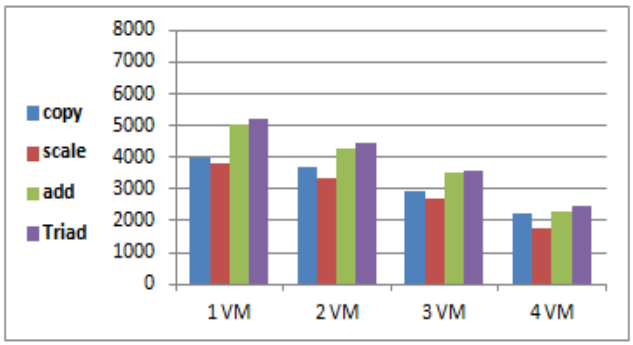


Fig. 13. Memory Isolation in Eucalyptus

E. Network Performance

The network performance tests are performed using Iperf. It is a network testing tool that allows the user to set various parameters that can be used for testing a network. It implements a client and server scheme to measure network performance between two ends, by creating a TCP and UDP data streams and measuring the throughput of network that is carrying them.

Three scenarios have been employed. First, bandwidth of VMs inside the cloud is measured by running two VMs, one as client and other as a server and TCP bandwidth between them is measured. Thereafter, the test is repeated when there are others VMs using the network. Second, packet loss is calculated at different bandwidths using UDP mode with a different number of VMs. Third, jitter is determined using UDP mode when there are more than one VM sending or receiving data over the network.

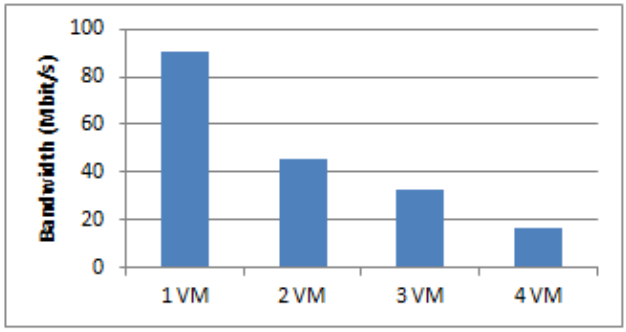


Fig. 14. Network Bandwidths Inside Eucalyptus Cloud.

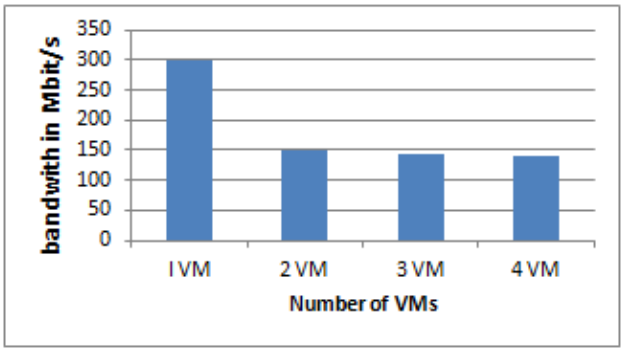


Fig. 15. Network Bandwidths Inside CloudStack Cloud.

Having seen the disk I/O interference problems, it is expected to find similar issues in the process of sharing another resource that is the network adapter. Figure 14 shows that performance of VM degrades as number of VMs increase. It proves that Eucalyptus has no built-in system of bandwidth fair-sharing between VMs. Every time concurrent TCP connections in the network are started from the VMs, each of them gets a different share of the link bandwidth and has the ability to starve the other depending on which connection begins first.

Figure 15 shows network performance of CloudStack. It reveals that when one VM is communicating, it utilizes all available network bandwidth but when there are others VMs using the network, the bandwidth is fairly divided among them.

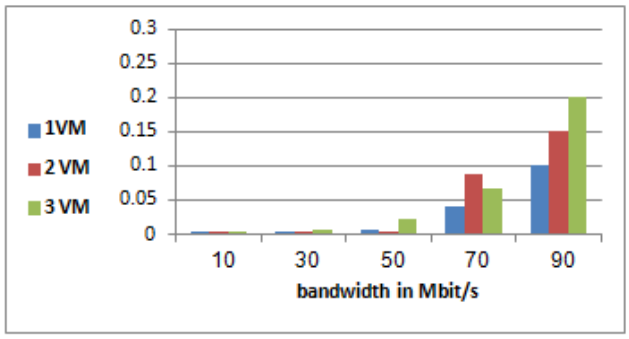


Fig. 16. Packets Loss in CloudStack

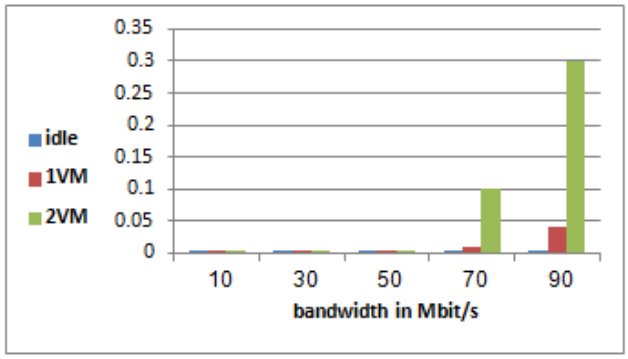


Fig. 17. Packets Loss in Eucalyptus

As depicted in figures 16 and 17, the packet loss is persisting around zero when each VM is communicating at a small bandwidth but as the bandwidth increases the packet loss increases considerably. However it does not arrive to a critical loss value in the both clouds.

Figure 18 expresses jitter when the VM is using 100Mbit/s bandwidth. As the number of VMs concurrently using network increases, the jitter is slightly increased in Eucalyptus while the jitter value is nearly the same in CloudStack. This is due to that bandwidth is fairly divided among VMs.

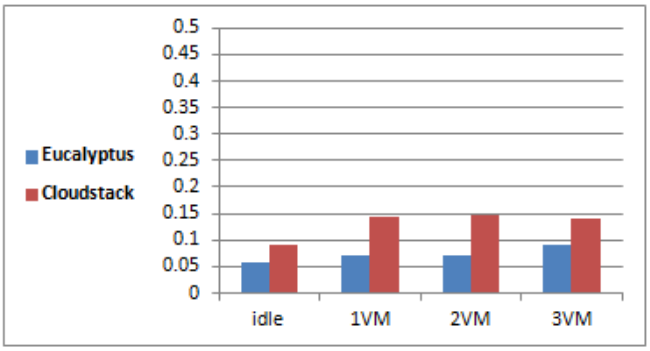


Fig. 18. Jitters in Eucalyptus and CloudStack

CloudStack has better network connection performance than Eucalyptus, due to better internal design and using of vRouter system virtual machine in cloudstak. Also the network internal traffic does not have to go through the master node (the CLC in Eucalyptus that act as internal router).

Therefore, the network connection between internal VMs will be solely determined by the physical network card which is around 1Gbps. CloudStack also provides a good bandwidth sharing among VMs. The network performances for all cloud solutions are restricted by the physical network environment.

V. VM Provisioning and Release Time

One of many advantages of the cloud is the elasticity that is the ability to dynamically acquire or release computing resources in response to demand. However, this elasticity is only meaningful to the cloud users when the acquired VMs can be provisioned in time and be ready to use within the user expectation. The long unexpected VM startup time could result in resource under-provisioning, which will inevitably hurt the application performance, hence it is required to evaluate the VM startup and release time to help cloud users to plan ahead and make in-time resource provisioning and releasing decisions [12]. A systematic study of VM provisioning and releasing time has been done for the Eucalyptus and CloudStack considering different related factors.

A. Number of VMs

The average provisioning time of VMs in CloudStack cloud is 16 seconds while in Eucalyptus, it is 127 seconds. This difference is due to CloudStack NFS storage configuration, in CloudStack the VM uses primary storage as its disk access via NFS while other resources (CPU, memory …) are provided by host server so there is no need to copy VM image file from image repository in primary storage to host machine disk. However in Eucalyptus VMs use host local disk. Therefore when a new VM is provisioning, the image file (size in Mbytes) is copied from Walrus storage to host machine (node controller) which is time consuming.

Figure 19 reveals that when the number of VMs requested increases, the launch time increases accordingly in both clouds. This is due to that both cloud platforms handle each VM requested as if it is launched individually (one requested after other). The provisioning time of 2 VMs request in CloudStack is 31 seconds which equals the sum of two VMs startup time requested alone, and the same applies for VMs request. In Eucalyptus, the launch time of multiple VMs shows a time difference; for example the time for 3 VM provisioning is 134 second, which it not a 3 times of provisioning one VM. This is due to Eucalyptus is not resending the image file for multiple VM. So when a new VM is creating, the Eucalyptus checks if the image file exists in the images cache on host server. Therefore there is no need to copy it again from the walrus. The little difference in multiple VMs provisioning is the time consumed in each VM resources allocation.

B. Type of VMs

The VM provisioning and release time in both cloud platforms is not influenced by its type as illustrated in figures 20 and 21. VMs with different types have nearly the same startup and release time around 16 and 28 seconds respectively in CloudStack and 128 and 10 seconds in Eucalyptus. This reveals the satisfactory and quick VM resource allocation schedulers of both Eucalyptus and CloudStack.

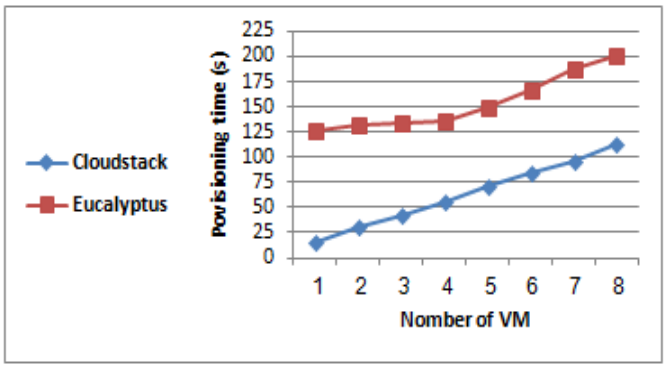


Fig. 19. VMs Launch Time vs. Number of VMs

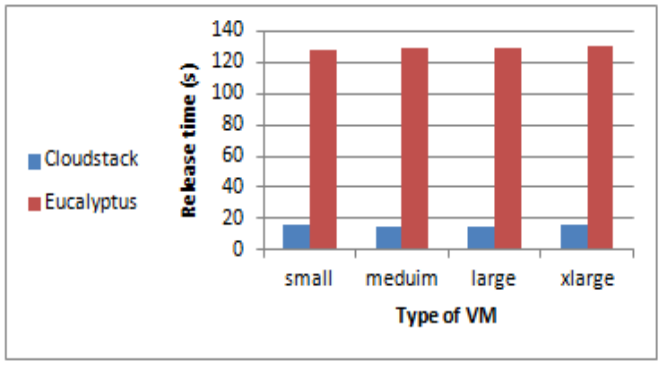


Fig. 20. VMs Startup Time vs. Type of VMs

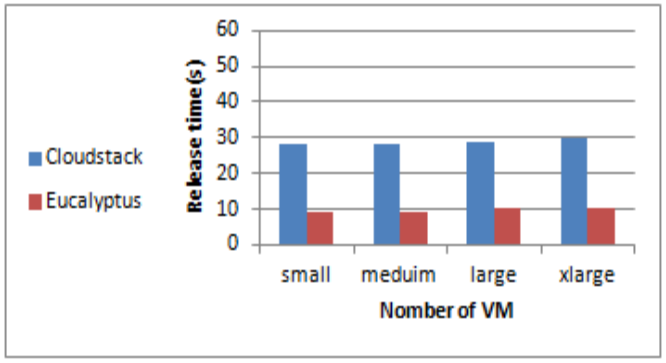


Fig. 21. VMs Release Time vs. Type of VMs

C. Image Size

The VM provisioning time is not influenced by size of image or template used to initiate it in CloudStack as depicted in figure 22. VMs with different image sizes have nearly the same startup time around 16 seconds. This is due to using primary storage as shared disk for VMs in CloudStak access via NFS. Therefore there is no need to copy templates (of different sizes) from primary storage to hosts which results in reducing the time of VM startup regardless of template size.

In Eucalyptus, the size of VM image (which depend mainly on OS) can largely impacts the provisioning time as it is shown in figure 23. This is due to local disk configuration of Eucalyptus which requires VM image copying from image repository in walrus to disk of sever that hosts VM. The larger the image file , the longer the VM provisioning time will be.

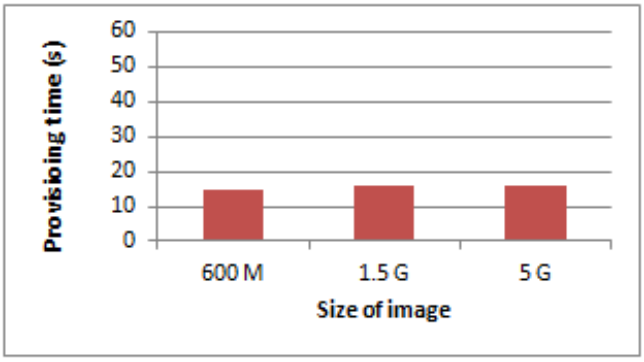


Fig. 22. VMs Startup Time with Different Image Size in CloudStack

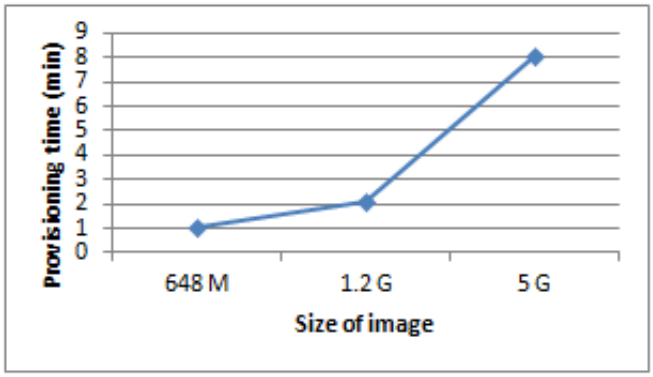


Fig. 23. VMs Startup Time with Different Image Size in Eucalyptus

D. Adding Additional Disk Space

CloudStack allows users to attach additional volume to VM disk at time of creation. The VM provisioning and release time is not affected by adding additional disk volumes as being requested by the user. The VM startup and release time is nearly the same when adding different disk size to the VM which is rated 16 and 29 seconds respectively as shown in figure 24. This is probably due to that CloudStack uses the primary storage to provide disks to VMs with a quick resource allocation scheduler. Eucalyptus allows attaching disk volume to running VM only.

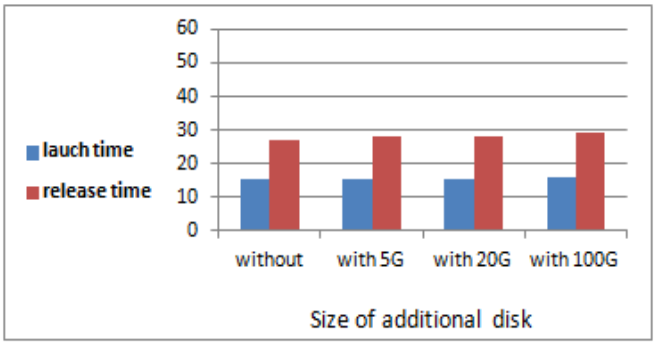


Fig. 24. VMs Startup and Release Time vs. Additional Disk

VI. Performance Evaluation of VM Live Migration

CloudStack supports live migration of VMs between host servers while Eucalyptus supports only cold migration of VMs due to local disk configuration and lack of central sharing of VMs. Cold migration requires stopping the running VM and then moving it with its data disk to another host machine where it starts and runs again. So the VM will have a downtime which may affect user works [16]. Cold migration has no advantage in disaster recovery since VM disk is located at the host machine. So if the host fails, the VM and its data disk will be lost. This is contrast to the live migration in CloudStack where the VM data disk is at high available primary storage. Therefore in case of host failure, the VM can migrate to another host and resume work and access its disk via NFS.

Time duration of VM live migration in CloudStack has been expressed considering different factors as pursued.

A. Image Size

Duration of VM migration is influenced by image or template size used to initiate it as shown in figure 25. There is no difference when using 1G and 5G image size for VMs using shared disk. There is no need to move data disk from source host to destination host. That is the size should not affect migration time. However, when 600M image has been deployed, it takes a shorter time than 1G and 5G. This is due to that these are GUI OS images while 600M is non. This means that it is lighter and its applications consume less CPU and memory; the context switch compromising CPU status and memory pages copied from source to the destination host, is of small size thus it migrates faster.

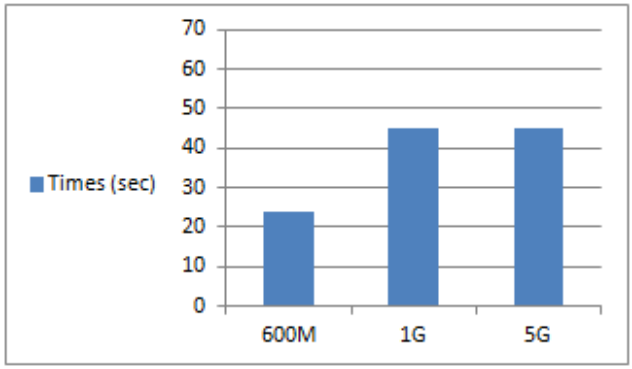


Fig. 25. Live Migration with Different Image Size

B. Types of VMs

We have measured migration time of different types of VMs running normal application. The type of VM can largely influence the duration of migration as shown in figure 26. This is due to increasing memory size assigned to VM in each type, so the duration of live migration increases linearly with it.

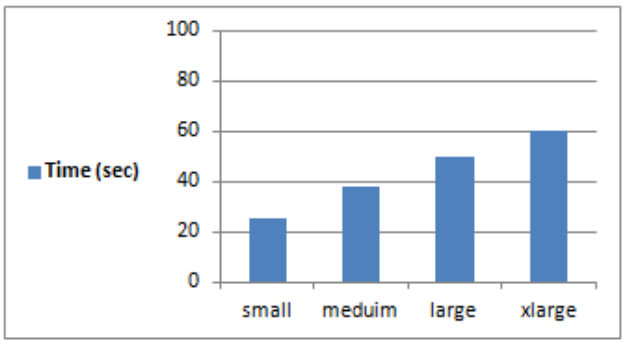


Fig. 26. Live Migration with Different VM types

C. Number of VMs

The average time of live migration of VM in the CloudStack cloud is 40 seconds. When the number of migrating VM increase, this time increases accordingly, as it is shown in figure 27.

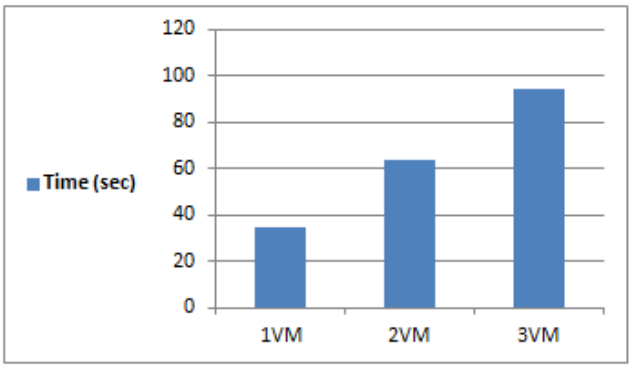


Fig. 27. Live Migration with Different Number of VMs

D. CPU load

We have measured the migration time of VM when the CPU is running an intensive application to assess its effect on migration as a relating factor. We have tested two types of VMs, medium and large and have used Lookbusy tool to generate a 90% CPU utilization. We have found that the CPU load can have an impact on the duration of migration as shown in figure 28.

We can conclude that live migration depends on CPU utilization and applications running on the VM.

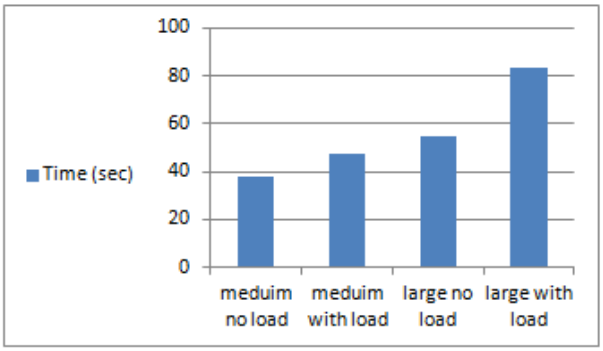


Fig. 28. Live Migration with CPU Load

VII. Web Application Over Cloud Computing

One of most popular usage of VM in cloud is web application hosting. Cloud hosting has many advantages over traditional web hosting like cost reduction, scalability, flexibility, backup, security and isolation, and unlimited storage capacity. The main issues of running web application over cloud are performance and stability.

To evaluate Eucalyptus and CloudStack Clouds in hosting a web application, we have implemented a web application on VM in both clouds and measured the response time of web application to test the stability of running the application on VM. Etherpad is an open source online office suite similar to Google Docs. It is a web-based collaborative real-time editor. Etherpad has been implemented on VMs of both Clouds with MySQL as the database and Nginx as the web server.

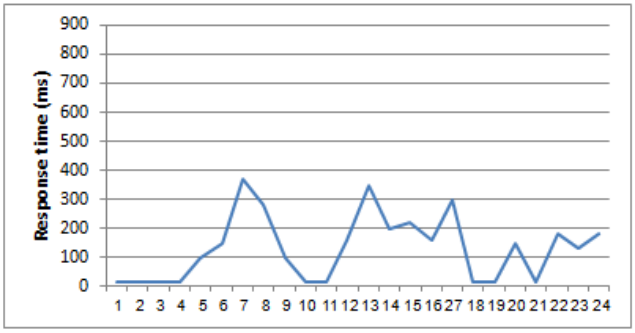


Fig. 29. Response Time in Eucalyptus Cloud

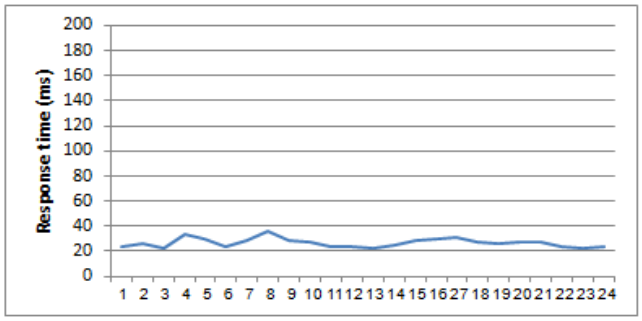


Fig. 30. Response Time in CloudStack Cloud

Figures 29 and 30 showed the response time of Etherpad which has been run on both clouds for 24 hours. The data is collected every one hour. The goal is to test VM stability on running the web application with changing the background load. Different numbers of VMs are let to run and invoke different disk and CPU intensive operations to test whether the cloud resource sharing could affect the running web application.

The figures 29 and 30 reveal that CloudStack VM is more stable than Eucalyptus VM in running the web application as the response time values over the 24 hours are nearly constant between 20 and 38 ms, while in Eucalyptus, this dramatically varies between 14 and 380 ms. Hence CloudStack is more suitable in hosting web application than Eucalyptus cloud.

VIII. CONCLUSION

In this paper, we analyzed and compared the performance of Eucalyptus and CloudStack cloud with different storage configuration thoroughly to assess its suitability to be adopted as an open source private cloud solution for different business and scientific purposes. We have considered the performance of VM as the key point of evaluation. It has been found that storage configuration of the cloud largely affects VMs performance. CloudStack NFS configuration is 69% faster in VMs provisioning than Eucalyptus local disk configuration, while VM disk I/O performance in Eucalyptus local disk configuration outperformed the VM disk performance in CloudStack NFS configuration.

VMs performance of both clouds was evaluated in regard to CPU utilization, disk I/O speed, Memory bandwidth, Network performance, and VM management operations such as VM provisioning time and live migration. The result showed that there is always a performance decrease due to co-located VMs running resource-intensive tasks. The drop in performance is slight for CPU and memory intensive workload and very significant for disk and network I/O intensive workloads. The major lessons learned related to the performance evaluation of VM management operation are: (1) the duration for the live migration changes with the CPU load; (2) the duration for the live migration increases linearly as the memory assigned to the VM increases; (3) the startup and release time have not been impaired by the VM type; (4) the startup and release time have not been impaired by image size or by adding additional disk volumes in CloudStack, while the startup time is largely affected by image size in Eucalyptus.

Also, Eucalyptus and CloudStack clouds ability in hosting web applications was tested by measuring the response time of web application that was hosted on their VMs. It has been found that CloudStack is more suitable in hosting web applications and as private cloud solution in general due to its stability and fair VMs performance. On the other hand, Eucalyptus is easier in deployment and more modular, it can be used in testing a specific application on the cloud so it's a good choice for developers and researchers in this field.

IX. FUTURE WORK

As a future work we intend to analyze security aspects of Eucalyptus and CloudStack by evaluating the compliance of them with security standards related to cloud security. Also, the methodology and the benchmarks used for performance evaluation in this paper can be used for different cloud management platforms whatever open source or commercial platforms (OpenNebula OpenStack, HP cloud, VMware, etc.) and compare their performance results with this paper to extend the evaluation of the cloud management platforms.

REFERENCES

[1] E. T. de Sousa, P. R. Maciel, E. M. Medeiros, D. S. de Souza, F. A. Lins,and E. A. Tavares, "Evaluating Eucalyptus Virtual Machine Instance Types: A Study Considering Distinct Workload Demand ", Third International Conference on Cloud Computing, GRIDs, and Virtualization pp. 130- 135, July 2012.

[2] H. Tianfield, "Cloud Computing Architectures", Proceeding of the 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1394-1399, Oct 2011.

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing" National Institute of Standards and Technology Special Publication 800-145, September 2011.

[4] Z. Pantic and M. A. Babar, "Guidelines for Building a Private Cloud Infrastructure ", Technical Report TR-2012-153, IT University of Copenhagen, Denmark, 2012.

[5] D. N. Chorafas, "Cloud computing strategies", CRC Press, 2011.

[6] A. S. Pilla, and L.S. Swasthimthi, "A Study on Open Source Cloud Computing Platforms ", EXCEL International Journal of Multidisciplinary Management Studies, Vol. 2, Issue 7, July 2012.

[7] I. Voras, et al., "Evaluating Open-Source Cloud Computing Solutions", Proceedings of the 34th International Convention MIPRO, pp. 209-214. May 2011.

[8] [8] P. Sempolinski and D. Thain, "A Comparison and Critique of Eucalyptus, OpenNebula and Nimbus", Proceedings of the IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), pp. 417-426, November 2010.

[9] [9] C. El Amrani, et al, "A Compartive Study of Cloud Computing Middleware", Proceeding of the12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 690-693, 2012.

[10] [10] G. V. Laszewski, , J. Diaz, F. Wang, G. C. Fox, "Comparison of Multiple Cloud Frameworks", proceeding of the 5th IEEE International Conference on Cloud Computing (CLOUD 2012), pp.734-741, June 2012.

[11] [11] O. Sefraoui, M. Aissaoui, and M. Eleuldg "Comparison of multiple IaaS Cloud platform Solutions", Recent Researches in Information Science and Applications, Proceedings of the 7th WSEAS International Conference on Computer Engineering and Applications (CEA '13), pp. 212-217, January 2013.

[12] M. Mao and M. Humphrey, "A Performance Study on the VM Startup Time in the Cloud ", Proceeding of the 5th IEEE International Conference on Cloud Computing (CLOUD 2012), pp.423-430, June 2012.

[13] D. Steinmetz, B. Perrault, R. Nordeen, J. Wilson, X. Wang,"Cloud Computing Performance Benchmarking and Virtual Machine Launch Time", Proceedings of the 13th annual conference on Information technology education, ACM New York, pp. 89-90, 2012.

[14] F. Gomez-Folgar, A. Garcia-Loureiro and T. F. Pena, R. Valin, " Performance of the CloudStack KVM Pod Primary Storage under NFS Version 3", Proceeding of 10th IEEE International Symposium on Parallel and Distributed Processing with Applications, pp. 845-846, 2012.

[15] J. J. Dongarra, "Performance of Various Computers using Standard Linear Equations Software", scientific report, University of Manchester, February 5, 2013.

[16] A. Agarwal and R. Shangruff, "Live Migration of Virtual Machines in Cloud", International Journal of Scientific and Research Publications, Volume 2, Issue 6, pp.1-5, June 2012.

An Enhanced Fuzzy Multi Criteria Decision Making Model with a proposed Polygon Fuzzy Number

Samah Bekheet
Department of Computer and Information Science, ISSR, Cairo University, Egypt

Ammar Mohammed
Department of Computer and Information Science, ISSR, Cairo University, Egypt

Hesham A. Hefny
Department of Computer and Information Science, ISSR, Cairo University, Egypt

***Abstract*—Decisions in real world applications are often made under the presence of conflicting, uncertain, incomplete and imprecise information. Fuzzy multi Criteria Decision making (FMCDM) approach provides a powerful approach for drawing rational decisions under uncertainty given in the form of linguistic values.**

Linguistic values are usually represented as fuzzy numbers. Most of researchers adopt either triangle or trapezoidal fuzzy numbers. Since triangle, intervals, and even singleton are special cases of Trapezoidal fuzzy numbers, so, for most researchers' Trapezoidal fuzzy numbers are considered, generalized fuzzy numbers (GFN). In this paper, we introduce polygon fuzzy number (PFN) as the actual form of GFN. The proposed form of PFN provides higher flexibility to decision makers to express their own linguistic rather than other form of fuzzy numbers. The given illustrative example ensures such ability for better handling of the FMCDM problems.

Keywords—Fuzzy multi criteria decision making; linguistic values; polygon fuzzy number; level set.

I. Introduction

Decision making is the procedure to find the best alternatives among a set of feasible alternatives and also ranking them as their priorities. Under a fuzzy environment, fuzzy multi criteria decision making (FMCDM) needs linguistic values (e.g. excellent, very good, good, bad) to enable the decision makers to express their own opinions. Such linguistic values need fuzzy tools to evaluate their calculations [1].

Examples of FMCDM tools are T-Norm Based, Gaussian fuzzy numbers, Interval fuzzy numbers, Interval type two fuzzy number, Triangle fuzzy numbers and Trapezoidal fuzzy numbers [2].

Interval, Triangle and Trapezoidal fuzzy numbers are more popular due to their conveniences of the arithmetic operations such as: addition, subtraction, multiplication, division, reciprocal, geometric mean, etc. Such operations enable the decision makers to determine the rank of criteria (alternatives) powerfully [3].

Several researchers consider Trapezoidal fuzzy numbers as Generalized fuzzy numbers [4,5,6]. This mainly due to the fact that other popular forms of specific fuzzy numbers including: triangles, intervals, or even singleton can be obtained as special cases of Trapezoidal fuzzy numbers. The main contribution of this paper is to introduce the polygon fuzzy number as a general fuzzy number rather than the trapezoidal one. Using fuzzy polygon number, several fuzzy numbers can be obtained (e.g. triangle, trapezoidal, pentagon, hexagon, heptagon, octagon and etc).

Such generalization introduces more flexibility to decision makers their own linguistic values. It also to express gives the possibility to standardize different fuzzy numbers into a specific polygon fuzzy number to solve various decision making problems based on different views of decision makers [7, 8, 9].

The rest of the paper is organized as follows: Section II defines the problem. Section III introduces the proposed model and the required definitions of the proposed model. Section IV introduces a numerical example. Finally section V presents the conclusion.

II. Problem Definition

A Trapezoidal fuzzy number with its four vertices, is considered a generalized number for other forms including: triangles, intervals and also singletons (i.e. crisp numbers) [4,5,6].

However, it is intuitively clear that allowing more vertices to the fuzzy number adds more flexibility to the decision maker to represent his own opinion to deal with the considered FMCDM problem.

Therefore, the ability to introduce generalized piece-wise membership function with n-vertices as a fuzzy number with its own arithmetic operations represents the typical unification of all other forms of fuzzy numbers. Adopting such new forms of generalized fuzzy numbers should considerably enhance modeling and solving FMCDM problems.

III. Polygon Fuzzy Number

A. Basic Definitions

Polygon fuzzy sets are firstly addressed in [10] in the context of fuzzy interpolative reasoning. A polygon fuzzy set *A* has n characteristic points (a_0, a_1,....,a_{n-1}) as shown in fig.

1. The core of the fuzzy set, at which the membership equals one, is represented by the interval $[a_{\lfloor (n-1/2) \rfloor}, a_{\lceil (n-1/2) \rceil}]$.

There are $\lfloor (n-1)/2 \rfloor + 1$ membership levels including bottom and top levels. Thus the cardinality of the level set of a polygon fuzzy set is denoted by V as given in (1)

$$V = \lfloor (n-1)/2 \rfloor + 1 \quad (1)$$

It is clear that V represents the number of α-cuts of the polygon fuzzy sets, namely:

$\alpha_0 = 0, \ldots, \alpha_{\lfloor (n-1)/2 \rfloor + 1} = 1$.

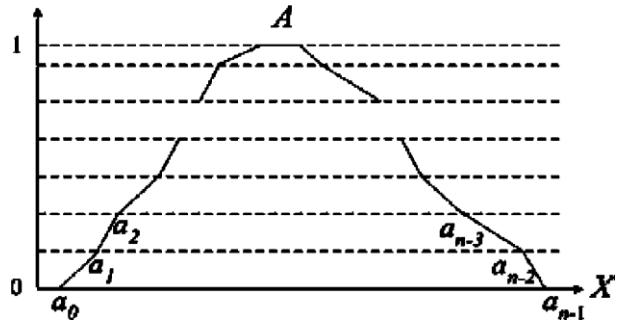


Fig. 1. Polygon fuzzy set

A polygon fuzzy number (PFN) is defined as a convex and normal polygon fuzzy set.

B. Ranking of Polygon fuzzy number

The centre of area of a fuzzy number is considered the most popular ranking method [11] [12].

Computing the centroid of a general polygon fuzzy number can be obtained by the following theorem.

***Theorem 1*:** Let A be a polygon fuzzy number that is characterized by the n-points $((x_0, x_1, \ldots, x_{n-1})$. Let $y_i = \mu_A(x_i)$, be the membership level for each point x_i . Then the centroid $\bar{X}$ of A is Computed using the following formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} \overline{X}_i A_i}{\sum_{i=1}^{n} A_i} \quad (2)$$

$$\overline{X}_i = \frac{1}{3}[(x_i + x_{i-1}) + (\frac{x_i y_i + x_{i-1} y_{i-1}}{y_i + y_{i-1}})] \quad (3)$$

$$A_i = \frac{(x_i - x_{i-1})(y_i + y_{i-1})}{2} \quad (4)$$

Proof:

The polygon fuzzy number A with its n vertices see fig 2, can be divided into (n-1) sub-polygons.

Each sub-polygon can generally be represented as a trapezoidal with 4-vertices. Thus the i^{th} sub-polygon as shown in fig.3, has an area equals A_i and its centroid $\overline{x_i}$ is computed as follows:

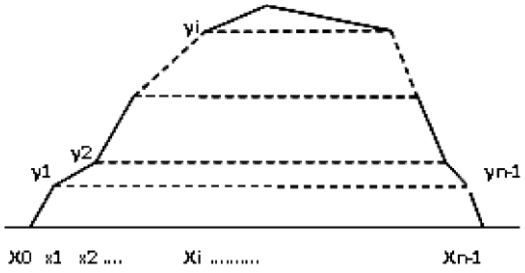


Fig. 2. Polygon characterized

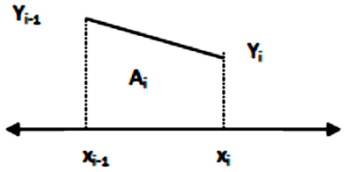


Fig. 3. Sub- polygon area

$$\overline{x_i} = \frac{\int_{x_{i-1}}^{x_i} x \, f_i(x) dx}{\int_{x_{i-1}}^{x_i} f_i(x) dx}$$

$$= \frac{I_1}{I_2} = \frac{f_i(x) - y_{i-1}}{x - x_{i-1}} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$$

$$f_i(x) = \frac{y_i - y_{i-1}}{x_i - x_{i-1}} (x - x_{i-1}) + y_{i-1}$$

$$= \frac{y_i - y_{i-1}}{x_i - x_{i-1}} x - \frac{y_i - y_{i-1}}{x_i - x_{i-1}} x_{i-1} + y_{i-1}$$

$$= \frac{\Delta y_i}{\Delta x_i} x + y_{i-1} - \frac{\Delta y_i}{\Delta x_i} x_{i-1}$$

Therefore,

$$f_i(x) = \alpha_i x + \beta_i$$

where

$$\alpha_i = \frac{\Delta y_i}{\Delta x_i}, \beta_i = y_{i-1} - \frac{\Delta y_i}{\Delta x_i} x_{i-1}$$

Then, performing the integrations for both I_1 and I_2, we have:

$$\overline{x_i} = \frac{1}{3}[(x_i + x_{i-1}) + (\frac{x_i y_i + x_{i-1} y_{i-1}}{y_i + y_{i-1}})]$$

Also, it is clear that the area of the i^{th} sub-polygon shown in fig. 3 is:

$$A_i = \frac{(x_i - x_{i-1})(y_i + y_{i-1})}{2}$$

C. Arithmetic Operations

The arithmetic operations of two polygon fuzzy numbers should satisfy the following two rules:

1) Both fuzzy numbers should have the same number of vertices.

2) Both fuzzy numbers should have the same level set.

Thus, if we have to add two different polygon fuzzy numbers, e.g triangle $T(a_0, a_1, a_2)$ and hexagon $H(b_0, b_1, b_2, b_3, b_4, b_5)$. If the level set of the triangle is: $\{T(a_0)=T(a_2)=0, (a_1)=1\}$, and the level set of the hexagon is: $\{H(b_0) =H(b_5)=0, H(b_1)=H(b_4)=0.6, H(b_2)=H(b_3)=1\}$.

Then additional level set value at 0.6 should be added to the triangle fuzzy number and consequently more vertices appears for that triangle to be in the form: $T(a_0, a_1, a_2, a_3, a_4, a_5)$ so that its level set becomes: $\{T(a_0)=T(a_5)=0, T(a_1)=T(a_4)=0.6, T(a_2)=T(a_3)=1\}$.

Thus, keeping the above two rules in mind,

Assume that there two PFNs $A(a_0, a_1, a_2, ..., a_{n-1})$ and $B(b_0, b_1, b_2, ..., b_{n-1})$, then the arithmetic operations can be defined as follows:

1- PFNs Addition

$A \oplus B = (a_0+b_0, a_1+b_1, a_2+b_2, ..., a_{n-1}+b_{n-1})$

2- PFNs Subtraction

$A \ominus B = (a_0-b_{n-1}, a_1-b_{n-2}, a_2-b_{n-3}, ..., a_{n-1}-b_0)$

3- PFNs Multiplication

$A \otimes B = (a_0\times b_0, a_1\times b_1, a_2\times b_2, ..., a_{n-1}\times b_{n-1})$

4- PFNs Divisions

$A \Phi B = (a_0/b_{n-1}, a_1/b_{n-2}, a_2/b_{n-3}, ..., a_{n-1}/b_0)$

Where $b_0\neq0$, $b_1\neq0$, ... and $b_{n-1}\neq0$.

IV. Illustrative Exampe

Assume two alternatives are evaluated w.r.t three criteria C1, C2 and C3 as shown in table 1. The scale of evaluation extends from 1 to 9. Let the decision maker put his evaluation values in the form of polygon fuzzy numbers as shown in table1.

According to table 2, the whole level set is {0,0.5,0.6,0.7,0.8,1}. Therefore, all the above PFNs should be rewritten according to the whole level set. This of course will add more vertices as shown in table 3. It is clear, that the obtained unified PFNs are all having twelve vertices at the above six values of the level set.

TABLE I. Criteria-Alternatives Evaluation Matrix

Criteria / Alternatives	C1	C2	C3
A	A_{11}	A_{12}	A_{13}
B	B_{11}	B_{12}	B_{13}

TABLE II. The Evaluation Pfns

Fuzzy number	Type	Level set
$A_{11}(1,2,3,4)$	Trapezoid	{ A_{11} (1)= A_{11} (4)=0, $A_{11}(2)$
A_{12} ((3,4,5)	Triangle	{ A_{12} (3)= A_{12} (5)=0, $A_{12}(4)$
A_{13} (4,5,6,7,8,9)	Hexagon	{ A_{13} (4)= A_{13} (9)=0, $A_{13}(5)$ =0.6, A_{13} (6)= A_{13}
$B_{11}(1,2,3,4, 5)$	Pentagon	{ $B_{11}(1)$= $B_{11}(5)$=0, $B_{11}(2)$=0.6, $B_{11}(3)$ =$B_{11}(4)$ =1}
B_{12} ((2, 3, 4, 6)	Trapezoid	{$B_{12}(2)$=$B_{12}(6)$ =0, $B_{12}(3)$=$B_{12}(4)$=
B_{13} (4,5,7,8,9)	Pentagon	{$B_{13}(4)$=$B_{13}(9)$=0, $B_{13}(5)$ = $B_{13}(8)$=0.5, $B_{13}(7)$=1}

Now, get the normalized ranked PFN for each alternative as:

$R(A) = \sum A_{1k} / (\sum A_{1k} + \sum B_{1k})$, k=1,2,3 (5) $R(B) = \sum B_{1k} / (\sum A_{1k} + \sum B_{1k})$, k=1,2,3 (6)

Table 4 shows the obtained normalized ranked PFNs.

TABLE III. The New Forms of Evaluation Pfns

PFN level	A_{11}	A_{12}	A_{13}	B_{11}	B_{12}	B_{13}
0	1	3	4	1	2	4
0.5	1.62	3.5	4.83	1.83	2.5	5
0.6	1.75	3.6	5	2	2.6	5.4
0.7	1.88	3.7	5.25	2.25	2.7	5.8
0.8	2	3.8	5.5	2.5	2.8	6.2
1	3	4	6	3	3	7
1	3	4	7	4	4	7
0.8	3.2	4.2	7.67	4.2	4.4	7.4
0.7	3.3	4.3	8	4.3	4.6	7.6
0.6	3.4	4.4	8.14	4.4	4.8	7.8
0.5	3.5	4.5	8.29	4.5	5	8
0	4	5	9	5	6	9

TABLE IV. THE NORMALIZED RANKING PFNS

Normalized Ranked PFNs / level	R(A)	R(B)
0	0.21	0.18
0.5	0.295	0.276
0.6	0.316	0.3
0.7	0.337	0.335
0.8	0.363	0.37
1	0.448	0.445
1	0.538	0.58
0.8	0.66	0.7
0.7	0.723	0.765
0.6	0.783	0.83
0.5	0.845	0.91
0	1.2	1.33

Now, applying the above centroid theorem for PFN, then the final crisp rank value for each alternative is found as follows:

COA(R(A)) = 0.5244 ,

COA(R(B)) = 0.468.

Thus it is clear that alternative A should be selected as the best choice.

V. CONCLUSION

This paper introduced a way for adopting PFNs in FMCDM problems. The proposed forms of PFN ensure its generality over other popular forms of fuzzy numbers. A general formula for ranking PFNs is given using the centroid method.

The arithmetic operations for PFNs are clearly illustrated and an illustrative example shows how to adopt such generalized fuzzy numbers for solving FMCDM problems.

REFERENCES

[1] Dug Hun Hong, Strong laws of large numbers for t-norm-based addition of fuzzy set-valued random variables, Fuzzy sets and systems,(2013) vol. 223, pp 449-728.

[2] Shyi-Ming Chen ,Cheng-Yi Wang, Fuzzy decision making systems based on interval type-2 fuzzy sets, Science Direct, Applied Mathematical Modeling, (2013).

[3] F. Herrera, E. Herrera-Viedma, Spain Linguistic decision analysis: steps for solving decision problems under linguistic information, Computer Science and Arti cial Intelligence (2000) 115, pp 67-82.

[4] Shi-Jay Chen and Shyi-Ming Chen, Fuzzy Risk Analysis Based on Similarity Measures of Generalized Fuzzy Numbers, IEEE Transaction on Fuzzy systems (2003), Vol. 11, No. 1 , pp. 45-56.

[5] Dong Yong, Shi Wenkang, Du Feng, Liu Qi, A new similarity measure of generalized fuzzy numbers and its application to pattern recognition, Pattern Recognition Letters, 25, (2004), pp. 875-883.

[6] Shyi-Ming Chen, Jim-Ho Chen, Fuzzy risk analysis based on ranking generalized fuzzy numbers with different heights and different spreads, Expert Systems with Applications, 36, (2009), pp. 6833-6842.

[7] Chia-Chi Sun, A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods, Expert Systems with Applications 37 (2010) 7745–7754.

[8] Mohammad Anisseh , Fatemeh Piri , Mohammad Reza Shahraki ,Fazlollah Agamohamadi, Fuzzy extension of TOPSIS model for group decision making under multiple criteria, Springer 2011.

[9] E. Roghanian, J. Rahimi, A. Ansari, Comparison of first aggregation and last aggregation in fuzzy group TOPSIS, Applied Mathematical Modelling 34 (2010) 3754–3766.

[10] Yu-Chuan Chang, Shyi-Ming Chen, Churn-Jung Liau, Fuzzy Interpolative Reasoning for Sparse Fuzzy-Rule-Based Systems Based on the Areas of Fuzzy Sets, IEEE Transaction on Fuzzy Systems, Vol. 16, No. 5, pp. 1285-1301, October 2008.

[11] Cheng, C. H., A new approach for ranking fuzzy numbers by distance method, Fuzzy Sets and Systems (1998). 95, 307–317.

[12] Wang, Y.M., Yang, J.B., Xu, D.L., and Chin, K.S., On the centroids of fuzzy numbers, Fuzzy Sets and Systems (2006) 157, pp: 919-926

Telugu Bigram Splitting using Consonant-based and Phrase-based Splitting

T. Kameswara Rao
Assoc. Professor and Head, CSE Dept
Brahma's Inst. of Engg. and Tech
Rajupalem, Nellore, AP, India

Dr. T. V. Prasad
Former Dean of Computing Sciences,
Visvodaya Technical Academy,
Kavali, AP, India

***Abstract*—Splitting is a conventional process in most of Indian languages according to their grammar rules. It is called '*pada vicchEdanam*' (a Sanskrit term for word splitting) and is widely used by most of the Indian languages. Splitting plays a key role in Machine Translation (MT) particularly when the source language (SL) is an Indian language. Though this splitting may not succeed completely in extracting the root words of which the compound is formed, but it shows considerable impact in Natural Language Processing (NLP) as an important phase. Though there are many types of splitting, this paper considers only consonant based and phrase based splitting.**

Keywords*—Bigram; n-gram; consonant based splitting; phrase based splitting

I. Introduction

Combining / conjunction of two or more words to form bigrams or n-grams is a conventional process in Indian languages which plays an important role [1], for instance, '*vibhakti*' (inflection) attachment to a root word that can be noun, pronoun, verb, etc. Inflections become postpositions and they are attached to the rear end of the root word. In many foreign languages like English, French, etc., inflections are prepositions and they are separate words. This is the reason why foreign languages strictly maintain word order. If the word order is changed, then the meaning of the sentence will be changed. For example, in the sentence 'Krishna is playing with snake', the word 'with' is preposition and related to 'snake'. The word order of 'with' and 'snake' should not be disturbed, and if changed, it may yield a sentence like 'snake is playing with Krishna'. Now, this sentence may be grammatically valid but gives incorrect meaning and objective of the sentence is changed.

But in most of the Indian languages, word order is negotiated [5] since there is no change in meaning as the inflections become part of the words. e.g. '*kRshNuDu pAmutO ADutunnADu*'. Here '*kRshNuDu*' is the Subject, '*pAmu*' is the Object, '*tO*' is the inflection, '*ADu*' is the verb and '*tunnADu*' is the tense and gender describer. Inflection and object are combined together to form single word using conjunction rules and meaning does not change whatever the word order may be. The above sentence can also be written with no change in meaning as:

1. '*pAmutO kRshNuDu ADutunnADu*'
2. '*pAmutO ADutunnADu kRshNuDu*'
3. '*ADutunnADu kRshNuDu pAmutO*'
4. '*ADutunnADu pAmutO kRshNuDu*'
5. '*kRshNuDu ADutunnADu pAmutO*'

This happens because of attaching inflection with root word i.e. '*tO*' with '*pAmu*'. If these two are not combined, word order affects the sentence considerably and may change the meaning or become meaningless. For example, '*kRshNuDu pAmu tO ADutunnADu*'. This can be written as '*pAmu kRshNuDu tO ADutunnADu*' (absurd meaning). Another example is '*rAmuDi valana rAvANuDi cAvu*' (*rAvaNa*'s death is due to *rAma*). Here '*valana*' is inflection and relates strictly with '*rAmuDi*'. Ignoring word order, in case this sentence is written as '*rAvANuDi valana rAmuDi cAvu*' (*rAma*'s death is due to *rAvaNa*) and the meaning is drastically changed.

All these examples conclude that when the inflections are properly attached to appropriate root words, then the word order cannot be an obligation, or else word order changes the meaning in a wrong direction and may render the sentence meaningless. The primary objective of the MT is to maintain the meaning. But in general, databases or dictionaries do not contain words with their inflection forms. As a consequence, splitting of those compounds is a mandatory step in MT to improve ease as well as accuracy in translation.

II. Issues in Conjunction and Splitting

Just as the issue of handling splitting leading to word order was discussed above, similar issue is faced with conjunctions also. Translators must be aware of when and where conjunctions are to be and not to be employed. If not, either unnecessary meanings are generated or sentence becomes either meaningless or non-informative. Two instances are given, one each for conjunction and splitting:

- ***Issues in conjunction:*** When the sentence. 'doMga rAmuNNi koTTADu' (literally meaning thief beat rAma) is examined, the word 'doMga' is subject and a noun, 'rAmuNNi' is inflected object and 'koTTADu' is verb and also acts as gender representative. There is nothing wrong in combining 'rAmuNNi' and 'koTTADu' to form a compound 'rAmuNNikoTTADu' from the meaning's perspective. Issues arise if the words 'doMga' and 'rAmuNNi' are combined to form the compound 'doMgarAmuNNi' which literally means 'thief natured rAma', a disturbed meaning. This happened since the noun 'doMga' is converted in to an adjective when it is combined with

'rAmuNNi'. Moreover, since the sentence missed the subject, one cannot understand who has hit 'rAma'.

- ***Issues in splitting:*** Examine the sentence 'lakshmi piccikukkani caMpinadi' (literally meaning lakshmi killed mad dog). Here 'lakshmi' is subject and is a noun, 'kukkani' is object and is inflected and 'caMpinadi' is verb as well as tense and gender representative. Splitting of 'piccikukkani' is to be done in such a way that it is to be considered as a whole word for translation. Otherwise, mere splitting in the sentence may result as 'lakshmi picci kukkani caMpinadi' (literally means lakshmi's madness killed the dog) which is absurd though grammatically correct.

III. SANDHIS AS AN AID FOR SPLITTING

Amongst all Indian languages, Sanskrit and Telugu have well structured and numerous grammar rules [4]. Especially, the richness of Telugu language is with its huge number of words which can help express the meaning and mood much precisely [6]. Translation of compound-words (or simply compounds) or n-grams can be an obligation in MT as they are not available in database as they are. It becomes an overhead to maintain a database that consists of every possible conjunctional combination of n-grams. It is, therefore, impossible to translate compounds without splitting.

Though conjunction (known as '*sandhi*' in Sanskrit, as well as in Telugu) is seemed to be combining of two words (*pUrva-pada and uttara-pada*), actual '*sandhi*' occurs between only two letters, i.e. last letter of the first word ('*pUrva-svara'*) and first letter of the second word ('*para-svara*'). A *sandhi* will results in at least one of the following

- Concatenation of pUrva-pada and para-pada
- Either pUrva-svara or para-svara is dropped
- A new vowel / consonant is inserted
- Some specific words ***are inserted***

This paper handles only consonant and specific word issues.

'*sandhis*' are categorized in to five types in Sanskrit. They are 1. '*ach sandhis*' 2. '*prakRti bhAva sandhis*' 3. '*hal sandhis*' 4. '*visarga sandhis*' and 5. '*svAdi sandhis*'.

Among all these categories, only '*hal sandhis*' are considered in this paper as they involves necessarily a consonant (consonant is called '*hal*' in Sanskrit) in compound as a result. These '*hal sandhis*' are listed in table 1.

TABLE I. SANSKRIT '*HAL SANDHIS*' AND THEIR RESULTANT CONSONANTS

S.No	'*sandhi*' name	Resultant Consonant
1	*Scutva sandhi*	*S, c, ch, j, Q*
2	*shTutva sandhi*	*sh, T*
3	*jastva sandhi*	*g, j, D, d, b*
4	*anunAsika sandhi*	*G, Q, N, n, m*
5	*pUrva savarNa sandhi*	*ggh, jjh, DDh, ddh, bbh*
6	*para savarNa sandhi*	*Ll*
7	*chatva sandhi*	*Cch*

Though Telugu adopted all '*sandhis*' from Sanskrit grammar, it has its own '*sandhis*' as well as their precise formulae. Some of them involve only vowels [7], some of them involve only consonants and some involves both vowels and consonants. Later two cases are briefly considered as consonant resultant '*sandhis'* in this paper. Table 2 describes the list of Telugu consonant '*sandhis*'.

There are some other '*sandhis'* which works with phrases. They are discussed in later sections.

TABLE II. TELUGU '*HAL SANDHIS*' AND THEIR RESULTANT CONSONANTS

S.No	'*sandhi*' name	Resultant Consonant
1	*yaDAgama sandhi*	*y*
2	*dviruktaTakAra sandhi*	*TT*
3	*gasaDadavAdESa sandhi*	*g, s, D, d, v*
4	*druta / saraLAdESa sandhi*	*g, j, D, d, b*
5	*pumpvAdESa sandhi*	*p / Mp*
6	*penvAdi sandhi**	*nn*
7	*AmrEDita sandhi*	*TT / rr / ll/ tt*
8	*pampavarNAdESa sandhi*	*Pa*
9	*trika sandhi*	Two consonants other than *S,sh,s,h*
10	*lu la na la sandhi*	Two consonants
11	*dugAgama sandhi*	*du*
12	*nakArAdESa sandhi*	*nn / NN*

*This '*sandhi*' is listed in phrase based splitting also

IV. CONSONANT BASED SPLITTING

Consonant based splitting is not a special kind of splitting rather finding the possibilities to split a compound with the help of a consonant. When the consonants which are listed in Table 2 are encountered in a compound, splitting the compound by using appropriate '*sandhi'* rules yields good results in extracting root words with which the compound is formed. Except in some special cases, in majority of cases, this process is successfully extracted the root words of the compound. This paper deals with splitting rather than '*sandhi*' formation. In view of this, deep explanation about '*sandhi'* or compound formation is negotiated.

1) yaDAgama sandhi: This 'sandhi' involves the consonant 'y' as a result in compound. (See Table 3)

TABLE III. EXAMPLES OF *YADAGAMA SANDHI*

S.No	Root words	Compound	Replace 'y' with
1	*mA + amma*	*mA**y**amma*	A whitespace
2	*mI + illu*	*mI**y**illu*	A whitespace
3	*mA + Uru*	*mA**y**Uru*	A whitespace
4	*hari + ataDu*	*hari**y**ataDu*	NA
5	*tella + Enugu*	*tella**y**Enugu*	NA

When a '***y***' is observed in the compound, Ensuring the previous vowel of the consonant '*y*' is a long vowel (in this case '*A*' or '*I*') then applying *yaDAgama sandhi* rules in splitting yields good results in extracting root words.

Splitting: In this case root words can be extracted by replacing '*y*' with a whitespace (i.e. removal of '*y*'). Splitting has to be done before '*y*' and '*y*' is replaced with a whitespace where second root word starts.

2) dviruktaTakAra sandhi: This 'sandhi' involves the consonant 'TT' as a result in compound. dviruktaTakAra

literally means the letter 'Ta' is twice derived. Examples are given in Table 4.

TABLE IV. EXAMPLES OF *DVIRUKTATAKARA SANDHI*

S.No	Root words	Compound	Replace 'TT' with
1	*kuru + usuru*	*ku**TT**usuru*	*ru* + whitespace
2	*ciru + aDavi*	*ci**TT**aDavi*	*ru* + whitespace
3	*kaDu + eduru*	*ka**TT**eduru*	*Du* + whitespace
4	*naDu + aDavi*	*na**TT**aDavi*	*Du* + whitespace
5	*niDu + Urpu*	*ni**TT**Urpu*	*Du* + whitespace
6	*naDu + illu*	*na**TT**illu*	*Du* + whitespace

When '*TT*' is observed in the compound, *dviruktaTakAra sandhi* rules are applied in splitting to extract root words.

Splitting: '*TT'* can be suitably replaced with '*ru / Du*' and split. There are two special cases in this '*TT*' issue, as mentioned in Table 5 and 6.

TABLE V. SPECIAL CASE 1 OF HANDLING '*TT*'

S.No	Root words	Compound	Replace with
1	*ciTTi + eluka*	*ci**TT**eluka*	*TTi* + whitespace
2	*ciTTi + aDavi*	*ci**TT**aDavi*	*TTi* + whitespace
3	*ciTTi + Amudamu*	*ci**TT**Amudamu*	*TTi* + whitespace
4	*ciTTi + Idu*	*ci**TT**Idu*	*TTi* + whitespace
5	*ciTTi + uDuku*	*ci**TT**uDuku*	*TTi* + whitespace

*In the above mentioned special cases, '*TT*' can be replaced with '*TTi*' and split.

TABLE VI. SPECIAL CASE 2 OF HANDLING '*TT*'

S.No	Root words	Compound	Replace 'TT' with
1	*ciTTi + cApa*	*ci**TT**icApa*	NA
2	*ciTTi + pApa*	*ci**TT**ipApa*	NA
3	*ciTTi + tALamu*	*ci**TT**itALamu*	NA

In the above case, '*parasvara*' is consonant. In a compound, '*parasvara*' is neither available nor identifiable. For this case, vowel based splitting gives better results. Replacement for '*TT*' is not applicable in this case, rather separating '*ciTTi*' as a root word. Splitting has to be done after 'ru/Du/TTi' where first root word ends.

Issue 1: '*TT*' based splitting fails in the case of '*ceTTekkaDa'* (literally means where tree is). It is known that its root words are '*ceTTu' + 'ekkaDa'*. But according to Table 4, it become '*ceDu' + 'ekkaDa'* (literally means where bad is) which gives an incorrect meaning after translation.

This can be avoided by applying vowel based splitting [7] which gives longest word among all combinations of root words, i.e '*ceTTu*' and '*ceDu*' can be formed according to split rules. But '*ceTTu*' is longer word than '*ceDu*' and is returned as a first root word.

Issue 2: This '*sandhi*' rules fail in splitting of the compound '*miTTamadhyAhnamu*', since it is formed with an unusual logic by combining the root words '*madhyAnamu' + 'madhyAnamu*'.

N.B: Better idea to translate this category of words is to consider them as a single word rather bigram and enter them in to database, e.g. *'ciTTaDavi, ciTTeluka, naTTaDavi.*

Note: More issues are discussed in Phrase based splitting.

1) gasaDadavAdESa sandhi: This 'sandhi' involves the consonant 'g/s/D/d/v' as a result in compound. (Table 7)

TABLE VII. EXAMPLES OF *GASADADAVADESA SANDHI*

S.No	Root words	Compound	Replace with
1	*rAru + kadA*	*rAru**g**adA*	A whitespace + *k*
2	*apuDu + caniye*	*apuDu**s**aniye*	A whitespace + *c*
3	*nIvu + Takkari*	*nIvu**D**akkari*	A whitespace + *T*
4	*mIru + talaci*	*mIru**d**alaci*	A whitespace + *t*
5	*vAru + pOduru*	*vAru**v**Oduru*	A whitespace + *p*
6	*ataDu + kalaDu*	*ataDu**g**alaDu*	A whitespace + *k*
7	*vADu + cEsenu*	*vADu**s**Esenu*	A whitespace + *c*
8	*Ame + tolagenu*	*Ame**d**olagenu*	A whitespace + *t*

When '*g/s/D/d/v*' is identified in the compound, *gasaDadavAdESa sandhi* rules are applied in splitting to extract root words.

Splitting: In this case, root words can be extracted by simply replacing '*g/s/D/d/v*' with '*k/c/T/t/p*' appropriately with a whitespace prefixed. Splitting has to be done before '*g/s/D/d/v*' where second root word starts.

Issue: '*tallidamDrulu*' can be split according to this '*sandhi'* rule by replacing '*d*' with a whitespace + '*t'*. But in the case of '*Akalidappulu*' (literally means hungry and thirsty) this sandhi fails in giving correct meaning after translation as it splits the word into '*Akali' + 'tappulu*'(literally means hungry and mistakes). This also fails when it try to split the compound '*cerukugaDa*'(literally means stem of sugarcane). It splits it into '*ceruku + kaDa*'(literally means near sugarcane / at the end of the sugar cane) which is an incorrect translation.

Same problem is repeated in the case of '*nOrujAru*' which is formed by two root words '*nOru, jAru*' (logically means tongue slip). If this rule is applied on this it splits as '*nOru, cAru*' (literally means juice of mouth, i.e. spit) which is an incorrect translation.

2) druta / saraLAdESa sandhi: This 'sandhi' involves the consonant 'g/j/D/d/b' as a result in compound. (Table 8)

TABLE VIII. EXAMPLES OF DRUTA / SARALADESA SANDHI

S.No	Root words	Compound	Replace with
1	*pUcenu + kamala*	*pUcenu**g**amala*	whitespace + *k*
2	*Kanenu + cukkalu*	*Kanenu**j**ukkalu*	whitespace + *c*
3	*cEsenu + Takkulu*	*cEsenu**D**akkulu*	whitespace + *T*
4	*namilenu + tamba*	*Namilenu**d**amba*	whitespace + *t*
5	*virisenu + padmam*	*Virisenu**b**admam*	whitespace + *p*
6	*nannu + cUci*	*nannu**j**Uci*	whitespace + *c*
7	*bhAryanu + cEse*	*bhAryanu**j**Ese*	whitespace + *c*
8	*kappanu + tine*	*Kappanu**d**ine*	whitespace + *t*

When 'g/j/D/d/b' is identified in the compound, *druta / saraLAdESa sandhi* rules are applied in splitting to extract root words.

Splitting: In this, root words can be extracted by simply replacing 'g/j/D/d/b' with 'k/c/T/t/p' appropriately with a whitespace prefixed. Splitting has to be done before 'g/j/D/d/b' and the consonant is replaced (as in Table 8) where second root word starts.

Issues: This sandhi can form the compounds in different ways. They are listed in Table 9.

TABLE IX. TYPES OF *DRUTA/ SARALADESA SANDHI* COMPOUND FORMATIONS

S.No	Root words	Compound	Replace with
1	*pUcenu + kamala*	*pUceM**g**amala/ pUcen**g**amala*	WS+ *k*
2	*kanenu + cukkalu*	*kaneM**j**ukkalu /kanen**j**ukkalu*	WS+ *c*
3	*cEsenu + Takku*	*cEseM**D**akkulu/cEsen**D**akku*	WS+ *T*
4	*namilenu + tamba*	*namileM**d**amba/namilen**d**amba*	WS+ *t*
5	*virisenu + padma*	*viriseM**b**adma /virisen**b**adma*	WS+ *p*
6	*nannun + cUci*	*nannuM**j**Uci /nannun**j**Uci*	WS+ *c*
7	*bhAryan + cEse*	*bhAryan**j**Ese /bhAryaM**j**Ese*	WS+ *c*
8	*kappan + tine*	*kappan**d**ine /kappaM**d**ine*	WS+ *t*

*WS stands for whitespace

N.B: To improve accuracy in splitting, it is to be ensured that if the previous vowel of '*g/j/D/d/b*' is '*M/n*' abefore applying *druta/ saraLAdESa sandhi* rule. If so, then first replace '*M/n*' with '*nu*' and then replace the consonant according to Table 9.

Issue 1: This '*sandhi'* rules fails in splitting the compound '*vEsenugAlamu*'. It is known that its root words are '*vEsenu + gAlamu*'(literally meaning anchored). But according to '*sandhi'* rules, root words become *'vEsenu + kAlamu'* (literally meaning time is thrown), an incorrect translation.

Issue 2: If the compound is '*vaccenugOvulu*' formed by the words '*vaccenu, gOvulu*' (literally meaning cows came). '*sandih*' rules changes '*gOvulu*' into '*kOvulu*' and then searches in Database which is not a proper word and is not available. As a consequence, the compound cannot be split. One more example for this type of failure is '*pADenugandharvuDu*' formed by '*pADenu, gandharvuDu*' (literally meaning '*gandharva'* sung).

Note: Sometimes '*pUrvapada*' is a past tense of a verb which cannot be a root word and is not available in Database.

E.g. '*pUcenu'* (blossomed), '*kanenu'* (saw / delivered), '*cEsenu'* (done). Their root words are '*pUyu*' (To blossom), '*kanu*' (To see / To deliver), '*cEyu*'(To do) respectively. Sometimes '*pUrva pada*' is terminated with a '*druta*'(half M i.e. *n,* e.g. *pucen*)[2] which is not available in Database.

If a morphological algorithm is developed to morph the word '*pUcenu*' into '*pUyu*' and '*nannun*' into '*nannu*' and so on properly, then this '*sandhi*' rules are suitable for splitting.

1) puMpvAdESa sandhi: This 'sandhi' involves 'pu/Mpu' as a result in compound. Examples are given in Table 10.

TABLE X. EXAMPLES OF *PUMPVADESA SANDHI*

S. No	Root words	Compound	Replace with
1	*sarasamu + mATa*	*sarasa**pu**mATa/sarasa**Mpu**mATa*	*mu* + W
2	*virasamu + cUpu*	*virasa**pu**cUpu /virasa**Mpu**cUpu*	*mu* + W
3	*nikkamu + nIlamu*	*nikka**pu**nIlamu /nikka**Mpu**nIlamu*	*mu* + W

*W stands for whitespace

When a '*pu/Mpu'* is observed in the compound, *puMpvAdESa sandhi* rules are applied for splitting.

Splitting: In this case root words can be extracted by replacing '*pu/Mpu*' with '*mu*' with a whitespace appended. Splitting has to be done next to '*mu'* after '*pu/Mpu*' replaced with '*mu*' appended with a whitespace where second root word starts.

2) penvAdi sandhi: This 'sandhi' involves 'nn' as a result in compound. (Table 11)

TABLE XI. EXAMPLES OF *PENVADI SANDHI*

S.No	Root words	Compound	Replace with
1	*penu + adurulu*	*Pe**nn**adurulu*	*nu* + whitespace
2	*kanu + Aku*	*ka**nn**Aku*	*nu* + whitespace
3	*anu + urvISuDu*	*a**nn**urvISuDu*	*nu* + whitespace
4	*penu + oDalu*	*pe**nn**oDalu*	*nu* + whitespace

If '*nn*' is observed in the compound, *penvAdi sandhi* rules are applied in splitting to extract root words.

Splitting: In this case root words can be extracted by replacing '*nn*' with 'n*u*' with a whitespace appended. Splitting has to be done next to '*nu'* after '*nn*' replaced with '*nu*' appended with a whitespace where second root word starts.

3) AmrEDita sandhi: This 'sandhi' also involves the consonant 'TT' as a result in compound like dviruktaTakAra sandhi. But this 'sandhi' forms compounds in three types but this paper considers only one type as the second type does not involve consonant as a result and third type is more ambiguous and the nature is unidentifiable which is discussed as a special issue in this concept. (Table 12)

TABLE XII. EXAMPLES OF *AMREDITA SANDHI*

S.No	Root words	Compound	Replace with
1	*kaDa + kaDa*	*Ka**TT**akaDa*	Handled with a different logic.
2	*civara + civara*	*Ci**TT**acivara*	
3	*eduru + eduru*	*E**TT**aeduru*	
4	*naDuma + naDuma*	*na**TT**anaDuma*	
5	*pagalu + pagalu*	*pa**TT**apagalu*	
6	*bayalu + bayalu*	*ba**TT**abayalu*	
7	*modaTa + modaTa*	*mo**TT**amodaTa*	
8	*tuda + tuda*	*Tu**TT**atuda*	
9	*kona + kona*	*Ko**TT**akona*	
10	*piDugu + piDugu*	*pi**TT**apiDugu*	

Case 1: When '*TT*' is observed in the compound then, search for the immediate next consonant in the compound not allowing to surpass two vowels. If only one vowel is placed between '*TT*' and its next immediate consonant, then check immediate previous consonant in the compound. If both are same then, extract a word starting from the next immediate consonant to the end of the compound; and identify the compound is formed of two words of that extracted word kind. See all examples of Table 12 except 3.

Case 2: If two vowels are passed in finding the next immediate consonant in compound, then identify the second vowel and identify the immediate previous vowel in the compound. If both are same, then extract a word starting from second vowel to the end of the compound; and identify the compound is formed of two words of that extracted word kind. See the third example in Table 12.

Issues: Special issues are given in Table 13.

TABLE XIII. SPECIAL ISSUES OF *AMREDITASANDHI*

S.No	Root words	Compound	Replace with
1	*iMkulu + iMkulu*	*irriMkulu*	NA
2	*iggulu + iggulu*	*irriggulu*	NA
3	*ceduru + ceduru*	*cellaceduru*	NA
4	*tuniyalu + tuniyalu*	*Tuttuniyalu*	NA
5	*miTlu + miTlu*	*mirumiTlu*	NA
6	*atuku + atuku*	*aMdatuku*	NA
7	*iMkulu + iMkulu*	*irriMkulu*	NA
8	*tumuru + tumuru*	*Tuttumuru*	NA

*NA stands for Not-Applicable

If these compounds are tried to split using this sandhi rules, they cannot extract proper root words as well as proper translation. However, if some rules are implemented to split above compounds, then there are very high chances of improper splitting of other compounds which are not formed based on this sandhi rule. For example, a rule is developed to split '*cellaceduru*' into '*ceduru + ceduru*' then it splits the compound '*tellateppalu*' which is formed of '*tella, teppalu*' (literally means white boats) into '*teppalu + teppalu*' which is an incorrect splitting.

NB: Entry of above category of words as it is in to the Database gives better results than splitting.

1) paMpavarNAdESa sandhi: This 'sandhi' involves 'pa' as a result in compound. Examples are given in Table 14.

TABLE XIV. EXAMPLES OF *PAMPAVARNADESA SANDHI*

S.No	Root words	Compound	Replace with
1	*nAmu + cEnu*	*nA**pa**cEnu*	*mu* + whitespace
2	*pAmu + rEDu*	*pA**pa**rEDu*	*mu* + whitespace
3	*janumu + nAra*	*janu**pa**nAra*	*mu* + whitespace
4	*vEmu + kAya*	*vE**pa**kAya*	*mu* + whitespace
5	*ammu + Sayya*	*am**pa**Sayya*	*mu* + whitespace
6	*inumu + kaDDi*	*inu**pa**kdaDDi*	*mu* + whitespace
7	*enumu + guMpu*	*enu**pa**guMpu*	*mu* + whitespace
8	*minumu + gAre*	*minu**pa**gAre*	*mu* + whitespace
9	*emmu + gUDu*	*eM**pa**gUDu*	*mu* + whitespace
10	*kanumu + cEnu*	*kanu**pa**cEnu*	*mu* + whitespace

When a '*pa*' is observed in the compound, *paMpavarNAdESa sandhi* rules are applied .

Splitting: In this case, root words can be extracted by replacing '*pa*' with 'm*u*' with a whitespace appended. Splitting has to be done next to '*pa'* after '*pa*' replaced with '*mu*' appended with a whitespace where second root word starts.

Issues: This '*sandhi'* rules fail when applied to split the compound '*pApakannu*' which is formed by two root words '*pApa, kannu*' (literally means baby's eye). '*pAmu, kannu*' (literally means snake's eye) will be the root words after splitting the compound using this rules which is an incorrect translation.

2) trika sandhi: This 'sandhi' involves two same consonants other than 'S,sh,s,h' as a result in compound. In Sanskrit grammar, 'a, i, e' are called 'trika'[3]. (See Table 15)

TABLE XV. EXAMPLES OF *TRIKA SANDHI*

S.No	Root words	Compound	Replace with
1	*A + kanya*	*A**kk**anya*	A precise procedure is used rather than mere replacement
2	*I + kAlamu*	*i**kk**Alamu*	
3	*E + lOkamu*	*e**ll**Okamu*	
4	*A + aSvamu*	*a**yy**aSvamu*	
5	*A + bhaMgi*	*a**bbh**aMgi*	
6	*I + adi*	*I**yy**adi*	
7	*I + dharaNi*	*i**ddh**araNi*	

When any two consonants other than '*S,sh,s,h*' are observed in the compound as in the above table, *trika sandhi* rules are applied in splitting to extract root words.

N.B: To apply this rule, ensure the immediate preceding character of two consonants is any one of the vowels '*a/i/e*'. Otherwise this rule is not applicable.

Splitting: In this case, root words are to be extracted by replacing first consonant along with its preceding short vowel with the long form of the vowel. Splitting has to be done between two consonants.

3) lu-la-na-la sandhi: This 'sandhi' involves two same consonants as a result in compound. This 'sandhi' is also called 'Du-varNalOpa sandhi' as it eliminates the rear consonant 'Du' of 'purvapada' in compound. In this case 'purvapada' is mandatorily 'mUDu'. (Table 16)

TABLE XVI. EXAMPLES OF *LU-LA-NA-LA SANDHI*

S.No	Root words	Compound	Replace with
1	*mUDu + jagamulu*	*mu**jj**agamulu*	First consonant of the two is replaced with '*Du'*
2	*mUDu + lOkamulu*	*mu**ll**Okamulu*	
3	*mUDu + Arulu*	*mu**yy**Arulu*	
4	*mUDu + maDugu*	*mu**mm**aDugu*	
5	*mUDu + kAru*	*mu**kk**Aru*	
6	*mUDu + pAtika*	*mu**pp**Atika*	

If two consonants are observed in the compound as in the above table, *lu-la-na-la sandhi* rules are applied if and only if the two preceding letters of two consonants are '*mu*'.

Splitting: In this case, root words are to be extracted by replacing first consonant with '*Du*'. Splitting has to be done between two consonants.

4) dugAgama sandhi: This 'sandhi' involves 'du' as a result in compound. Examples are given in Table 15.

TABLE XVII. EXAMPLES OF *DUGAGAMA SANDHI*

S.No	Root words	Compound	Replace with
1	*nI + karuNa*	*nI**du**karuNa*	'*du*' is replaced with a white space
2	*nA + nEramu*	*nA**du**nEramu*	
3	*tana + rUpu*	*tana**du**rUpu*	
4	*mana + sAyamu*	*mana**du**sAyamu*	
5	*tama + karuNa*	*tama**du**karuNa*	
6	*nA + rUpu*	*nA**du**rUpu*	
7	*mI + cUpu*	*mI**du**cUpu*	
8	*mA + snEhamu*	*mA**du**snEhamu*	

When 'du' is observed in the compound, *dugAgama sandhi* rules are applied in splitting to extract root words.

Splitting: In this case, root words are to be extracted by replacing '*du*' with a whitespace.

5) nakArAdESa sandhi: This 'sandhi' involves two same consonants either 'nn' or 'NN' as a result in compound. Examples are given in Table 18.

TABLE XVIII. EXAMPLES OF NAKARADESA SANDHI

S.No	Root words	Compound	Replace with
1	*mUDu + nALLu*	*mU**nn**ALLu / mU**NN**ALLu*	Different logic is applied
2	*reMDu + nALLu*	*re**nn**ALLu / re**NN**ALLu*	

When '*nn/NN*' observed in the compound as in the above table, *nakArAdESa sandhi* rule is applied in splitting to extract root words.

Splitting: In this case, root words are to be extracted by replacing first '*n/N*' of the pattern '*nn/NN*' with '*MDu'* + a whitespace. If the second consonant of the pattern is '*N*', then, change it in to '*n*'. Split two consonants.

Issues: This '*sandhi*' rule fails in yielding proper words in the case of '*vENNILLu*'(literally meaning hot water), a compound formed by two proper words '*vEDi, nILLu*'. This rule can split the compound as '*vEDu, nILLu*' (literally meaning requesting water) which is an improper splitting.

V. PHRASE BASED SPLITTING

Here the phrase is referred to be the '*pUrvapada'*. The '*sandhis*', like '*rugAgama sandhi'* of Telugu, directly deal with phrases instead of single vowel or consonant. If the phrase is identified, compound can be split by applying these '*sandhi*' rules. But over dependency on these rules leads to improper results. These '*sandhis'* are listed in Table 19.

TABLE XIX. LIST OF PHRASE BASED '*SANDHIS'*

S. No	'*sandhi*'	*sandhi* Phrases
1	*rugAgama sandhi*	*pEda, bIda, mudda, bAliMta, manuma, goDDu, komma, vidhava, dhIra and some words which ends with 'aMta' like SrImaMta, guNavaMta, dhImaMta, bhAgyavaMta,* etc.
2	*prAtAdi sandhi*	*prAta, krotta, lEta, pUta, pUvu, mIdu, muMdu, keMpu, cennu, kriMdu,* etc.
3	*penvAdi sandhi*	*penu, kanu, anu,* etc
4	*lu – la – na – la sandhi*	*mUDu*
5	*dviruktaTakAra sandhi*	*kuru, ciru, kaDu, niDu, naDu*

*4 and 5 '*sandhis'* are listed in '*hal sandhis'* section also.

In this phrase based '*sandhis*', phrase is '*pUrva-pada*' in majority of cases. In many instances, these '*sandhis'* does not result in the '*pUrva-pada*' as it is in the compound. For this reason, it is required to create a new phrase as an independent phrase or as a reference of '*pUrva-pada*'. A search is done for this phrase in compound for better results. Table 20 a,b,c, d & e describes the possibility of phrase based splitting.

TABLE XX. (A). POSSIBILITIES OF '*RUGAGAMA SANDHI'*

S.No	Root words	Compound	Phrase searched for	A/NA
1	*pEda + Alu*	*pEdarAlu*	*arAlu*	A
2	*bIda + Alu*	*bIdarAlu*	*arAlu*	A
3	*bAliMta + Alu*	*bAliMtarAlu*	*arAlu*	A
4	*goDDu + Alu*	*goDDurAlu*	*urAlu*	NA
5	*komma + Alu*	*kommarAlu*	*urAlu*	NA

Splitting: According to grammar rules, the '*para-pada*' of this sandhi is always '*Alu*'. For better results, ensuring the presence of the new phrase '*arAl/urAl*' in the compound is essential. Then, separate the root word from the compound as in the Table 20 (a). Next, as per rules, the second part's first letter i.e. '*r*' is to be removed to make it '*Alu*'. As the word '*Alu*' have two meanings – 'wife and lady', replacing '*r*' with a meaningless character (it becomes *@Alu/ #Alu / *Alu etc.*) avoids the ambiguity in the selection of the suitable meaning of '*Alu*'. For this, a rule is to be implemented to pick the word 'lady' from the database when '*Alu*' is preceded with a special symbol. If not, it becomes 'poor wife' instead of 'poor lady'.

In the case of '*dhanavaMturAlu/ guNavaMturAlu/ dhIrurAlu / bhAgyavaMturAlu*' etc. replace '*u*' with '*a*' of '*urAl*', and separate the root word. Remaining process can be same. These cases give improper meaning after splitting. Because '*goDDurAlu*' literally means a married lady who have no children. It can be merely split into '*goDDu*' + '*Alu',* when this rule is applied. This is acceptable according to grammar rule, but the meaning becomes 'buffalo lady'. Though ordinary splitting is applied on this compound, it can be split in to '*goDDu'* + *'rAlu*' literally means 'buffalo to fell' which is a misinterpretation. Better way to solve these splitting problems is, to include these compounds in database not considering them as bigrams.

TABLE XX. (B). POSSIBILITIES OF '*PRATADI SANDHI'*

S.No	Root words	Compound	Phrase searched	A/NA
1	*kriMdu + kaDupu*	*krIgaDupu*	*krI*	A
2	*kriMdu + kannu*	*krIgannu*	*krI*	A
3	*kriMdu + kAlu*	*krIgAlu*	*krI*	A
4	*kriMdu + toDa*	*krIdoDa*	*krI*	A
5	*krotta + cAya*	*kroMjAya*	*kroM*	A
6	*krotta + pasiDi*	*kroMbasiDi*	*kroM*	A
7	*krotta + mAvi*	*krommAvi*	*kro**	NA
8	*prAta + illu*	*prAyillu*	*prA**	NA
9	*lEta + dUDa*	*lEdUDa*	*lE**	NA
10	*pUvu + tOTa*	*pUdOTa*	*pU**	NA
11	*mIdu + kaDa*	*mIgADa*	*mI*	NA
12	*krotta + gaMDi*	*kroggaMDi*	*kro**	NA

Splitting: If the first three characters of a compound are '*krI',* then, they are to be replaced with '*krimda*' and if first four characters of a compound are '*kroM*', then they are to be replaced with '*krotta*' if and only if the immediate next character of the phrase are '*g/j/D/d/b*'. For these two cases. change it to '*k/c/T/t/p*' respectively.

Sometimes '*g/j/D/d/b*' is also included into the phrase to improve the quality in splitting. For example, '*muMgOpamu*' is the bigram formed by the root words '*muMdu' + 'kOpamu*'. For this case, phrase will be '*muMg*'. But only '*muM*' is replaced with '*muMdu*' and '*g*' *(/j/D/d/b*) of the phrase is changed to '*k*' (*/c/T/t/p*). Remaining process is same.

**Issues:* Though they are acceptable according to the grammar rules, sometimes they mislead the proper words. For example, '*prAraMbhasamayamu*' is the compound (literally means starting time). It can become '*prAta' + 'raMbhasamayamu'* (literally means the time of old *raMbha* – a divine dancer) when split using this rule where the phrase

used is '*prA*' as in the above table. This is incorrect translation.

TABLE XX. (C). POSSIBILITIES OF '*PENVADI SANDHI'*

S.No	Root words	Compound	Phrase searched for	A/NA
1	*penu + aduru*	*Pennaduru*	*Penn*	A
2	*penu + Uta*	*pennUta*	*Penn*	A
3	*kannu + Aku*	*kannAku*	*kann*	A
4	*penu + Eru*	*pennEru*	*penn*	A

Splitting: If the first four characters of a compound are '*penn/kann*' then replacing with '*penu/kannu*' is the phrase based splitting using '*penvAdi sandhi'* as in the above table.

Issues: '*pennidhi*' (literally means great treasure) is a compound formed by two root words '*penu'+ 'nidhi*'. For this case '*n*' is prefixed to the second word after splitting.

TABLE XX. (D). POSSIBILITIES OF '*LU- LA- NA- LA SANDHI*

S.No	Root words	Compound	Phrase to be searched for	A/NA
1	*mUDu + jagamulu*	*mujjagamulu*	*mujj*	A
2	*mUDu + yEDulu*	*muyyEDulu*	*muyy*	A
3	*mUDu + lOkamulu*	*mullOkamulu*	*mull*	A
4	*mUDu + pAtika*	*muppAtika*	*mupp**	A
5	*mUDu + kArulu*	*mukkArulu*	*mukk**	A
6	*mUDu + ciccu*	*muccciccu*	*mucc**	A
7	*mUDu + trIva*	*muttrOva*	*mutt*	A
8	*mUDu + cemaTalu*	*muccemaTalu*	*mucc**	A

Splitting: When first two characters are '*mu*' and next two characters are any same consonants, then '*mu*' is replaced with '*mUDu*' and separated. First character from the second word i.e. first consonant is removed as in the above table.

**Issues:* Sometimes a root word like '*muccaTa*' (literally means fondness) can also be split into '*mUDu' + 'caTa*' (an absurd) lead to incorrect translation.

TABLE XX. (E). POSSIBILITIES OF '*DVIRUKTATAKARA SANDHI'*

S.No	Root words	Compound	Phrase searched for	A/NA
1	*kuru + usuru*	*ku**TT**usuru*	*kuTT**	A
2	*ciru + aDavi*	*ci**TT**aDavi*	*ciTT**	A
3	*kaDu + eduru*	*ka**TT**eduru*	*kaTT**	A
4	*naDu + aDavi*	*na**TT**aDavi*	*naTT**	A
5	*niDu + Urpu*	*ni**TT**Urpu*	*niTT**	A

Some descriptions are given about this in previous section.

**Issues:* This sort of compounds highly conflicts with '*AmrEDita sandhi*' as in the Table 21.

TABLE XXI. COMPOUNDS OF DVIRUKTATAKARA AND AMREDITA SANDHIS

S.No	Compounds of *dviruktaTakAra sandhi* rule	Compounds of *AmrEDita sandhi* rule
1	*ka**TT**eduTa(kaDu + eduTa)*	*ka**TT**akaDa(kaDa + kaDa)*
2	*ci**TT**aDavi(ciru + aDavi)*	*ci**TT**acivara(civara + civara)*
3	*na**TT**aDavi (naDu + aDavi)*	*na**TT**anaDi(naDi+ naDi)*
4	*ni**TT**Urpu (niDu + Urpu)*	*ni**TT**aniluvu(niluvu + niluvu)*

Some issues of '*dviruktaTakAra sandhi'* are described above.

As the '*pUrva-pada*' does not appear in the compound except some portion, the compounds formed by the *AmrEDita sandhi* rule also looking same and is not possible to decide which rule is to be used to split. To avoid unnecessary issues raised in splitting the compounds, enter them in database as they are.

VI. CONCLUSION

Though there are many issues involved in consonant and phrase based splitting that lead to incorrect translation, these rules extracts root words from the bigrams considerably. Many problems can be solved either by implementing a rule which finds out the longest root word among many combinations or entry of critical compounds in to database. For instance, the compound '*muppYokaTi*' (literally means thirty one) formed by root words '*muppY' + 'okaTi*'. It can be split as '*mUDu'+ 'pYokaTi*'. If further splitting process continues, the word '*muppY*' will be available in database as it is a root word, and it can be separated. This becomes '*muppY' + 'okaTi*'. As the word '*muppY'* is longer than '*mupp'*, it becomes the result.

REFERENCES

[1] Malladi Krishna Prasad, "Telugu Vyaakaranamu", Sri Venkateswara Book Depot, 2012.

[2] Dr. Samudrala Vemkata Ramga Ramanujacharya, "Samskruta Vaani" Rohini Publications, 1997.

[3] Kambhampati Ramagopala Krishnamurti, "Telugu Vyaakaranamu", Sri Sailaja Publications, 1991.

[4] A.H. Arden, "A Progressive Grammar of the Telugu Language", 2nd Edition, Society for promoting Christian Knowledge, Madras, 1905.

[5] Robert Caldwell, "A Comparative Grammar of The Dravidian or South-Indian Family of Languages", 2nd Edition, 1875.

[6] T. Venkateswara Prasad and G. Mayil Muthukumaran Telugu to English Translation Using Direct Machine Translation Approach, Int. J. of Sc. & Engg. Investigations, Vol.2 Issue 12, Jan 2012

[7] T. Kameswara Rao and Dr. T.V. Prasad, " Key Issues in Vowel based Splitting of Telugu Bigrams", Int. J. of Adv. Computer Sc. App., Vol. 5, No. 3, 2014

APPENDIX

Pronunciation: Letter pronunciation is as in Table 20 and 21.

TABLE XXII. VOWEL PRONUNCIATION

RT	Usage as	RT	Usage as	RT	Usage as
a	a in – That	R	Ru in – Ruk	o	O in - Obey
A	a in - Father	Ru	roo in – roof	O	oa in - Roar
i	i in - His	e	e in - When	W	ou in - Shout
I	Ea in - Eagle	U	oo in – fool	aM	um in - sum
u	u in – Put	E	a in - Hate	aH	aH in – aH
U	oo in – fool	Y	I in - Ice		

RT means Roman Telugu

TABLE XXIII. CONSONANT PRONUNCIATION

RT	Usage as	RT	Usage as	RT	Usage as
k	C in – Cut	Dh	Dh in – Dhamaru	y	Y in – Yak
kh	che in – Ache	N	Na in – Jana gaNa mana	r	R in – Rat
g	g in - Dog	t	th in – Path	l	L in – Lip
gh	gh in -Vagha	th	th in - sthambh	L	l in – mahila
G	Gy in - Gyan	d	Th in – The	v	V in – Van
c	ch in – Catch	dh	dh in – dharma	S	S in – Sand
Ch	Ch in – Chunk	n	n in – Pen	sh	Sh in - Sharp
J	J in – Jar	p	P in – Pot	h	H in – Hen
Jh	Jh in - Jhaveri	ph	Ph in – Phal		
T	t in – Cat	b	B in – Bat		
Th	T in - Tagore	bh	bh in – prabhu		
D	d in – God	M	M in - Mat		

The capitalized letters should be pronounced with greater emphasis on them.

A Review of Text Messaging (SMS) as a Communication Tool for Higher Education

Dr. Daragh Naughton
Limerick Institute of Technology
School of Applied Science, Engineering & Technology
Department of Mechanical & Automobile Engineering
Moylish, Limerick, Ireland

***Abstract*—Since 2011, The Limerick Institute of Technology (School of Applied Science Engineering & Technology) has actively engaged in a course of research to determine if SMS can be used to (1) increase student's preparation for class, (2) increase their motivational levels towards learning and (3) to assist with memory retention. The purpose of this paper is to introduce the concept of the technology and to summarise the academic arguments that have been made both for and against the use of such technology for teaching and learning activities in higher education. A full quantitate and qualitative analysis of the research will take place in 2014. The project forms part of a school wide Scholarship of Teaching & Learning approach (SoTL). This paper will be of interest to academic managers, program managers, e-learning support staff, administrators and lecturers.**

Keywords—SMS; text message; higher education; motivation

I. INTRODUCTION

Since 2011, the School of Applied Science, Engineering and Technology at the Limerick Institute of Technology has used SMS as a communication tool as part of its institute wide Scholarship towards Teaching and Learning (SoTL). Almost 10,000 text messages have been delivered to 1st and 2nd year mechanical engineering students in that period. The pilot scheme was designed to increase student's motivation and preparedness for class. Motivational texts were scheduled and delivered using a unique SMS portal. The following discussion presents the background to the research and should be of interest to educational managers and lecturers in the cognitive areas of science, engineering, information technology and math. A full statistical analysis of the results will be available publically in January of 2014.

Supporting students at vulnerable points in their academic career has always presented challenges. The step-up in maturity required from post-primary to higher education can be difficult for students to manage. Reaching this generation irrespective of the maturity challenge also presents its own difficulties. Fortunately, the latest generation of undergraduates has grown up in a world of persuasive digital technology which has caused them to develop fundamentally different ways of thinking and processing information from their predecessors and teachers [1]. SMS offers the ability to communicate directly with the students in a personal manner and can assist with not only enhancing the first year learning environment [2] but with a raft of other directly related areas including (1) communication and administrative support [3], (2) teaching and learning support [3] and (3) encouraging interactivity [4].

As a communication tool, SMS allows senders to transmit short messages of 160 alphanumeric characters to any suitable receiving device operating over the GSM network. The fact that children are growing up with wireless mobile technologies with positive outcomes in modern society is an indicator that SMS technologies should be embedded in the epistemologies of modern universities [5]. The potential success of any SMS education program is reinforced by the ever-growing popularity of mobile devices. In the UK alone, OFCOM reported that 96% of 15-24 age groups own GSM enabled devices [6] while a study at the University of Dundee found that 96% of students in their university owned a mobile phone [7]. Further research by Harley *et al.* reported that students were happy to receive pastoral support via SMS from university staff [8].

Use of SMS in education supports students to learn in "no fixed location or time" [9] and therefore the learning is facilitated at a time and location which suits the students (mobile devices have the capability to store and retrieve information). The learning aspect of SMS can operate on either a *push* model or a *pull* model [10]. In a push model, the teacher dictates which information is sent and at what time; the student is not afforded the opportunity to reply. Conversations are one way and are limited to motivational and preparatory elements of the lesson plan. The pull model is a closed loop system where students can reply, give answers and receive feedback. Other usages of SMS services include fill in the blanks, true false questions, multiple choice questions etc. [11]. The current research suggests that being accessible, contextualised and collaborated are the main appeal for mobile phones in learning [12]. The learning curve and familiarity for the periphery devices is also reduced given the overwhelming number of students who own and operate phones, therefore, mobile devices can be more easily integrated across the curriculum than desktops or SMART™ boards [13]. Also, as these devices are convenient and provide an expedited learning experience in student appropriate locations.

II. WHERE IS THE TEACHING & LEARNING EVIDENCE?

Evidence within the higher education sphere includes but is not limited to the following:

Evidence of SMS as a teaching aid in higher education: Wallace has used SMS successfully to teach business

information systems development to undergraduates at the University of the West of England. He concluded that the simplicity of the technology made it viable and that it "encouraged good development practices" amongst the final year undergraduates [14]. Hagos successfully used SMS technology to promote on-going learning in the area of mathematics. The research concluded that SMS proved to be a valuable asset in focusing the students' attention towards the learning of mathematics even when the lecturer was late or absent from class [15]. So, a researcher at the Hong Kong Institute of Education developed a bespoke SMS texting service and used the system to increase participation amongst students on a Bachelor of Education degree program. Although the research lacked supporting data, it is an excellent example of the technical development required to integrate SMS into the Teaching Tutee Tool paradigm [3].

Evidence of SMS as a transitional support technology: Jones *et al.* used SMS as a support tool to aid the transition of post primary students to higher education. The research suggested that students reacted differently to the SMS service but that overall the advantages outweighed the disadvantages. Interestingly, Jones *et al.* found no evidence to suggest that students saw texting as in intrusion into their personal space [2].

Evidence of SMS as an encouraging communication tool: In a survey of 532 students, Leung found those students who used SMS the most were motivated by its convenience, low-cost and its utility for coordinating events. The research also found evidence to suggest that SMS helped students overcome shyness and an un-willingness to communicate in an educational setting [16].

Evidence of SMS as an advocacy service: Young *et al.* used SMS texting to provide an advocacy service for student nurses, occupational therapists and radiographers on clinical placement in the United Kingdom. The research found that the students and staff embraced the technology but concerns were raised about the uptake of the service. The research concluded that safeguards must be established to ensure inclusion of all (even those who do not own their own personal phones) [17].

Evidence of SMS as an administrative support tool: Nordin *et al.*, albeit with a limited sample size, suggests that school administrators view the use of SMS as an appropriate communication tool for learner-teacher activities. 80% of those administrators surveyed responded positively to SMS as a teaching and learning assistive protocol [18].

Example of SMS as a tool for encouraging interactivity: Researchers from Trinity College Dublin successfully used SMS to successfully increase the interactivity of higher education students drawn from undergraduate and postgraduate sample pools. The "PLS TXT UR *Thoughts*" project concluded that SMS led to a more active learning environment, a greater provision of feedback for lecturers and increased student interest and motivation.

III. Advantages of SMS For Higher Education

Advantages of SMS technology within the educational context include:

- Low cost of operation from an academic institution's perspective as the technology is owned by the students [6].
- Instant communication and real time feedback is possible [19].
- Transactional distances and geographical barriers are no longer constraints [20].

Students readily accept the use of SMS as a communication device and research has indicated that communicating on such a personal level helps to foster a sense of community and gives a sense of "belonging to the university" [21] [8]. Jones et al. produced a body of research which indicated that SMS usage makes a valuable contribution towards the teacher as a "facilitator" or "broker" and can help to motivate participation in appropriate activities [2]. Overall, it would appear that the advantages of using SMS for supporting higher education outweigh the disadvantages.

IV. Disadvantages of SMS For Higher Education

The disadvantages of SMS within the learning environment have been extensively reviewed by Jones *et al.* [2]. Issues can arise over the ownership and control of the periphery device. Despite the teachers best intentions, intrusions into class time can detract from the teachers overall control of the class and the students focus. Mifsud has termed this the "intruder" effect [19]. Also, mobile phones tend to have small screens and the length of the messages is limited to 160 characters. This makes it difficult to communicate as effectively compared to conventional devices (e-mail etc.). Other researchers have raised concerns about the cost of such a program although the cost issue has failed to make a strong argument against the use of the technology when comparisons are drawn against other technological devices such as student response systems and PDA's etc. [4].

Sharples has called for caution and suggests that the introduction of mobile learning into an academic environment is not a panacea as it can bring problems as well as solutions; in particular the perception that young people may see the use of SMS for formal learning as an attempt to colonise and intrude on social spaces [20]. Markett *et al.* found that the interaction between face-to-face teaching methods and challenges set by SMS protocols became blurred and provided an unwelcome distraction [4]. He concluded that this may make the transition to higher education more challenging for the user. Horstmanshof has concluded that the additional workload in establishing a new communication media may prove difficult for existing teaching staff [21]. He also concluded that students may become addicted to the over-dependency that SMS communication may foster; and that this in turn may hamper the "natural development of more self-regulatory strategies". Noble has concluded that the use of such technology can turn universities into "digital diploma mills" and reminds the reader that education is a process that necessarily entails an interpersonal (not merely interactive) relationship between people [22].

V. Conclusions

Although SMS has some detractors, there is no doubt that it can have significant benefits for educators and education management. Anecdotal evidence from the research currently being conducted at the Limerick Institute of Technology suggests that students associate similar gratification elements of SMS usage to other communication media such as TV or the internet. Significantly, students have also indicated that they do not consider SMS as an intrusion into their personal

space. Some interesting topics have been raised by the current research; such as the ability of SMS to function equally well in classrooms where varying contradictory educational paradigms exist, in particular both behaviorism and constructivism. When available and fully analysed, the results of the current research and its limitations will be made public in keeping with The Limerick Institute of Technology's commitment to the SoTL framework.

ACKNOWLEDGMENT

The author wishes to express gratitude to the Limerick Institute of Technology Development Office for facilitating this course of research.

REFERENCES

[1] D. O. J. Oblinger, "Educating the next generation," 2005. [Online]. Available: www.educause.edu/books/educatingthenextgeneration/5989.

[2] G. E. G. R. A. Jones, "How can mobile SMS support and enhance a first year undergraduate learning environment?," Research in learning technology, vol. 17, no. 3, pp. 201-218, 2009.

[3] S. So, "The development of a SMS based teaching and learning system," Journal of Educational Technology Development and Exchange, vol. 2, no. 1, pp. 113-124, 2009.

[4] C. S. I. W. S. T. B. Markett, "Using short message service to encourage interactivity in the classroom," Computers and Education, vol. 46, pp. 280-293, 2006.

[5] I. Y. S. L. C. Tsai, "Exploring the course development model for the mobile learning context: a preliminary study," Taiwan, 2005.

[6] OFCOM, "The consumer experience research report," 2007. [Online]. Available: www.ofcom.org.uk/research/tce/ce07/researcho7.pdf.

[7] University of Dundee, "Mobile Communications Survey," 2006. [Online]. Available: www.dundee.ac.uk/elecengyphics/mobilesurveyresults.php.

[8] D. W. D. P. S. W. P. Harley, "Using texting to support students' transition to university," Innovations in Education and Teaching, vol. 44, no. 3, pp. 229-241, 2007.

[9] Kinshunk., "Adaptive mobile learning technologies," 2003. [Online]. Available: www.kcweb.org.uk/weblibrary/M-learning.pdf. [Accessed 12 01 2012].

[10] I. K. A. K. D. U. H. Yengin, "Is SMS still alive for education: Analysis of Educational potentials of SMS technology?," Procedia Computer Science, vol. 3, pp. 1439-1445, 2011.

[11] TxtTools, "SMS Text Messaging for Education," 2008. [Online]. Available: www.texttools.co.uk.

[12] S. Geddes, "Mobile learning in the 21st century; Benefits for Learners," The Knowledge Tree, vol. 6, 2006.

[13] D. H. S. Mosoley, "Ways forward with ICT effective pedagogy using information and communications technology for literacy and numeracy in primary schools," Newcastle, 1999.

[14] C. Wallace, "Teaching information systems development with SMS," in The 9th Java and the Internet Computing Curriculum Conference (JICC9), London Metropolitan University, London , 2005.

[15] L. Hagos, "Enhancing teaching and learning through SMS-mediated lectures in mathematics," in Hybrid learning: A new frontier ICHL, Hong Kong, 2008.

[16] L. Leung, "Unwillingness-to-communicate and college student's motives in SMS mobile messaging," Telematics and Informatics, vol. 24, pp. 115-129, 2007.

[17] P. M. E. G. G. R. R. S. R. C. M. F.-P. M. Young, "Help is just a text away: The use of short message service texting to provide an additional means of support for health care students during practice placements," Nurse Education Today, vol. 30, pp. 118-123, 2010.

[18] M. H. I. Y. N. E. M. Nordin, "The mobile learning environment for the in-service school adminstrators," Procedia social and behavioural sciences, vol. 7, no. C, pp. 671-679, 2010.

[19] J. Attewell, "Mobile Technologies and Learning; a technological update and m-learning project summary," Learning and Skills Development Agency, London, UK, 2005.

[20] P. C. A. Gorsky, "A critical analysis of transactional distance theory," The Quarterly Review of Distance Education, vol. 6, no. 1, pp. 1-11, 2005.

[21] L. Horstmanshof, "Using SMS as a way of providing connection and communication for first year students," in Beyond the comfort zone; Proceedings of the the 21st ASCILITE conference., Perth, Western Australia, 2004.

[22] L. Mifsud, "Alternative learning areanas-pedagogical challenges to mobile learning in education," in IEEE international workshop on wireless and mobile technologies in Sweden, Vaxjo, Sewdwn, 2008.

[23] M. Sharples, "Big issues in mobil learning," University of Nottingham, Learning Sciences Research Institute, Nottingham, UK, 2007.

[24] D. Noble, "Rehersal for the revolution," in Theoretical principles of distance education, Oxon, Routledge, 2002, pp. 22-38.

Web and Telco Service Integration: A Dynamic and Adaptable Approach

Julián Rojas, Leandro Ordóñez-Ante, Juan Carlos Corrales
Telematics Engineering Group
University of Cauca
Popayán, Colombia

***Abstract*—The current evolution of the Web, known as Web 2.0 and characterized by providing a diverse global service ecosystem, has marked a change in the role played by telecom operators. In order to maintain high competitive market dynamism and generate new revenue sources, many operators seek to leverage the wide variety of existing Web services and integrate them with its infrastructure capabilities. Such integration leads to various challenges from a technological perspective, where the heterogeneity on networks and the need for highly qualified personnel for the development of these services are highlighted. This paper is propose the definition of a mechanism for integrating both Web and Telco services which facilitates and speeds the development of new services, considering the dynamic conditions of its execution.**

Keywords*—Web Services; Telco Services; JAIN SLEE; Integration; Adaptation

I. Introduction

Currently there is a trend in the telecommunications industry that has created a scenario in which a new model known as Telco 2.0 [1] has been defined. This model relates the concepts, services and Web 2.0 technologies with traditional telecommunications features (Telco services), allowing operators to expand their service portfolio and have a greater impact on the market by reaching end-users with more complex and personalized services. These new kind of services are known as converged services due to its integration of functionalities form the telecommunications domain (voice, video and data) with Information Technology (IT) services from the Web domain.

Such integration demands complex and robust platforms that support the interoperability of different technologies and communication protocols, characteristic of services from both Telco and Web domains. Different approaches have been proposed such as the SIP Servlets Specification [2], the Ericsson Converged Service Studio [3], Alcatel-Lucent uReach CSF (Converged Services Framework) [4], among others. An alternative that stands out is the JAIN SLEE Specification [5], which proposes a standard and robust environment for the creation and execution of converged services, meeting the rigorous performance requirements typical of Telco services (high availability, low latency, asynchronous behavior, etc.) and allowing its integration with Web technologies.

Generally Telco services belonging to an operator reside on platforms located within its network, where they are managed and executed. On the other hand, Web services tend to be distributed applications that belong to 3rd party providers, therefore integrating them to the operator infrastructure, requires the development of modules that represent them and manage the data interaction present during their invocation. This implies an extra effort on developing terms considering that, for each Web service that wants to be integrated, a module must be designed and built.

Another issue related to the integration of Web and Telco services over an operator network is that due to the distributed nature of Web services, when failures occur on runtime, the operator usually does not have access to the platforms where Web services reside to address and correct them. Such failures become a major issue and could compromise the correct operation of converged services that relay on those Web services. This issue poses the need for operators to implement contingency mechanisms that adapt to Web service invocation failures at runtime, in order to provide high quality converged services.

This paper introduces the definition of a mechanism for integrating Web and Telco services over a converged platform, specifically JAIN SLEE. The proposed mechanism facilitates the inclusion of Web services into telecommunications networks by defining an interaction model between JAIN SLEE and Web environments that allows integrating different Web services without having to develop additional modules for each Web service that wants to be integrated. The interaction model also considers the dynamic conditions of Web services invocation by detecting possible failures at runtime and adapting to alternative ones, defined by the converged service designer at creation time. The rest of the paper is arranged as follows. Next section presents a conceptual base of the different technologies related with this work. Section 3 describes different proposals that address the integration of Web and Telco services. On section 4 is presented a detailed description of the proposed integration mechanism. Section 5 presents the fault handling and adaptation functions included in the mechanisms. Section 6 presents a case study through which the proposed mechanism is evaluated. Finally, on section 7 are presented the conclusions and future work.

II. Background

The integration of Web and Telco services contemplates the use of diverse technologies that comprise both domains. Telecommunications protocols such as SIP, SS7 stack and

SMPP, and Web frameworks like WSDL/SOAP services and Apache Axis2 are among the technologies that must be taken into consideration. Is also important the platform that supports the integration of these services, which determines how converged services are built and executed. As stated before, the selected platform for developing this proposal is the JAIN SLEE specification. Next is presented a description of the main technologies and concepts that comprises this proposal.

A. Web Services

According to the W3C (World Wide Web Consortium) definition, a Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards [6]. Following such definition, Web services can be characterized as distributed software components which can be described, published, discovered and invoked with standard protocols. Web services communicate using XML and Web protocols, working internally and across the Internet. They support heterogeneous interoperability and use SOAP for service calls and WSDL for service descriptions [7].

B. Apache Axis2

The Apache Axis2 project is a Java-based implementation of both the client and server sides of the Web services equation. Designed to take advantage of the lessons learned from Apache Axis 1.0, Apache Axis2 provides a complete object model and a modular architecture that makes it easy to add functionality and support for new Web services-related specifications and recommendations [8].

Among the functionalities provided by Apache Axis2 is the creation of implementation classes for both the server and client using WSDL documents. This is a significant advantage for dynamic invocation of Web Services considering that Web service clients usually are hard coded implementations that cannot be modified at runtime and are designed to deal with the invocation and data interaction of a single Web service. Therefore, Apache Axis2 is able to infer data requirements and interactions of Web services from its WSDL description documents and dynamically implement the necessary classes for its invocation at runtime.

C. JAIN SLEE Specification

JAIN SLEE aims at defining a new kind of application server designed for hosting carrier-grade Telco services. In particular, a JSLEE container is designed for hosting communication applications while typical application servers have been designed for enterprise applications and they usually do not consider high-availability and performance concerns. JSLEE containers relay on an event based model, with asynchronous interactions among components [9]. The atomic element define by JSLEE is the SBB (Service Building Block).

An SBB is a software component that sends and receives events, and performs computations based on the receipt of such events and its current state. Events are used to represent occurrences of importance that may occur at arbitrary points of time. An event may asynchronously originate from different sources such as communications protocol stacks, network elements or from application components within the SLEE. The SLEE deals with those events through elements called resource Adaptors which adapt the particular interfaces of an external resource into the interfaces and requirements of the SLEE [10].

D. TelComp 2.0 Project

The present work is framed within the Project TelComp 2.0: *Retrieval and Composition of Complex Components for the Creation of Telco 2.0 Services* [11], funded by COLCIENCIAS and developed by the Telematics Engineering Group of the University of Cauca. The TelComp 2.0 project proposes the generation of a platform aimed to support the process of creation, composition and execution of new converged services, providing developers with tools that allow them to articulate atomic services (Web/Telco) over a unified environment for defining new value-added functionalities. TelComp 2.0 execution environment is based on the JAIN SLEE specification which must support the execution of Telco services and its integration with Web services. Therefore, the present proposal represents a major contribution to this project by defining a dynamic and adaptable mechanism for integrating Web and Telco services that facilitates and allows automating the creation of converged services.

III. Related Works

There are several proposals that address the integration of Web and Telco services, defining interaction models that mainly focus on representing and exporting Telco service functionalities as Web services so they can be composed in business processes platforms such as BPEL (Business Process Execution Language) or ESB (Enterprise Service Bus).

One of these proposals is the Parlay X specification which details a set of simple web services that can be used as building blocks for telecom applications. The key design point for all of these web services was simplicity. Parlay-X combines sets of communications functions into useful but non application specific building blocks. The capabilities are restricted to those which can be performed with a single SOAP message exchange, since this simplifies use for the non-professional programmer. The Parlay-X functions are documented in a self-documenting XML interface [12]. Following this line of development, in [7] is proposed a mechanism to encapsulate Telco functionalities running on JAIN SLEE platforms as WSDL/SOAP Web services, for them to be integrated with other Web services through an ESB. However, such encapsulation enforces a synchronous behavior of Telco services and does not allow managing communication sessions during the execution of converged services. Another approach is the one presented in [13] where a set of design patterns to represent Telco services behavior are proposed. The patterns are implemented using Parlay X interfaces and composite applications are built over a BPEL engine. This work address the complexity of managing Telco service transactions on a synchronous environment but this requires the extension of the BPEL standard by adding additional parameters that consider the asynchronous nature of Telco services, significantly increasing the complexity of implementing such system.

On [14] is presented a description and analysis of different approaches to formally describe Telco service functionalities for its integration on Web based execution environments. This work presents an UML based modelling approach to describe Telco services, that considers asynchronous behavior but does not address execution or integration issues. Another type of analysis is made on [15] where the advantages and disadvantages of different programming frameworks are considered for integrating Web applications on IMS (IP Multimedia Subsystem) based telecommunications architectures. This work presents a more convenient approach from operators perspective, considering that Web applications are integrated into its telecommunications infrastructure without having to design or implement Web based platforms for composing new converged services. However, the analysis made mainly focus on the programming complexity of using languages such as Java or Ruby for developing Web applications that could be integrated into IMS environments, but does not specifies how such integration should be carried out. It is also important to highlight the work presented in [16] where a model to describe and export Telco Services as Web services is defined. Such model is based on the development and inclusion of a SOAP resource adaptor to a JAIN SLEE environment which receives invocation requests for activating Telco services on the SLEE. As pointed out before, exporting Telco services as Web services allows to integrate them into Web based composition engines but does not permit to manage asynchronous transactions from the converged services execution flow perspective. Finally, the work presented in [17] defines a converged platform for executing composed services, comprising both Web and Telco domains. In this approach, integration is achieved through a composition engine which manages service invocations and data interactions in a centralized manner. However, undesired scenarios during the execution of converged services, such as possible invocation failures are not considered. Without relying on adaptable mechanisms, problems presented at runtime may cause major failures, preventing converged services to carry out its purposes.

Related works show a strong trend towards the integration of Web and Telco services, by representing and describing Telco services as Web services through initiatives such as Parlay X, so they can be included in Web based composition engines. Our approach pretends to achieve such integration by defining an interaction mechanism to include Web services into a Telco service capable environment like JAIN SLEE, where asynchronous behavior and transaction based interactions can be managed and executed. This type of integration helps to reduce the complexity for operators to develop converged services by not having to include Web based platforms within its network infrastructures and using additional interfaces for describing its Telco capabilities. We also consider the dynamic behavior of Web service invocations by adding failure detection and adaptive capabilities to the proposed interaction mechanism.

IV. Integration Mechanism

As stated before, this proposal is framed within the project TelComp 2.0. Within this project a generalized structure for building converged services has been defined, as shown on Figure 1. The execution logic of a converged service is managed by an orchestrator SBB which communicates with a set of atomic services (Web or Telco) through firing and receiving events. These events may be provided by resource adaptors or defined by each atomic service as custom events developed for specific tasks. An entity called *Event Router*, defined by the JAIN SLEE specification, is responsible for delivering events to its corresponding destinies.

Telco services implementation reside along with converged services in the JAIN SLEE environment due to the capabilities of this specification for supporting the execution of this kind of services. On the other hand, Web services implementations are distributed across the Internet, generally belonging to 3rd party providers. Therefore, invoking Web services from within the SLEE during the execution flow of converged services requires a mechanism to generate requests and receive responses while managing its data interactions. One possible approach is to implement a JAIN SLEE module for each Web service that wants to be invoked. However this is a complex and time consuming task due to the large number of existing Web Services.

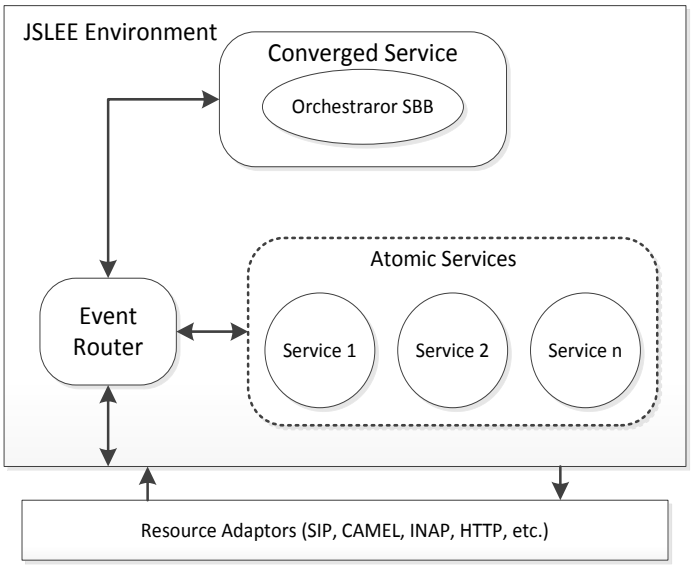


Fig. 1. Converged service general structure.

The proposed integration mechanism defines a generalized and unique component to perform Web service invocations and manage its data interactions. For this, two main modules are defined. The first module is called *Web Service Invocator* and is implemented as a JAIN SLEE application. This module is responsible for representing Web services within the SLEE and for communicating with other Telco services involved in the execution flow of converged services. The second module is called *Dynamic WS Client* which is implemented as a Java Web application, based on the Apache Axis2 framework. This module is responsible for dynamically creating Web services invocation clients from its WSDL description documents. Figure 2 presents a modular scheme of the integration mechanism.

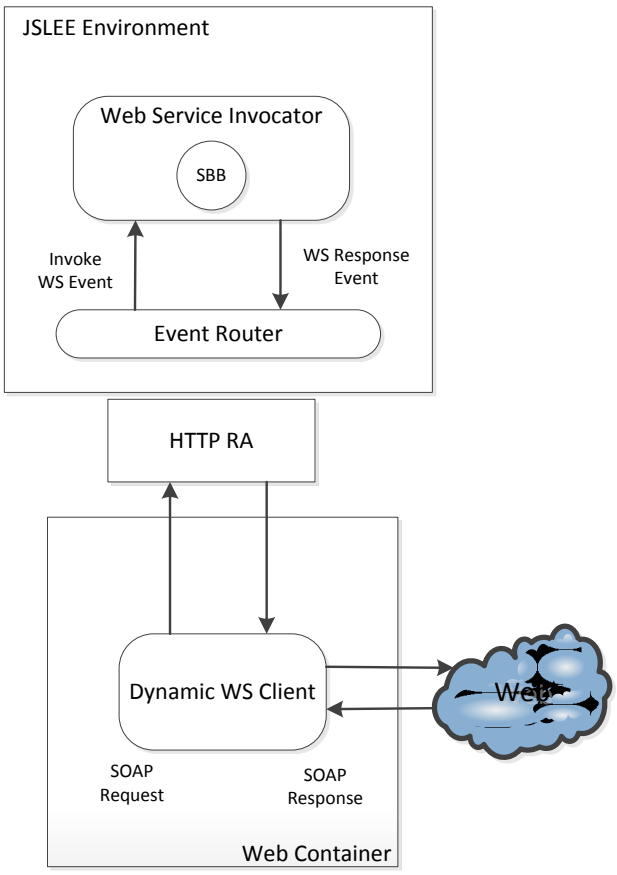


Fig. 2. Modular scheme of the integration mechasnism.

For defining a generalized module for invoking different Web services, it must be considered that despite been described in a standardized manner using WSDL documents; they present a high heterogeneity regarding its data interactions. Input and output parameters of Web services are of different types, being simple data as integers or strings, or more complex such as arrays. The number of inputs and outputs of Web services is another important factor to be considered. For example, a weather Web service may require one input (*location*) to generate a weather forecast, returning two outputs (*temperature and forecast*), while a currency convertor Web service may require three inputs (*source, target and value*) to generate the equivalency between to different currencies, returning one output (*result*). Invoking these Web services requires different methods to create the invocation requests with the corresponding input parameters and receive its responses with the corresponding output data. Below are described the *Web Service Invocation* and *Dynamic WS Client* modules and how is addressed Web service data interaction heterogeneity.

A. Web Service Invocator

As pointed out above, *Web Service Invocation* module has been implemented as a JAIN SLEE application. The interaction with this module is made through an initial event which contains the required data for a Web service invocation, and a response event which contains the invocation results, as shown on Figure 2. To deal with different number of input and output parameters with different data types, it has been defined a data managing interaction mechanism using Java HashMaps. A HashMap is a data structure used to implement associative arrays based on a Key/Value model. The initial event contains a HashMap which includes all the input parameters (name and value) of a Web service for its invocation. Same as the initial event, the response event also contains a HashMap which includes all the output parameters, associating its names and values.

Once an initial event is received by the module, it extracts all the parameters contained in the HashMap and builds a HTTP GET request addressed to the *Dynamic WS Client* module to perform the Web service invocation. Table I shows the parameters that are extracted from the HashMap and included in the HTTP request.

TABLE I. WEB SERVICE INVOCATION PARAMETERS

Parameter	Description
serviceWSDL	This parameter reference the URL of the Web service WSDL. For example: $?serviceWSDL = http://address/wheather?wsdl$
operationName	This parameter contains the name of the specific operation that wants to be invoked form the Web service. For example: $\&operationName = getForecast$
inputs	This parameter contains the name and value of the different inputs required for the Web service for its execution. For example: $\&location = Bogota$

The HTTP response returned by the *Dynamic WS Client*, which contains the output parameters resulting from invoking a Web service, presents such information using a predefined XML structure which will be described later. *Web Service Invocator* module extracts the data contained in the HTTP response through an XML parser and generates a HashMap that presents the outputs in a Name/Value manner which is included in the response event. Table II presents an example of the output HashMap, resulting from the invocation of a weather Web Service.

TABLE II. OUTPUT HASHMAP STRUCTURE EXAMPLE

Key	Value
$temperature$	$(String)"50°F - 77°F"$
$forecast$	$(String)"Cloudy\ with\ 60\%\ chance\ of\ light\ rain"$

HashMaps represent a very convenient mechanism to present the output parameters of a Web service due to their capacity of relating different type of data such as primitive variables, or arrays.

B. Dynamic WS Client

This module receives the HTTP GET request coming from the *Web Service Invocator* module which contains all the parameters needed to invoke a Web service. Using the Apache Axis2 functions, it retrieves the WSDL of the Web service and

maps the input parameter values to create an appropriate client that generates SOAP requests to the Web service. Upon receiving a SOAP response, it takes the output parameter values and generates an XML document which organizes the information. Figure 3 presents an example of a Web service response structured through an XML document.

```
<?xml version="1.0" encoding="UTF-8"?>
<outputs>
    <output>
        <name>temperature</name>
        <value>50°F-77°F</value>
    </output>
    <output>
        <name>forecast</name>
        <value>Cloudy with 60% chance of light rain</valu
    </output>
</outputs>
```

Fig. 3. Example of response structured as an XML document.

The response XML documents are defined by a structure which contains a main tag called *outputs* that encapsulate all the output parameters of a Web service response. A specific parameter is represented by the tag *output* which contains the *name* of the parameter and its corresponding *value*. To represent parameters as arrays, its content is separated in different *output* structures which have the same *name* but possibly a different *value*. Having *output* structures with the same *name* allows the *Web Service Invocator* module to identify arrays and build them to be included in the response HashMap, preventing them to be considered as independent parameters. Once the XML document including all the output parameter has been created, the *Dynamic WS Client* module generates an HTTP 200 OK response including the XML document and addressed to the *Web Service Invocator Module*.

V. Fault Handling and Adaptation

The fault handling and adaptation functions added to the integration mechanism are aimed to detect possible failures during the invocation of Web services and to adapt the mechanism for invocating backup services which are defined at creation time, allowing converged services to carry out its purpose. Based on the service fault taxonomy defined in [18], the fault handling mechanism identifies two different types of faults:

- *Service Provider Faults:* reference possible faults that occur on platforms hosting Web services implementation. For example a Web service being unavailable won't be able to process and respond invocation requests.
- *Communication Faults:* reference possible faults that may occur due to network issues. For example sending to a Web service an invocation request over a network holding heavy traffic may result on a delayed response or no response at all.

The detection of *Service Provider Faults* is made by capturing exceptions generated in the *Dynamic WS Client* module during the invocation of Web services. Once an exception is captured, it sends a HTTP error response to the *Web Service Invocator* module. On the other hand, *Communication Faults* are detected by establishing an execution timer for every invocation request sent by the *Web Service Invocator* module. The maximum waiting time for a Web service response is configured by the converged service developer at the creation stage. The timer is set using the Timer Facility defined by the JAIN SLEE specification. Once a fault is detected, the *Web Service Invocator* module adapts itself to perform the invocation of backup Web services defined at creation time through a list provided by the converged service developer. The adaptation process is formally defined in algorithm 1.

Algorithm 1: Adaptation Process for Web Service Invocation

INPUTS: $WSInvocationData, backupServiceList[\,], timeLimit$

OUTPUTS: $WSOutputData, responseEvent$

BEGIN

$WSOutputData := \emptyset$

$httpReponse := \emptyset$

$httpRequest = createReq(WSInvocationData)$

$timer = setTimer(timeLimit)$

$fireHTTPRequest(httpRequest)$

while $timer$ is active **do**

if $httpResponse \neq \emptyset$ **then**

$deactivateTimer(timer)$

end if

end while

if $httpResponse \neq \emptyset$ **then**

if $httpResponse = 200\ OK$ **then**

$WSOutputData = parseData(httpResponse)$

else do

for ws_i in $backupServiceList[\,]$ **do**

$WSOutputData = invokeWS(ws_i)$

if $WSOutputData \neq \emptyset$ **then**

break

end if

end for

end if

else do

for ws_i in $backupServiceList[\,]$ **do**

$WSOutputData = invokeWS(ws_i)$

if $WSOutputData \neq \emptyset$ **then**

break

end if

end for

end if

$responseEvent = createEvent(WSOutputData)$

$fireEvent(responseEvent)$

END

As stated before, *Web Service Invocator* module initiates a Web service invocation by creating an HTTP request which includes the invocation parameters ($WSInvocationData$). Immediately after sending the request to the *Dynamic WS Client* module, it sets a timer which indicates the maximum waiting time ($timeLimit$) for a response. While waiting, if a response is received the timer is deactivated and the response is analyzed. If the response corresponds to an HTTP 200 OK response, the output parameters are retrieved ($WSOutputData$) and the response event is fired. Otherwise, either if the response did not arrive within the established time or if it corresponds to an HTTP error response, backup Web services (ws_i) retrieved from the list ($backupServiceList[\,]$) provided by the converged service developer are invoked until a successful invocation is achieved. However if no successful invocation is completed an error in the execution flow of the converged service will be produced. This adaptation mechanism helps to reduce the probability of major faults occurring during the execution of converged services.

VI. Implementation and Case Study

The implementation of the *Web Service Invocation* and *Dynamic Web Service* modules was made over the Mobicents JAIN SLEE server and Apache Tomcat Web container respectively. A functional test of the integration mechanism was performed through a converged service called *Twitter Financial Message* which is composed of two Web services (Finance WS and Twitter WS) and two Telco services (Receive IM and Send IM). This service initiates its execution upon receiving a SIP instant message (Receive IM) from the user containing the code name of a NASDAQ stock from which the user wants to know its current value. With the code name is invoked a Web service (Finance WS) that returns the stock current value. Then, the stock information is sent simultaneously to the user as an instant message (Send IM) and as a private message in the Twitter account of the user through a Web service (Twitter WS). Figure 4 shows the converged service diagram and Figure 5 presents the implementation modules needed for its execution.

Through the modules developed, invoking and integrating Web services into a Telco environment requires only for converged service developers to specify data flow between component services, leaving Web service clients' implementation details to the modules.

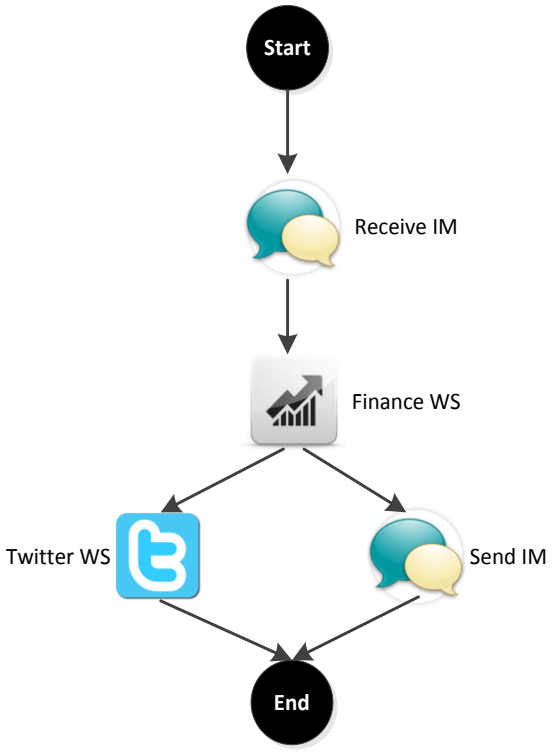


Fig. 4. Twitter Financial Message diagram.

To test the adaptation functions of the integration mechanism was set an undesired scenario where Twitter WS was made unavailable. As a backup Web service, another instance of Twitter WS was configured with a different name.

The test results show that the mechanism adapts to invoke a backup Web service in an average time of 41.9 milliseconds after detecting a fault, indicating a high performance of the adaptation process as seen in Figure 6.

VII. Conclusions and Future Work

This paper presented a dynamic and adaptable approach for integrating Web and Telco services in JAIN SLEE environments. This approach defines a generalized set of modules which allow invoking Web services regardless of their functionality or required data structure. The proposed mechanism facilitates for developers to invoke Web services and integrate them into the execution flow of converged services, by only specifying the data flow required by each one without having to deal with the implementation details of Web service clients. An adaptation mechanism that detects faults occurred during the invocation of Web services is also implemented with very satisfying results. The adaptation process is carried out in a transparent manner from the user perspective, adapting to invoke backup Web services defined at the creation stage, upon the detection of failures with very low execution times. Such process helps reducing the probability of major faults during the execution of converged services.

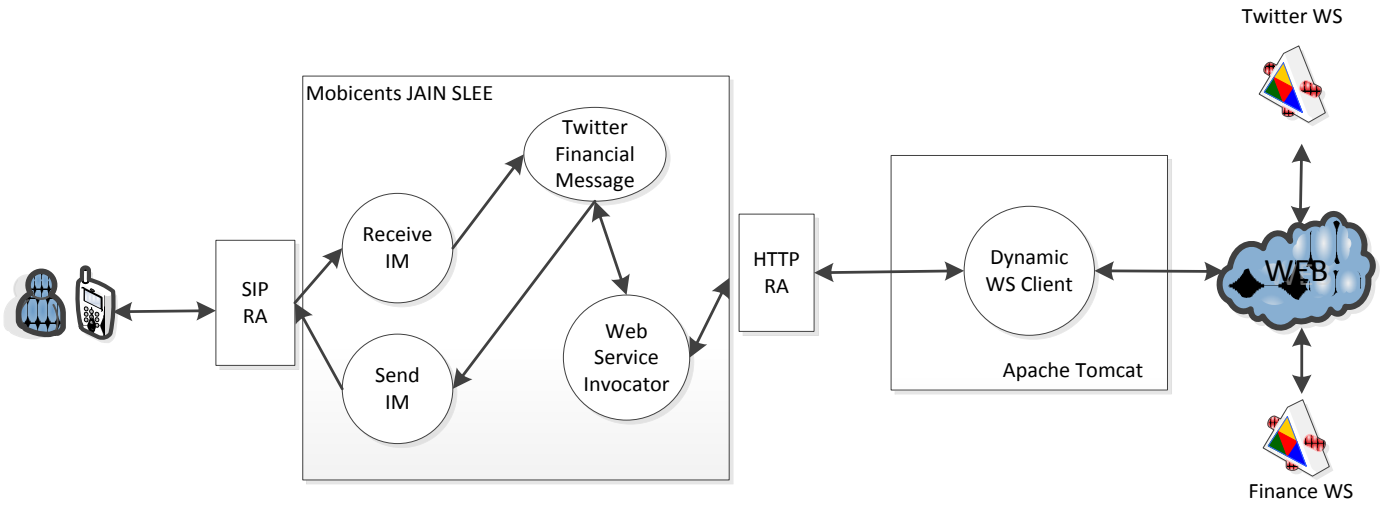


Fig. 5. Implementation of Twitter Financial Message.

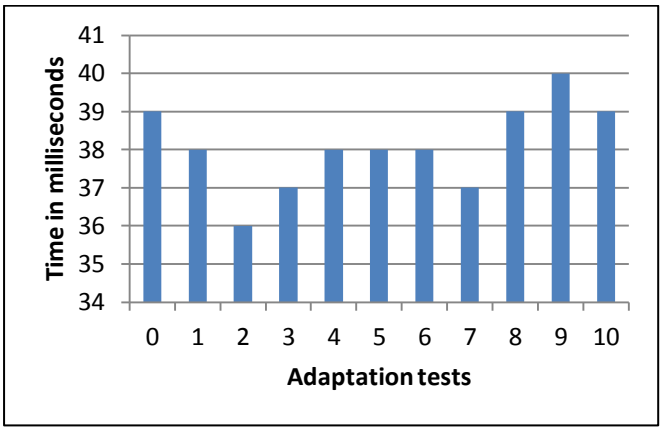


Fig. 6. Performance Adaptation Tests.

This work comprises an important contribution to the TelComp 2.0 project, specifically on the automation of converged service creation and composition. As future work, is proposed to automate the process of retrieving backup services based on functional and semantic analysis, without needing to be specified by the developer at creation time. Other approach that could be addressed is to include REST services into the integration mechanism, considering the large proliferation of this type of services. Another future approach is the design and development of a monitoring mechanism which enables to be aware of the current status of all services and detect undesired behaviors that may compromise the proper operation of converged services.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank University of Cauca and TelComp2.0 project (Code: 1103-521-28338 CT458-2011) for supporting and financing this work and the MSc. students Julián Andrés Rojas and Leandro Ordoñez.

REFERENCES

[1] Jong-Lok Yoon, "Telco 2.0: a new role and business model," *Commun. Mag. IEEE*, vol. 45, no. 1, pp. 10–12, 2007.

[2] JSRs: Java Specification Requests, "JSR 289: SIP Servlet v1.1." 2008.

[3] Ericsson, "Ericsson Converged Service Studio." 2013.

[4] uReach Technologies, "Converged Services Framework." 2013.

[5] JSRs: Java Specification Requests, "JSR 240: JAIN SLEE (JSLEE) v1.1." 2008.

[6] D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard, "Web Services Architecture, W3C Working Group Note 11." W3C Technical Reports and Publications, 2004.

[7] K. Rezabeigi, A. Vafei, and N. Movahhedinia, "A Web Services based Architecture for NGN Services Delivery," *World Acad. Sci. Egineering Technol.*, vol. 43, 2008.

[8] Apache Software Foundation, "Apache Axis2 User's Guide." 2012.

[9] P. Falcarin and C. Venezia, "Communication Web Services and JAIN-SLEE Integration Challenges," *Int. J. Web Serv. Res.*, vol. 5, no. 4, p. 5978, 2008.

[10] P. Falcarin and L. Walter, "An Aspect-Oriented Approach for Dynamic Monitoring of a Service Logic Excecution Environment," *IEC Annu. Rev. Commun.*, vol. 59, pp. 237–242, 2006.

[11] Grupo de Ingeniería Telemática, "TelComp2.0 Project Website," 2013. [Online]. Available: http://190.90.112.7:8080/TelComp-SCE/.

[12] L. Zygmunt, "Parlay/OSA - a New Way to Create Wireless Services." The Parlay Group, 2003.

[13] P. Baglietto, M. Maresca, M. Stecca, A. Manzalini, R. Minerva, and C. Moiso, "Analysis of design patterns for composite telco services," in *Intelligence in Next Generation Networks (ICIN), 2010 14th International Conference on*, 2010, pp. 1–6.

[14] E. Bertin and N. Crespi, "Describing Next Generation Communication Services: A Usage Perspective," *Lect. Notes Comput. Sci.*, vol. 5377, pp. 86–97, 2008.

[15] A. Hasanović, N. Suljanović, A. Mujčić, and R. Sernec, "Dynamic Languages Integration Path for Telecom Applications," in *Digital Telecommunications, 2009. ICDT '09. Fourth International Conference on*, 2009, pp. 133–137.

[16] C. Venezia and P. Falcarin, "Communication Web Services Composition and Integration," in *Web Services, 2006. ICWS '06. International Conference on*, 2006, pp. 523–530.

[17] J. Niemoller, E. Freiter, K. Vandikas, R. Quinet, R. Levenshteyn, and I. Fikouras, "Composition in Converged Service Networks: Requirements and Solutions," in *International Workshop on Business Systems Management and Engineering*, 2010.

[18] I.-Y. Chen, G.-K. Ni, C.-H. Kuo, and C.-Y. Lin, "A BPEL-Based Fault-Handling Architecture for Telecom Operation Support Systems.," *JACIII*, vol. 14, no. 5, pp. 523–530, 2010.

Improving Forecasting Accuracy in the Case of Intermittent Demand Forecasting

Daisuke Takeyasu[1]
[1]The Open University of Japan, Chiba City, Japan

Asami Shitara[2]
[2]Tax Corporation Arknet, Shizuoka City, Japan

Kazuhiro Takeyasu[3]
[3]College of Business Administration, Tokoha University, Shizuoka, Japan

***Abstract*—In making forecasting, there are many kinds of data. Stationary time series data are relatively easy to make forecasting but random data are very difficult in its execution for forecasting. Intermittent data are often seen in industries. But it is rather difficult to make forecasting in general. In recent years, the needs for intermittent demand forecasting are increasing because of the constraints of strict Supply Chain Management. How to improve the forecasting accuracy is an important issue. There are many researches made on this. But there are rooms for improvement. In this paper, a new method for cumulative forecasting method is proposed. The data is cumulated and to this cumulated time series, the following method is applied to improve the forecasting accuracy. Trend removing by the combination of linear and 2nd order non-linear function and 3rd order non-linear function is executed to the production data of X-ray image intensifier tube device and Diagnostic X-ray image processing apparatus. The forecasting result is compared with those of the non-cumulative forecasting method. The new method shows that it is useful for the forecasting of intermittent demand data. The effectiveness of this method should be examined in various cases.**

Keywords—intermittent demand forecasting; minimum variance; exponential smoothing method; trend

I. INTRODUCTION

Supply chain management is inevitable in industries in recent years. Demand forecasting is the basis in supply chain management. In industries, how to improve forecasting accuracy such as sales, shipping is an important issue. There are cases that intermittent demand forecasting is required. But the mere application of the past method does not bear good estimation of parameters and exquisite forecasting.

There are many researchers made on this.

Based upon the Croston's model (Box et al.2008), Shenstone and Hyndma (2005) analyzed the intermittent demand forecasting. Troung et al. (2011) applied Neural Network to intermittent demand forecasting. Tanaka et al. (2012) has built sales forecasting model for book publishing, where they have devised cumulative forecasting method.

In this paper, we further develop this cumulative forecasting method in order to improve the forecasting accuracy for intermittent demand.

A new method for cumulative forecasting method is proposed.

The data is cumulated and to this cumulated time series, the following method is applied to improve the forecasting accuracy. Focusing that the equation of exponential smoothing method(ESM) is equivalent to (1,1) order ARMA model equation, a new method of estimation of smoothing constant in exponential smoothing method is proposed before by us which satisfies minimum variance of forecasting error[7]. Generally, smoothing constant is selected arbitrarily. But in this paper, we utilize above stated theoretical solution. Firstly, we make estimation of ARMA model parameter and then estimate smoothing constants.

Thus theoretical solution is derived in a simple way and it may be utilized in various fields. Furthermore, combining the trend removing method with this method, we aim to improve the forecasting accuracy. An approach to this method is executed in the following method. Trend removing by the combination of linear and 2nd order non-linear function and 3rd order non-linear function is executed to the data of X-ray image intensifier tube device and Diagnostic X-ray image processing apparatus. The weights for these functions are set 0.5 for two patterns at first and then varied by 0.01 increment for three patterns and optimal weights are searched. For the comparison, monthly trend is removed after that. Theoretical solution of smoothing constant of ESM is calculated for both of the monthly trend removing data and the non-monthly trend removing data. Then forecasting is executed on these data.

The forecasting result is compared with those of the non-cumulative forecasting method. The new method shows that it is useful for the forecasting of intermittent demand data. The effectiveness of this method should be examined in various cases.

The rest of the paper is organized as follows. In section 2, the new method is described. ESM is stated by ARMA model and estimation method of smoothing constant is derived using ARMA model identification. The combination of linear and non-linear function is introduced for trend removing and the Monthly Ratio is also referred. Forecasting is executed in section 3, and estimation accuracy is examined, which is followed by the Discussion of section 4

II. DESCRIPTION OF THE NEW METHOD

A. Description of ESM Using ARMA Model[5]

In ESM, forecasting at time $t+1$ is stated in the following equation.

$$\begin{aligned}\hat{x}_{t+1} &= \hat{x}_t + \alpha(x_t - \hat{x}_t) \\ &= \alpha x_t + (1-\alpha)\hat{x}_t\end{aligned} \tag{1}$$

Here,

$\hat{x}_{t+1}$: forecasting at $t+1$
x_t : realized value at t
α : smoothing constant $(0<\alpha<1)$

(1) is re-stated as

$$\hat{x}_{t+1} = \sum_{l=0}^{\infty} \alpha(1-\alpha)^l x_{t-l} \tag{2}$$

By the way, we consider the following (1,1) order ARMA model.

$$x_t - x_{t-1} = e_t - \beta e_{t-1} \tag{3}$$

Generally, (p,q) order ARMA model is stated as

$$x_t + \sum_{i=1}^{p} a_i x_{t-i} = e_t + \sum_{j=1}^{q} b_j e_{t-j} \tag{4}$$

Here,

$\{x_t\}$: Sample process of Stationary Ergodic Gaussian Process $x(t)$ $t = 1,2,\cdots,N,\cdots$

$\{e_t\}$: Gaussian White Noise with 0 mean σ_e^2 variance

MA process in (4) is supposed to satisfy convertibility condition. Utilizing the relation that

$$E[e_t | e_{t-1}, e_{t-2}, \cdots] = 0$$

we get the following equation from (3).

$$\hat{x}_t = x_{t-1} - \beta e_{t-1} \tag{5}$$

Operating this scheme on $t+1$, we finally get

$$\begin{aligned}\hat{x}_{t+1} &= \hat{x}_t + (1-\beta)e_t \\ &= \hat{x}_t + (1-\beta)(x_t - \hat{x}_t)\end{aligned} \tag{6}$$

If we set $1-\beta=\alpha$, the above equation is the same with (1), i.e., equation of ESM is equivalent to (1,1) order ARMA model, or is said to be (0,1,1) order ARIMA model because 1st order AR parameter is -1 [1].

Comparing with (3) and (4), we obtain

$$\begin{cases} a_1 = -1 \\ b_1 = -\beta \end{cases}$$

From (1), (6),

$$\alpha = 1-\beta$$

Therefore, we get

$$\begin{cases} a_1 = -1 \\ b_1 = -\beta = \alpha - 1 \end{cases} \tag{7}$$

From above, we can get estimation of smoothing constant after we identify the parameter of MA part of ARMA model. But, generally MA part of ARMA model become non-linear equations which are described below. Let (4) be

$$\tilde{x}_t = x_t + \sum_{i=1}^{p} a_i x_{t-i} \tag{8}$$

$$\tilde{x}_t = e_t + \sum_{j=1}^{q} b_j e_{t-j} \tag{9}$$

We express the autocorrelation function of $\tilde{x}_t$ as $\tilde{r}_k$ and from (8), (9), we get the following non-linear equations which are well known.

$$\left\{ \begin{array}{ll} \tilde{r}_k = \sigma_e^2 \sum_{j=0}^{q-k} b_j b_{k+j} & (k \le q) \\ 0 & (k \ge q+1) \\ \tilde{r}_0 = \sigma_e^2 \sum_{j=0}^{q} b_j^2 & \end{array} \right\} \tag{10}$$

For these equations, a recursive algorithm has been developed. In this paper, parameter to be estimated is only b_1, so it can be solved in the following way.

From (3) (4) (7) (10), we get

$$\left. \begin{array}{l} q = 1 \\ a_1 = -1 \\ b_1 = -\beta = \alpha - 1 \\ \tilde{r}_0 = (1+b_1^2)\sigma_e^2 \\ \tilde{r}_1 = b_1 \sigma_e^2 \end{array} \right\} \tag{11}$$

If we set

$$\rho_k = \frac{\tilde{r}_k}{\tilde{r}_0} \tag{12}$$

the following equation is derived.

$$\rho_1 = \frac{b_1}{1+b_1^2} \tag{13}$$

We can get b_1 as follows.

$$b_1 = \frac{1 \pm \sqrt{1-4\rho_1^2}}{2\rho_1} \tag{14}$$

In order to have real roots, ρ_1 must satisfy

$$|\rho_1| \le \frac{1}{2} \tag{15}$$

From invertibility condition, b_1 must satisfy

$$|b_1| < 1$$

From (13), using the next relation,

$$(1-b_1)^2 \ge 0$$
$$(1+b_1)^2 \ge 0$$

(15) always holds. As

$$\alpha = b_1 + 1$$

b_1 is within the range of

$$-1 < b_1 < 0$$

Finally we get

$$\left.\begin{aligned} b_1 &= \frac{1-\sqrt{1-4\rho_1^2}}{2\rho_1} \\ \alpha &= \frac{1+2\rho_1-\sqrt{1-4\rho_1^2}}{2\rho_1} \end{aligned}\right\} \tag{16}$$

which satisfy above condition. Thus we can obtain a theoretical solution by a simple way.

Here ρ_1 must satisfy

$$-\frac{1}{2} < \rho_1 < 0 \tag{17}$$

in order to satisfy $0 < \alpha < 1$.

Focusing on the idea that the equation of ESM is equivalent to (1,1) order ARMA model equation, we can estimate smoothing constant after estimating ARMA model parameter. It can be estimated only by calculating 0th and 1st order autocorrelation function.

B. Trend Removal Method[5]

As trend removal method, we describe the combination of linear and non-linear function.

[1] Linear function

We set

$$y = a_1 x + b_1 \tag{18}$$

as a linear function.

[2] Non-linear function

We set

$$y = a_2 x^2 + b_2 x + c_2 \tag{19}$$

$$y = a_3 x^3 + b_3 x^2 + c_3 x + d_3 \tag{20}$$

as a 2nd and a 3rd order non-linear function.

[3] The combination of linear and non-linear function

We set

$$y = \alpha_1 (a_1 x + b_1) + \alpha_2 (a_2 x^2 + b_2 x + c_2) \tag{21}$$

$$y = \beta_1 (a_1 x + b_1) + \beta_2 (a_3 x^3 + b_3 x^2 + c_3 x + d_3) \tag{22}$$

$$\begin{aligned} y = {} & \gamma_1 (a_1 x + b_1) + \gamma_2 (a_2 x^2 + b_2 x + c_2) \\ & + \gamma_3 (a_3 x^3 + b_3 x^2 + c_3 x + d_3) \end{aligned} \tag{23}$$

as the combination of linear and 2nd order non-linear and 3rd order non-linear function. Here, $\alpha_2 = 1-\alpha_1$, $\beta_2 = 1-\beta_1$, $\gamma_3 = 1-(\gamma_1+\gamma_2)$. Comparative discussion concerning (21), (22) and (23) are described in section 5.

C. Monthly Ratio[5]

For example, if there is the monthly data of L years as stated bellow:

$$\{x_{ij}\}\ (i = 1, \cdots, L)\ (j = 1, \cdots, 12)$$

Where, $x_{ij} \in R$ in which j means month and i means year and x_{ij} is a shipping data of i-th year, j-th month. Then, monthly ratio $\tilde{x}_j\ (j = 1, \cdots, 12)$ is calculated as follows.

$$\tilde{x}_j = \frac{\frac{1}{L}\sum_{i=1}^{L} x_{ij}}{\frac{1}{L} \cdot \frac{1}{12} \sum_{i=1}^{L} \sum_{j=1}^{12} x_{ij}} \tag{24}$$

III. Forecasting The Production Data

A. Analysis Procedure

Sum total data of production data of X-ray image intensifier tube device and Diagnostic X-ray image processing apparatus from January 2010 to December 2012 are analyzed. These data are obtained from the Annual Report of Statistical Investigation on Statistical-Survey-on-Trends-in-Pharmaceutical-Production by Ministry of Health, Labour and Welfare in Japan.

The original data are accumulated for X-ray image intensifier tube device data and Diagnostic X-ray image processing apparatus data.

Analysis procedure is as follows. There are 36 monthly data for each case. We use 24 data (1 to 24) and remove trend by the method stated in 2.2. Then we calculate monthly ratio

by the method stated in 2.3. After removing monthly trend, the method stated in 2 is applied and Exponential Smoothing Constant with minimum variance of forecasting error is estimated. Then 1 step forecast is executed. Thus, data is shifted to 2nd to 25th and the forecast for 26th data is executed consecutively, which finally reaches forecast of 36th data. To examine the accuracy of forecasting, variance of forecasting error is calculated for the data of 25th to 36th data. Final forecasting data is obtained by multiplying monthly ratio and trend. Forecasting error is expressed as:

$$\varepsilon_i = \hat{x}_i - x_i \tag{25}$$

$$\bar{\varepsilon} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i \tag{26}$$

Variance of forecasting error is calculated by:

$$\sigma_\varepsilon^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\varepsilon_i - \bar{\varepsilon}\right)^2 \tag{27}$$

B. Trend Removing

Trend is removed by dividing original data by,(21),(22),(23). The patterns of trend removal are exhibited in Table 1.

TABLE I. THE PATTERNS OF TREND REMOVAL

Pattern1	α_1, α_2 are set 0.5 in the equation (21)
Pattern2	β_1, β_2 are set 0.5 in the equation (22)
Pattern3	α_1 is shifted by 0.01 increment in (21)
Pattern4	β_1 is shifted by 0.01 increment in (22)
Pattern5	γ_1 and γ_2 are shifted by 0.01 increment in (23)

In pattern1 and 2, the weight of α_1, α_2, β_1, β_2 are set 0.5 in the equation (21),(22). In pattern3, the weight of α_1 is shifted by 0.01 increment in (21) which satisfy the range $0 \le \alpha_1 \le 1.00$. In pattern4, the weight of β_1 is shifted in the same way which satisfy the range $0 \le \beta_1 \le 1.00$. In pattern5, the weight of γ_1 and γ_2 are shifted by 0.01 increment in (23) which satisfy the range $0 \le \gamma_1 \le 1.00$, $0 \le \gamma_2 \le 1.00$ .The best solution is selected which minimizes the variance of forecasting error.

C. Removing trend of monthly ratio

After removing trend, monthly ratio is calculated by the method stated in 2.3.

D. Estimation of Smoothing Constant with Minimum Variance of Forecasting Error

After removing monthly trend, Smoothing Constant with minimum variance of forecasting error is estimated utilizing (16). There are cases that we cannot obtain a theoretical solution because they do not satisfy the condition of (15).

In those cases, Smoothing Constant with minimum variance of forecasting error is derived by shifting variable from 0.01 to 0.99 with 0.01 interval.

The intermittent demand data often include 0 data. If there are so many 0 data, there is a case we cannot calculate the theoretical solation of smoothing constant.

In that case, we add very tiny data which is not 0 but close to 0 that does not affect anything in calculating parameters (i.e. negligible small).

E. Forecasting AND Variance of Forecasting Error

Utilizing smoothing constant estimated in the previous section, forecasting is executed for the data of 25th to 36th

data. Final forecasting data is obtained by multiplying monthly ratio and trend. Variance of forecasting error is calculated by (27).

As we have made accumulated data case and tiny data close to 0 added case, we have the following cases altogether.

1. Non Monthly Trend Removal

(1) Accumulated Data

(2) Non Accumulated Data

(2-1) Forecasting from the Accumulated data (Accumulated forecasting data at time *n*－Accumulated data (at time *n-1*))

A. Pattern1 B. Pattern2 C. Pattern3 D. Pattern4 E. Pattern5

(2-2) Forecasting from the tiny data close to 0 added case

A. Pattern1 B. Pattern2 C. Pattern3 D. Pattern4 E. Pattern5

2. Monthly Trend Removal

(1) Accumulated Data

(2) Non Accumulated Data

(2-1) Forecasting from the Accumulated data (Accumulated forecasting data at time *n*－Accumulated data (at time *n-1*))

A. Pattern1 B. Pattern2 C. Pattern3 D. Pattern4 E. Pattern5

(2-2) Forecasting from the tiny data close to 0 added case

A. Pattern1 B. Pattern2 C. Pattern3 D. Pattern4 E. Pattern5

We can make forecasting by reversely making the data from the forecasting accumulated data, i.e., that is shown at (2-1).

Now, we show them at Figure1 through 6.

Figure 1,2 and 3 show the Non-monthly Trend Removal Case in X-ray image intensifier tube device.

It includes all cases classified above.

Figure 1 shows the Accumulated Data Case in Non-Monthly Trend Removal.

Figure 2 shows the Forecasting from the Accumulated Data Case in Non-Monthly Trend Removal.

Figure 3 shows the Forecasting from the tiny data close to 0 added case in Non-Monthly Trend Removal.

Table 2,3 and 4 show the corresponding variance of forecasting error for each Figure 1,2 and 3.

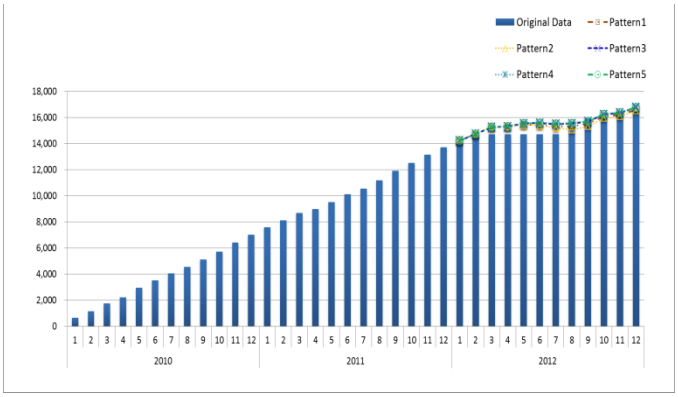


Fig. 1. Forecasting from the Accumulated Data Case in Non-Monthly Trend Removal (1-(1))

TABLE II. VARIANCE OF FORECASTING ERROR (1-(1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
582758.5794	553598.6015	655627.9015	672911.8873	655627.9015

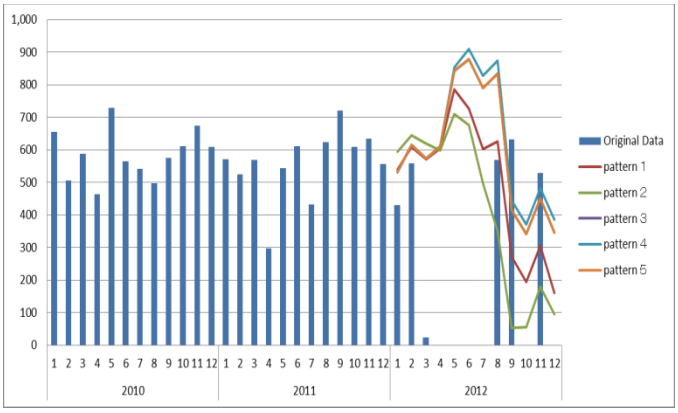


Fig. 2. Forecasting from the Accumulated Data Case in Non-monthly Trend Removal (1-(2-1))

TABLE III. VARIANCE OF FORECASTING ERROR (1-(2-1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
140940.069	183864.2572	134016.6212	134302.4864	134016.6212

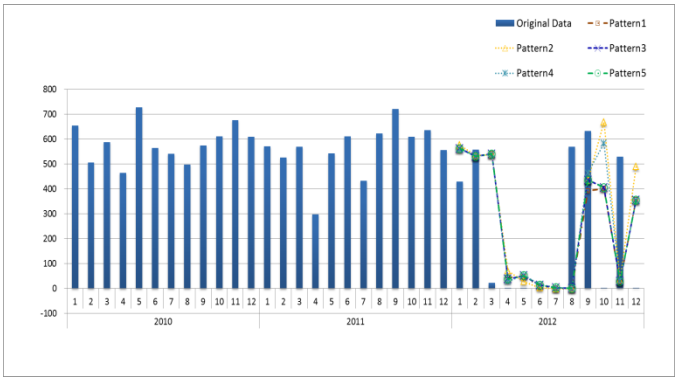


Fig. 3. Forecasting from the Tiny Data close to 0 Added case in Non-Monthly Trend Removal (1-(2-2))

TABLE IV. VARIANCE OF FORECASTING ERROR (1-(2-2))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
110237.5688	141410.2624	108230.8417	121685.9633	108230.8417

Next, we see the Monthly Trend Removal case.

Figure 4,5 and 6 show the Monthly Trend Removal Case in X-ray image intensifier tube device.

It includes all cases classified above.

Figure 4 shows the Accumulated Data Case in Monthly Trend Removal.

Figure 5 shows the Forecasting from the Accumulated Data Case in Monthly Trend Removal.

Figure 6 shows the Forecasting from the tiny data close to 0 added case in Monthly Trend Removal.

Table 5,6 and 7 show the corresponding variance of forecasting error for each Figure 4,5 and 6.

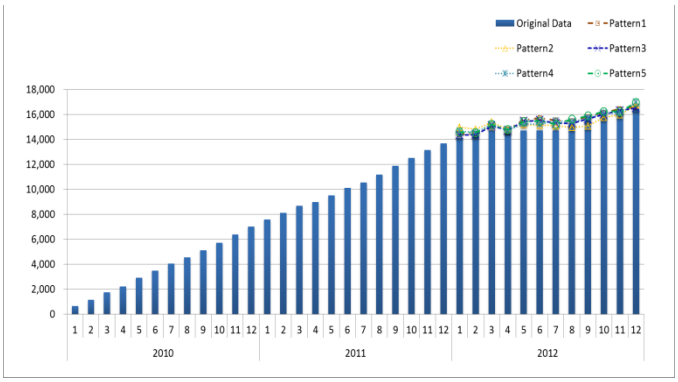


Fig. 4. Accumulated Data case in Monthly Trend Removal (2-(1))

TABLE V. VARIANCE OF FORECASTING ERROR (2-(1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
657839.2024	505482.4188	621118.7598	636984.8635	655587.073

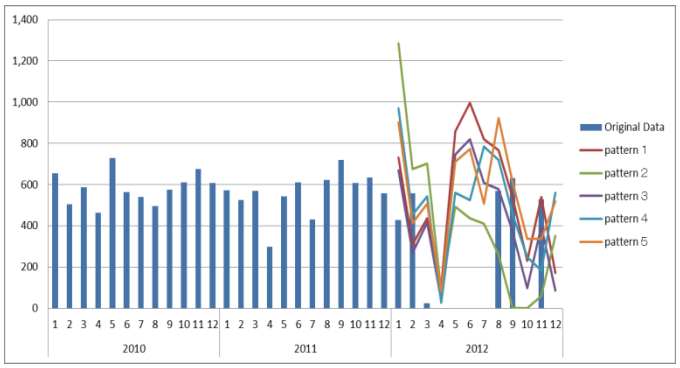


Fig. 5. Forecasting from the accumulated Data case in Monthly Trend Removal (2-(2-1))

TABLE VI. VARIANCE OF FORECASTING ERROR (2-(2-1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
153367.9683	210960.3255	139068.3784	129896.7705	103268.0731

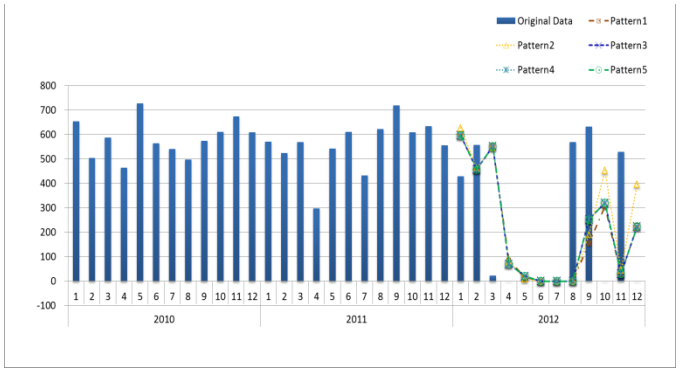


Fig. 6. Forecasting from the Tiny Data close to 0 Added case in Monthly Trend Removal (2-(2-2))

TABLE VII. VARIANCE OF FORECASTING ERROR (2-(2-2))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
115474.9994	131347.7067	108078.5346	108078.5346	108078.5346

Table 8 shows the summary for X-ray image intensifier tube devic by the Variance of forecasting error.

TABLE VIII. SUMMARY FOR X-RAY IMAGE INTENSIFIER TUBE DEVICE

		Monthly Trend Removal			Non Monthly Trend Removal		
Name	X-ray image intensifier tube device	Accumulated Data	Forecasting Value－Accumulated Value	Tiny data close to 0 added case	Accumulated Data	Forecasting Value－Accumulated Value	Tiny data close to 0 added case
Minimum variance of Forecasting Error		505482.4188	103268.0731	108078.5346	553598.6015	134016.6212	108230.8417

Now, we proceed to the case of Diagnostic X-ray image processing apparatus. Figure 7,8 and 9 show the Non-monthly Trend Removal Case in Diagnostic X-ray image processing apparatus.

It includes all cases classified above.

Figure 7 shows the Accumulated Data Case in Non-Monthly Trend Removal.

Figure 8 shows the Forecasting from the Accumulated Data Case in Non-Monthly Trend Removal.

Figure 9 shows the Forecasting from the tiny data close to 0 added case in Non-Monthly Trend Removal.

Table 9,10 and 11 show the corresponding variance of forecasting error for each Figure 7,8 and 9.

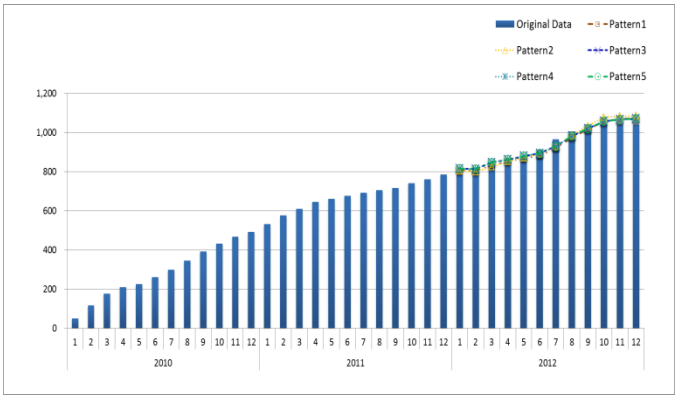


Fig. 7. Accumulated Data case in Non-Monthly Trend Removal (1-(1))

TABLE IX. VARIANCE OF FORECASTING ERROR (1-(1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
12144.26114	13893.64686	11033.83677	11033.83677	11033.83677

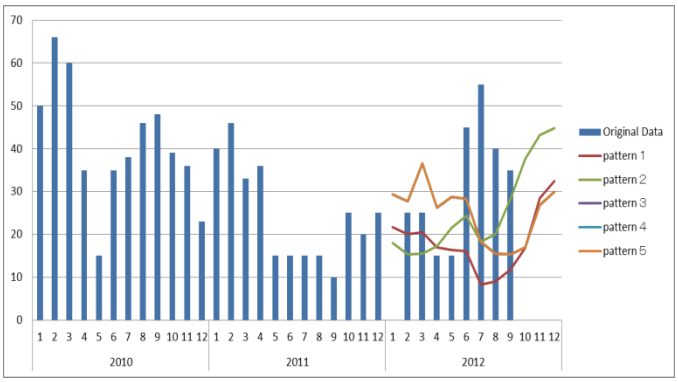


Fig. 8. Forecasting from the accumulated Data case in Non-Monthly Trend Removal (1-(2-1))

TABLE X. VARIANCE OF FORECASTING ERROR (1-(2-1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
642.6679766	714.8801978	515.2469097	515.2469097	515.2469097

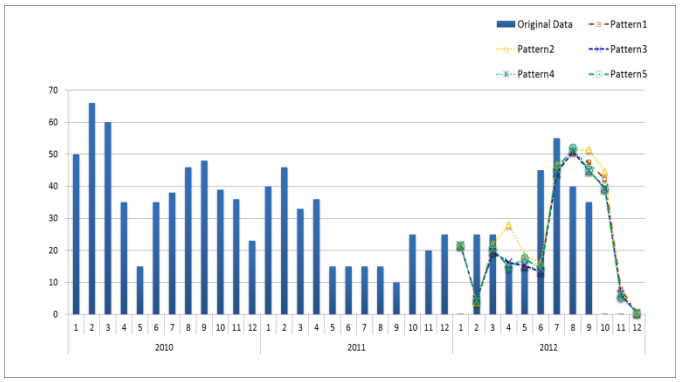


Fig. 9. Forecasting from the Tiny Data close to 0 Added case in Non-Monthly Trend Removal (1-(2-2))

TABLE XI. VARIANCE OF FORECASTING ERROR (1-(2-2))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
365.3419523	381.0283782	345.9402412	336.4640944	336.4640944

Next, we see the Monthly Trend Removal case.

Figure 10,11 and 12 show the Monthly Trend Removal Case in Diagnostic X-ray image processing apparatus..

It includes all cases classified above.

Figure 10 shows the Accumulated Data Case in Monthly Trend Removal.

Figure 11 shows the Forecasting from the Accumulated Data Case in Monthly Trend Removal.

Figure 12 shows the Forecasting from the tiny data close to 0 added case in Monthly Trend Removal.

Table 12,13 and 14 show the corresponding variance of forecasting error for each Figure 10,11 and 12.

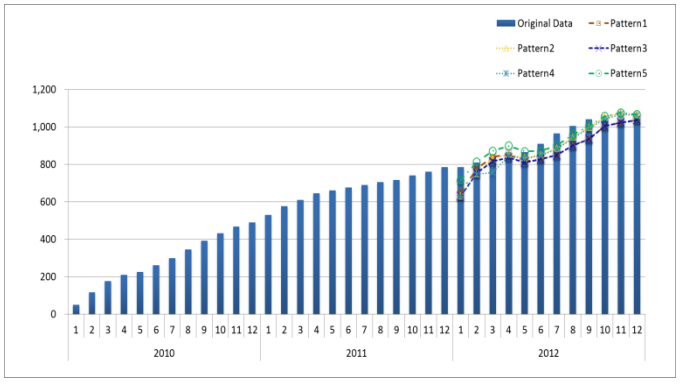


Fig. 10. Accumulated Data case in Monthly Trend Removal (2-(1))

TABLE XII. VARIANCE OF FORECASTING ERROR (2-(1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
13726.72311	19516.74892	15344.88764	20527.68403	13453.32873

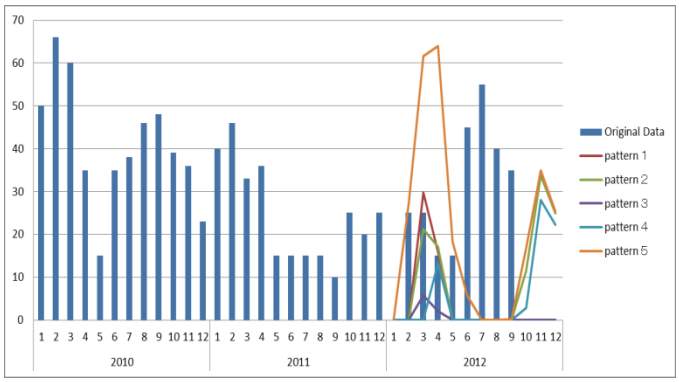


Fig. 11. Forecasting from the accumulated Data case in Monthly Trend Removal (2-(2-1))

TABLE XIII. VARIANCE OF FORECASTING ERROR (2-(2-1))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
463.5294454	804.9766978	379.2820022	697.53127	1206.975423

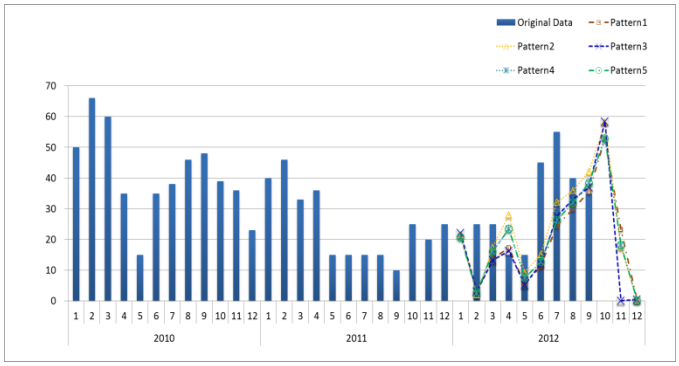


Fig. 12. Forecasting from the Tiny Data close to 0 Added case in Monthly Trend Removal (2-(2-2))

TABLE XIV. VARIANCE OF FORECASTING ERROR (2-(2-2))

Pattern1	Pattern2	Pattern3	Pattern4	Pattern5
603.3037984	579.9163939	585.7549099	565.5003321	565.5003321

Table 17 shows the summary for Diagnostic X-ray image processing apparatus by the Variance of forecasting error.

TABLE XV. SUMMARY FOR DIAGNOSTIC X-RAY IMAGE PROCESSING APPARATUS

		Monthly Trend Removal			Non Monthly Trend Removal		
Name	Diagnostic X-ray image processing apparatus of other	Accumulated Data	Forecasting Value－Accumulated Value	Tiny data close to 0 added case	Accumulated Data	Forecasting Value－Accumulated Value	Tiny data close to 0 added case
Minimum variance of Forecasting Error		13453.32873	379.2820022	565.5003321	11033.83677	515.2469097	336.4640944

IV. DISCUSSION

In the case of X-ray image intensifier tube device, Monthly Trend Removal case was better than Non-Monthly Trend

Removal case. This time series had a rather clear monthly trend and the result had reflected them. Forecasting from the accumulated data case (2-(2-1)) was better than those of the tiny data close to 0 added case (2-(2-2)) in this Monthly Trend Removal case for X-ray image intensifier tube device. On the other hand, in the case of Diagnostic X-ray image processing apparatus, Non-Monthly Trend Removal case was better than Monthly Trend Removal case. The time series of Diagnostic X-ray image processing apparatus does not have clear monthly trend. Forecasting from the tiny data close to 0 added case (1-(2-2)) was better than those of Accumulated data case (1-(2-1)). By the way, forecasting of accumulated data (1-(1), 2-(1)) shows rather good result. It can be used as one of the tool to decide when and how much volume to procure the materials etc.. It can be utilized as a new method to procure in supply chain management.

V. CONCLUSION

The needs for intermittent demand forecasting are increasing. In this paper, a new method for cumulative forecasting method was proposed. The data was cumulated and to this cumulated time series, the new method was applied to improve the forecasting accuracy. The forecasting result was compared with those of the non-cumulative forecasting method. The new method shows that it is useful for the forecasting of intermittent demand data. Forecasting from the accumulated data case (2-(2-1)) was better than those of the tiny data close to 0 added case (2-(2-2)) in this Monthly Trend Removal case for X-ray image intensifier tube device. On the other hand, in the case of Diagnostic X-ray image processing apparatus, forecasting from the tiny data close to 0 added case (1-(2-2)) was better than those of Accumulated data case (1-(2-1)). Among them, forecasting of accumulated data (1-(1), 2-(1)) shows rather good result. It can be used as one of the tool to decide when and how much volume to procure the materials etc.. It can be utilized as a new method to procure in supply chain management.

VI. FUTURE WORKS

It is our future works to investigate much further cases to confirm the effectiveness of our new method. The effectiveness of this method should be examined in various cases.

REFERENCES

[1] Box, G..E.P., Jenkins, G.M.& Reinsel, G.C.,Time Series analysis: forecasting and control, Wiley, 4th edn.,2008.

[2] Lydia Shenstone and Rob J. Hyndma, “Stochastic models underlying Croston’s method for intermittent demand forecasting”, Journal of Forecasting, 24:389-402,2005.

[3] Kenji Tanaka, Yukihiro Miyamura and Jing Zhang, “The Cluster Grouping Approach of Sales Forecasting Model for Book Publishing”, International Journal of Japan Association for Management Systems, Vol.4, No.1,pp.31-35,2012.

[4] Nguyen Khoa Viet Froung, Shin Sangmun, Vo Thanh Nha, Kwon Ichon, “Intermittent Demand forecasting by using Neural Network with simulated data”, Proceedings of the 2011 International Engineering and Operations Management Kuala Lumpur, Malasia, pp.723-728,2011.

[5] Kazuhiro Takeyasu and Keiko Nagata, “Estimation of Smoothing Constant of Minimum Variance with Optimal Parameters of Weight”, International Journal of Computational Science Vol.4,No.5, pp. 411-425,2010

An Algorithm Research for Supply Chain Management Optimization Model

Ruomeng Kong
School of sciences
Linyi University
Linyi, Shandong, P.R. China

Chengjiang Yin
Feixian school
Linyi University
Feixian, Shandong, P.R. China

***Abstract*—In this paper, we consider the extended linear complementarity problem on supply chain management optimization model. We first give a global error bound for the extended linear complementarity problem, and then propose a new type of algorithm based on the error bound estimation. Both the global and quadratic rate of convergence are established. These conclusions can be viewed as extensions of previously known results.**

Keywords—supply chain management optimization model; the extended linear complementarity problem; error bound; algorithm; quadratical convergence

I. INTRODUCTION

We consider a solution method for the extended linear complementarity problem on supply chain management optimization model. Letting $F(x) = Mx + p$, $G(x) = Nx + q$, the extended linear complementarity problem, abbreviated as ELCP, is to find a vector $x^* \in R^n$ such that

$$F(x^*) \in K, G(x^*) \in K^0, F(x^*)^T G(x^*) = 0, \qquad (1)$$

where M, $N \in R^{m\times n}$, $p, q \in R^m$, $A \in R^{s\times m}$, $B \in R^{t\times m}$, and

$$K = \{v \in R^m \mid Av \geq 0, Bv = 0\},$$

$$K^0 = \{u \in R^m \mid u = A^T\lambda_1 + B^T\lambda_2, \lambda_1 \in R^s_+, \lambda_2 \in R^t\}.$$

The solution set of the ELCP is denoted by X^*, which is assumed to be nonempty throughout this paper.

As is well-known, the extended linear complementarity problem (ELCP) is a special case of the extended nonlinear complementarity (ENCP) which plays a significant role in supply chain management. The topics of supply chain modeling, analysis, computation, and management are of great interests, both from practical and research perspectives. Research in this area is interdisciplinary by nature since it involves manufacturing, transportation, logistics, and retailing/marketing.

A lot of literatures have paid much attention to this area. See [1,2,3] for a recent surveys. Nagurney et al. ([4]) developed a variational inequality based supply chain network equilibrium model consisting of three tiers of decision-makers in the network. They established some governing equilibrium conditions based on the optimality conditions of the decision-makers along with the market equilibrium conditions. Dong et al.([5]) establish the finite-dimensional variational inequality formulation for a supply chain network model consisting of manufacturers and retailers in which the demands associated with the retail outlets are random. Nagurney et al. ([6]) establish the finite-dimensional variational inequality formulation for a supply chain network model in which both physical and electronic transactions are allowed and in which supply side risk as well as demand side risk are included in the formulation. The model consists of three tiers of decision-makers: the manufacturers, the distributors, and the retailers, with the demands associated with the retail outlets being random.

In recent years, many efficient solution methods have been proposed for solving it ([7, 8]). The basic idea of these methods is to reformulate the problem as an unconstrained or simply constrained optimization problem ([7,8]).

It is well-known that nonsingularity of Jacobian at a solution guarantees that the famous Levenberg-Marquardt (L-M) method for ELCP has a quadratic rate of convergence ([8]). Recently, Yamashita and Fukushima showed that the L-M method has a quadratic rate of convergence under the assumption of local error bound, which is much weaker than the nonsingularity of Jacobian([9]). This motivates us to consider the error bound estimation for the ELCP.

The paper is organized as follows. In Section 2, we recall the error bound for the ELCP. In Section 3, using the obtained result of error bound, the famousL-M algorithm is employed for obtaining solution of the ELCP, and we establish its the global and quadratic convergence based on the established error bound. Section 4 concludes this paper. Moreover, we do not require M and N to be square, and compared with the algorithm converges in [8], our conditions are weaker. These conclusions can be viewed as extensions of results in [8].

Some notations used in this paper are in order. Use R^n_+ to denote the nonnegative orthant in R^n; x_+ and x_- denote the orthogonal projections of vector $x \in R^n$ onto R^n_+, that is, $(x_+)_i := \max\{x_i, 0\}, (x_-)_i := \max\{-x_i, 0\}, 1 \leq i \leq n$; the norm $\|\cdot\|$ denotes the Euclidean 2-norm, the transpose of a matrix M be denoted by M^T. Without of making confusion, we denote a nonnegative vector $x \in R^n$ by $x \geq 0$.

II. PRELIMINARY

In this section, we mainly quote some known results on the error bound from [10] for ELCP. First, we give the needed assumptions.

Assumption 1 For A, M, N are the matrices defined in (1).

(A1) The matrix $M^T N$ is semi-definite (not necessarily symmetric);

(A2) The matrix A^T is column-full rank.

Under Assumption (A2), we can establish the following equivalent formulation of the ELCP([4]).

$$\begin{cases} AF(x) \geq 0, \\ BF(x) = 0, \\ (F(x))^{\bullet}\, G(x) = 0, \\ UG(x) \geq 0, \\ VG(x) = 0, \end{cases} \tag{2}$$

where

$$U = \{-A_L^{-1}B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I] + A_L^{-1}\},$$
$$V = \{A^{\bullet}\,\{-A_L^{-1}B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I] + A_L^{-1}\}$$
$$+B^{\bullet}\,[(A^{\bullet}\,A_L^{-1} - I)B^{\bullet}\,]^{+}[A^{\bullet}\,A_L^{-1} - I] - I\}.$$

The following result from Ref.4 mainly discusses the error bound for ELCP which will be applied to convergence of algorithm in next section.

Assumption 2 For system (2), there exists point $\hat{x} \in \Omega$, such that $AF(\hat{x}) > 0, UG(\hat{x}) > 0,$

$$\Omega = \{x \in R^n \mid BF(x) = 0, VG(x) = 0\}.$$

Theorem 1 Suppose that that Assumption 1(A1) and (A2) hold, and matrix $\left((AM)^T, (UN)^T\right)^T$ is of column full rank. Then there exists constant $\eta_1 > 0$ such that

$$\operatorname{dist}(x, X^*) \leq \eta_1 \{\| B(Mx+p)\| + \| V(Nx+q)\| + r(x)\}, \forall x \in R^n, \tag{3}$$

where $r(x) = \| \min\{Mx+p), Nx+q\} \|$.

III. ALGORITHM AND CONVERGENCE

In this section, we propose a new type of solution method to solve the ELCP based on the error bound results in Theorem 1, and the global and quadratic rate of convergence is also established, which was introduced first by Wang ([8]) for ENCP, but result of it was not given.

We now formulate the ELCP as a system of equations via the Fischer function ([11]) $\phi : R^2 \to R^1$ defined by

$$\phi(a,b) = \sqrt{a^2 + b^2} - a - b, a, b \in R.$$

A basic property of this function is that

$$\phi(a,b) = 0 \Leftrightarrow a \geq 0, b \geq 0, ab = 0.$$

For arbitrary vectors $a, b \in R^n$, we define a vector-valued function as follows

$$\Phi(a,b) = \begin{pmatrix} \phi(a_1, b_1) \\ \phi(a_2, b_2) \\ \cdots\cdots \\ \phi(a_n, b_n) \end{pmatrix}, \tag{4}$$

where $a = (a_1, a_2, \cdots, a_n)^T, b = (b_1, b_2, \cdots, b_n)^T$. Obviously,

$$\Phi(a,b) = 0 \Leftrightarrow a \geq 0, b \geq 0, a^T b = 0.$$

Using (4), we define a vector-valued function

$$\Psi : R^n \to R^{s+m+t}$$

and a real-valued function $f : R^n \to R$ as follows:

$$\Psi(x) = \begin{pmatrix} \Phi(AF(x), UG(x)) \\ BF(x) \\ VG(x) \end{pmatrix}, \tag{5}$$

$$f(x) = \frac{1}{2}\Psi(x)^T \Psi(x) = \frac{1}{2} \| \Psi(x) \|^2, \tag{6}$$

then the following result is straightforward.

Theorem 2 x^* is a solution of the ELCP if and only if $\Psi(x^*) = 0.$

In this following, allows us to extend above error bound in Theorem 1 to another residual function $\Psi(x)$. First, we give the following result in which Tseng ([12]) showed.

Lemma 1 For any $(a,b) \in R^2$, we have

$$(2 - \sqrt{2}) \mid \min\{a,b\} \mid \leq \mid \phi(a,b) \mid \leq (\sqrt{2} + 2) \mid \min\{a,b\} \mid.$$

By Lemma 1 and Theorem 1, we have the following result.

Theorem 3 Suppose that the conditions of Theorem 1 hold, then there exists a constant $\eta_2 > 0$ such that

$$dist(x, X^*) \leq \eta_2 \| \Psi(x) \|, \forall x \in R^n.$$

Proof Using Theorem 1, we have

$$\begin{aligned} \operatorname{dist}(x, X^*) &\leq \eta_1 \{\| B(Mx+p)\| + \| V(Nx+q)\| + r(x)\} \\ &\leq \eta_1 \{\| B(Mx+p)\|_1 + \| V(Nx+q)\|_1 \\ &\quad + c_1 \left\| \Phi(AF(x), UG(x)) \right\|_1 \} \\ &\leq \eta_1 \max\{1, c_1\} \{\| B(Mx+p)\|_1 \\ &\quad + \| V(Nx+q)\|_1 + \left\| \Phi(AF(x), UG(x)) \right\|_1 \} \\ &= \eta_1 \max\{1, c_1\} \{\| \Psi(x)\|_1\} \\ &\leq \eta_1 \max\{1, c_1\} \sqrt{2m+t} \{\| \Psi(x)\|\}, \end{aligned}$$

where the second inequality follows from Lemma 1 with constant $c_1 > 0$, the third inequality follows from the fact that $\| x\| \le \| x\|_1, \forall x \in R^n$, the last inequality follows from the fact that

$$\| x\|_1 \le \sqrt{n}\| x\|, \forall x \in R^n,$$

by letting $\eta_2 = \eta_1 \max\{1, c_1\}\sqrt{2m+t}$, then the desired result follows.

Clearly, this bound is an extensions of Theorem 2.1 in Mangasarian and Ren ([13]), Lemma 1 in Pang ([14]), and Corollary 3.2 in Xiu and Zhang ([15]).

Next, we review some definitions and basic results which will be used in the sequel.

The function $\Phi(x)$ is not differentiable everywhere with respect to $x \in R^n$. However, it is locally Lipschitzian, and therefore has a nonempty generalized Jacobian in the sense of Clarke ([16]). In the following, for a locally Lipschitzian mapping $\Theta: R^n \to R^m$, we let $\partial\Theta(x)$ to denote the Clarke's generalized Jacobian of $\Theta(x)$ at $x \in R^n$ which can be expressed as the convex hull of the set $\partial_B\Theta(x)$ ([17]), where

$$\partial_B\Theta(x) = \{V \in R^{m\times n} \mid V = \lim_{x^k \to \infty} \Theta'(x^k),$$
$$\Theta(x) \text{ is differentiable at } x^k \text{ for all k}\}.$$

Now, we recall some basic definitions about semi-smoothness and strong semi-smoothness.

A locally Lipschitz continuous vector valued function $\Theta: R^n \to R^m$ is said to be semi-smooth at $x \in R^n$, if the limit

$$\lim_{\substack{V \in \partial\Theta(x+th') \\ h' \to h, t \downarrow 0}} \{Vh'\}$$

exists for any $h \in R^n$.

It is well known that the directional derivative, denoted by $\Theta'(x;h)$, of Θ at x in the direction h exists for any $h \in R^n$ if Θ is semi-smooth at x. The following properties about the semi-smooth function are due to Qi and Sun in [18].

Lemma 2 Suppose that $\Theta: R^n \to R^m$ is a locally Lipschitz function and semi-smooth, then

a) for any $V \in \partial\Theta(x+h), h \to 0,$

$$Vh - \Theta'(x;h) = o(\| h\|);$$

b) for any $h \to 0,$

$$\Theta(x+h) - \Theta(x) - \Theta'(x;h) = o(\| h\|).$$

Semi-smooth functions lie between Lipschitz functions and continuously differentiable functions, and both continuously differentiable functions and convex functions are semi-smooth. A stronger notion than semi-smoothness is strong semi-smoothness.

The function $\Theta: R^n \to R^m$ is said to be strongly semi-smooth at x if Θ is semi-smooth at x and for any $V \in \partial\Theta(x+h), h \to 0,$ it holds that

$$Vh - \Theta'(x;h) = o(\| h\|^2).$$

A favorable property of the function $f(x)$ is that it is continuously differentiable on the whole space R^n although $\Psi(x)$ is not in general. We summarize the differential properties of Ψ and f defined by (5) and (6) in the following lemma ([19,20]).

Lemma 3 For the vector-valued function Ψ and real-valued function f defined by (5) and (6), the following statements hold.

(a) Ψ is strongly semi-smooth.

(b) f is continuously differentiable, and its gradient at a point $x \in R^n$ is given by $\nabla f(x) = V^T\Psi(x)$, where V is an arbitrary element belonging to $V \in \partial\Psi(x)$.

From Lemma 3 and discussion above, we can obtain the following result.

Theorem 4 For $x \in X^*$, there exist constants $\delta \in (0,1)$ and $\eta_3 > 0$ such that

$$\| \Psi(x+h) - \Psi(x) - Vh\| \le \| h\|^2,$$
$$x+h, x \in \{x \in R^n \mid \| x - x^*\| \le \delta\}.$$

In this following, a method for solving the ELCP is outlined. It is similar to that in [8, 9], But we consider method for ELCP with Armijo step size rule, and discuss its global convergence.

Algorithm 1

Step 1: Choose any point $x_0 \in R^n$, parameters $\sigma, \beta, \gamma \in (0,1)$ and $\varepsilon \ge 0$. Let $k = 0$.

Step 2: If $\| \nabla f(x^k)\| \le \varepsilon$, stop; Otherwise, go to Step 3.

Step 3: Choose an element $V^k \in \partial\Psi(x^k)$. Let $d^k \in R^n$ be the solution of the linear system

$$((V^k)^T V^k + \mu^k I)d = -(V^k)^T \Psi(x^k).$$

If d^k satisfies

$$\| \Psi(x^k + d^k)\| \le \gamma \| \Psi(x^k)\|,$$

then $x^{k+1} = x^k + d^k$, $k := k+1$, go to Step 5. Otherwise, go to Step 4.

Step 4: Let m_k be the smallest non-negative integer m such that

$$f(x^k + \sigma^m d^k) \le f(x^k) + \beta\sigma^m \nabla f(x^k)^T d^k.$$

Let $x^{k+1} := x^k + \sigma^m d^k$.

Step 5: Let $\mu^{k+1} = \| \Psi(x^{k+1}) \|^2$, $k := k+1$, go to Step 2.

For the above Algorithm 1, we assume that Algorithm 1 generates an infinite sequence $\{x^k\}$. By Theorem 3, Theorem 4, combining the proof of Theorem 3.1 in [9], we can obtain the following global convergence theorem.

Theorem 5 Let $\{x^k\}$ be generated by Algorithm 1 for ELCP with line search, then any accumulation point of the sequence $\{x^k\}$ is a stationary point of f. Moreover, if an accumulation point x^* of the sequence $\{x^k\}$ is a solution of (5). Then $dist(x^k, X^*)$ converges to 0 quadratically.

In Theorem 5, we have showed that Algorithm 1 has a quadratic rate of convergence under local error bound, which is much weaker than the nonsingularity of Jacobian. it is an extensions of the algorithm converges conclusion in [8], which is a new result for ELCP.

IV. Conclusion

In this paper, we consider an algorithm for the extended linear complementarity problem on supply chain management optimization model. To this end, we first give the global error bound for the ELCP, and use the error bound estimation to establish the global and quadratic convergence of algorithm for solving the ELCP.

Surely, under milder conditions, we may established global error bound for ELCP with the mapping being nonmonotone, and may use the error bound estimation to establish quick convergence rate of the Newton-type method for solving the ELCP instead of the nonsingular assumption just as was done for nonlinear equations in [9], this is a topic for future research.

Acknowledgment

The authors wish to give their sincere thanks to the editor and the anonymous referees for their valuable suggestions and helpful comments which improved the presentation of the paper. This work was supported by the Logistics Teaching and Research Reformation Projects for Chinese Universities (JZW2014048, JZW2014049),the Shandong Province Science and Technology Development Projects (2013GGA13034), the Natural Science Foundation of Shandong Province (ZR2010AL005, ZR2011FL017).

References

[1] H. Stadtler and C. Kilger, Supply Chain Management and Advanced Planning, Springer-Verlag, Berlin, Germany, 2002.

[2] C. C. Poirier, Supply Chain Optimization: Building a Total Business Network, Berrett-Kochler Publishers, San Francisco, California, 1996.

[3] C. C. Poirier, Advanced Supply Chain Management: How to Build a Sustained Competitive Advantage, Berrett-Kochler Publishers, San Francisco, California, 1999.

[4] A. Nagurney, J. Dong, and D. Zhang, "A Supply Chain Network Equilibrium Model", Transportation Research E, 38, pp. 281-303,2002.

[5] J. Dong, D. Zhang, A. Nagurney, "A Supply Chain Network Equilibrium Model with Random Demands", European Journal of Operational Research, 156, pp.194-212, 2004.

[6] A. Nagurney, J. Cruz, J. Dong and D. Zhang, "Supply Chain Networks, Electronic Commerce, and Supply Side and Demand Side Risk", European Journal of Operational Research, 164, pp.120-142,2005.

[7] F. Facchinei, J.S. Pang, Finite-Dimensional Variational Inequality and Complementarity Problems, Springer, New York, 2003

[8] Y.J. Wang, F.M. Ma and J.Z. Zhang, "A nonsmooth L-M method for solving the generalized nonlinear complementarity problem over a polyhedral cone", Appl. Math. Optim., 52(1), pp.73-92,2005.

[9] N. Yamashita and M. Fukushima, "On the rate of convergence of the Levenberg-Marquardt method", Computing [Suppl], 15, pp.239-249,2001.

[10] H.C. Sun, Y.J. Wang, "Further discussion on the error bound for generalized linear complementarity problem over a polyhedral cone", J. Optim. Theory Appl., 159(1):93–107, 2013.

[11] A. Fischer, "A special Newton-type optimization method", Optim., 24, pp. 269-284,1992.

[12] P. Tseng, "Growth behavior of a class of merit function for the nonlinear complementarity problem", J. Optim.Theory Appl., 89, pp.17-37,1996.

[13] O.L. Mangasarian and J. Ren , "New improved error bound for the linear complementtarity problem", Math. Programming, 66, pp. 241-255, 1994.

[14] J.S. Pang, "Inexact Newton methods for the nonlinear complementarity problem", Math. Programming, 36, pp.54-71, 1986.

[15] N.H. Xiu and J.Z. Zhang, "Global projection-type error bound for general variational inequlities", J. Optim. Theory Appl., 112(1), pp. 213-228, 2002.

[16] F.H. Clarke, Optimization and Nonsmooth Analysis, John Wiley and Sons, New York, NY, 1983.

[17] L. Qi, "Convergence analysis of some algorithms for solving nonsmooth equations", Math. Oper. Res., 18, pp.227-244,1993.

[18] L. Qi, J. Sun, "A nonsmooth version of Newton's method", Math. Programming, 58, pp. 353-367, 1993.

[19] F. Facchinei and J. Soares, "A new merit function for nonliner complementarity problems and a related algorithm", SIAM J. Optim., 7, pp.225--247,1997.

[20] N. Yamashita, M. Fukushima, "Modified Newton methods for solving a semismooth reformulation of monotone complementarity problems", Math. Programming, 76, pp.469-491,1997.

An Adaptive Hybrid Controller for DBMS Performance Tuning

Sherif Mosaad Abdel Fattah, Maha Attia Mahmoud, Laila Abd-Ellatif Abd-Elmegid
Department of Information Systems
Faculty of Computers and Information
Helwan University
Helwan, Egypt

***Abstract*—Performance tuning process of database management system (DBMS) is an expensive, complex and time consuming process to be handled by human experts. A proposed adaptive controller is developed that utilizes a hybrid model from fuzzy logic and regression analysis to tune the memory-resident data structures of DBMS. The fuzzy logic module uses flexible rule matrix with adaption techniques to deal with fluctuations and abrupt changes in the operation environment. The regression module predicts fluctuations in operation environment so the controller can take former action. Experimental results on standard benchmarks showed significant performance enhancement as compared to built-in self-tuning features.**

Keywords—automatic database tuning; fuzzy logic; adaptive controller; regression; self-tuning; DBMS

I. INTRODUCTION

Database management system performance tuning is a complex process with multiple objectives and tuning parameters. To know how to enhance such a process we need first to understand its characteristics and components. DBMS performance tuning can be generally described as a group of six activities to optimize the performance of a database[1].

Design Tuning tries to follow DB design best practices and normalizing DB tables to reveal un-optimized design issues that can degrade the performance. *SQL Tuning* tries to enhance the formulation of SQL statements to optimize the execution of the queries. *Memory Tuning* deals with allocating suitable values to the DB memory-resident data structures such as Shared Pool, Buffer Cache or Redo Log Buffer. *I/O Tuning* deals with I/O read/write anomalies such as disk fragmentation levels and tries to adjust its parameters for performance enhancements. *Connection Tuning* monitors network bandwidth and traffic and tries to optimize communication. *OS Tuning* investigates the system parameters and tries to adjust operation parameters such as virtual memory amount or size of memory page to enhance the performance of the DB environment.

DBMS performance tuning isn't an atomic process and it has a dynamic nature which makes its management harder and expensive due to need for an expert Database Administrator (DBA). The changes in the operation environment such as number of concurrent users, queries load, available memory space or network bandwidth can tend any performance tuning model to be unfeasible and outdated quickly if it can't adapt with these changes.

The term self-tuning databases[2] was coined for the aim of having a database that can learn and adapt with its environment with low or no interference from the human experts. To achieve this goal we have to depend on dynamic and adaptive control techniques such as fuzzy logic and nonlinear regression analysis.

In this paper, an adaptive hybrid controller (AHC) for DBMS memory-resident data structures is introduced. The controller utilizes hybrid model derived from fuzzy logic and regression analysis. The controller periodically monitors and feeds performance indicators of DBMS memory-resident data structures into fuzzy logic engine. The fuzzy logic engine fires corrective actions rules. The regression analysis module provides the controller with the ability to predict abrupt changes in the operation environment to further enhance the tuning process.

The rest of this paper is organized as follows: Section II describes preliminary concepts. Section III reviews previous work. Section IV introduces our proposed solution. Section V illustrates the experimental evaluations and results. Finally, Section VI concludes the paper and lists future work.

II. BACKGROUND

A. DBMS Memory-Resident Data Structures

DBMS memory resident data structures play a critical role in the process of tuning the DBMS performance. As it may decrease/increase the time and memory needed to execute queries and transaction on the database. There are three common data structures in any modern DBMS; Redo Log Buffer, Shared SQL Pool and Data Block Buffer [3] we are going to introduce the Data Block Buffer as it is the focus of this research in the following section.

The data block buffer cache (DBB) is the space reserved in memory for holding data blocks. The larger the DBB parameter value, the more memory is available for holding data blocks. The actual size of the DBB in bytes is computed as follows:

$$DBB = DB\_BLOCK\_BUFFERS \times DB\_BLOCK\_SIZE \quad (1)$$

The efficiency of the cache is measured by a metric called the data block buffer hit ratio (DBB-HR) that records the percentage of times a requested data block is available in the cache out of the total number of requests. When a data block is read in cache, it is called a logical read (LR). When the block is read from disk, it is called a physical read (PR).

$$(DBB\ Hit\ Ratio) = \frac{LR-\mathrm{PR}}{LR} \qquad (2)$$

For less than 20 concurrent users DBB-HR should be between 91% and 94%. Otherwise, it should exceed 94% in a healthy DBMS instance [3].

B. Fuzzy logic

Fuzzy logic (FL) mimics the ability of human brain in the usage of reasoning modes that are approximate rather than exact [4]. In traditional computing models, decisions are based on certainty and vigor but, this carries a cost of failure to deal with non-linear and complex problems that involve uncertainty in its characteristics. Examples to those problems can be, understanding human speech, sloppy handwriting, summarizing text or recognizing images.

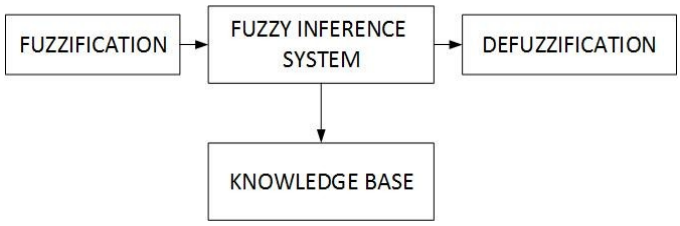


Fig. 1. Fuzzy Control Process

With Fuzzy Logic, decision rules are mapped to words rather than numbers. Computing based on words rather than exact number has tolerance to deal with uncertainty[5]. Broadly described, FL working scenario involves converting inputs of the problem from numerical nature (exact) to word based (approximate) nature in a process called Fuzzification. Then, the fuzzified inputs are supplied to the fuzzy inference engine which contains the inference rules to reach conclusions. Finally, the outputs are transformed from their approximate nature to an exact nature in a process called Defuzzification.

C. Regression Analysis

Regression analysis is a statistical technique used to predict the value of dependent variable (Y) given the values of independent variables (X1 … Xn)[6]. If the relation between the dependent and the independent variables is following a linear equation it is called linear regression and it can be represented by the following equation[7]:

$$Y = a + b_1 X_1 + \cdots + b_n X_n \qquad (3)$$

Where, *a* is intercept (the value of Y when x =0) and set (b_1 … b_n) represents slope of the line according to multiple (n) dimensions [X_1… X_n]

If the relation between the dependent and the independent variables is following a non-linear equation it is called non-linear regression. There are multiple models for nonlinear regression for example exponential model, power model or polynomial model[8].

III. Releated Wrok

The work in databases performance tuning started from decades and has been refined many times starting from the relational databases design concepts such as normalization forms and relational constraints to self-tuning databases ideas.

Ways in databases design tuning such as index pruning table and materialized views were addressed in [9][10]. Physical database tuning and the use of self-healing performance tuning methodologies were introduced in [11][12]. In [13], an modular approach was presented for providing self-healing database functionalities.

Each module in the system is assigned to a specific monitoring handler. In [14], a new way for physical data file organization based on search queries was proposed. Search queries are used to cluster similar records and to store them in one cluster block. So, I/O operations can be optimized in the physical layer. [1] Introduced a statistical approach to rank and evaluate the effect of database performance tuning parameters. In [15], operation research (OR) techniques were used to probe the SQL queries to optimize database logical design structures (schema) such as indexes or materialized views. [16] Introduced a neural networks based controller. Data mining techniques were used to analyze the database's log file to extract operation features.

Then, the result is used to train the neural network for controlling database's buffer cache levels. In [17], a fuzzy logic controller was introduced to tune the performance of web servers in terms of request response-time over multiple service level classes. Each Service level class will be assigned response-time level. The controller task is to maintain those service levels of response-time for each class when the server is heavy loaded. The work in [17] was extended in [18] to manage configuration of virtual machines on cloud-computing environments according to user's quality of service classes.

IV. Contribution

This research proposes a controller that employs hybrid criteria between fuzzy logic and regression analysis to adaptively tune the size of DBMS memory-resident data structures. The following figure describes the main components of the proposed controller:

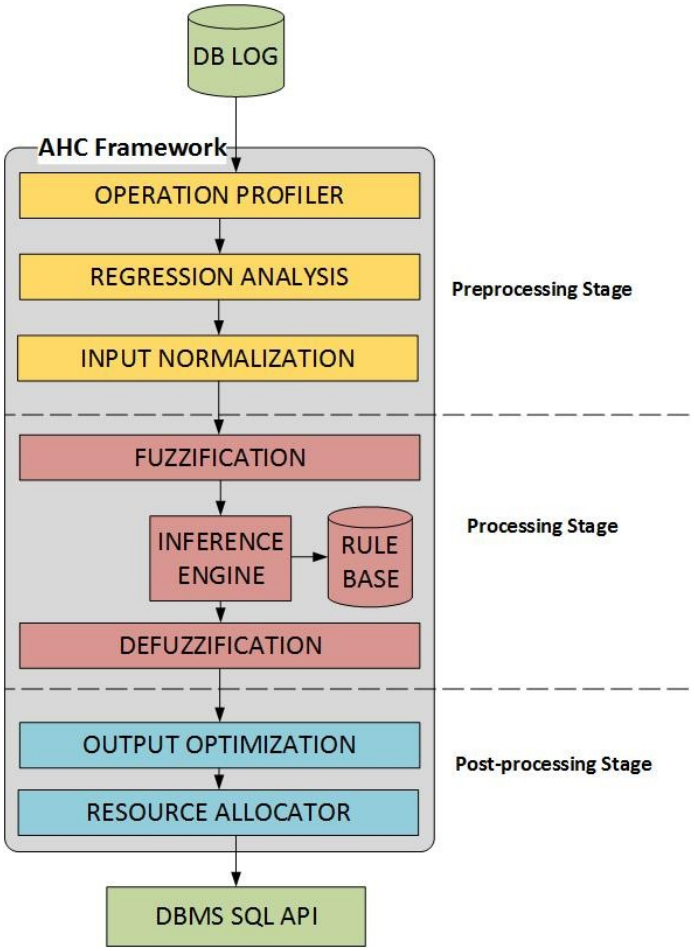


Fig. 2. Proposed System Architecture

In the next section we are going to explore AHC framework in a modular approach.

A. Preprocessing Stage

1) Operation Profiler

That module is responsible for collecting the current error values for each configured performance parameter. Performance parameters are configured in XML file. The DB admin configures the set of performance parameters along with their reference values. The error calculation depends on the current parameter value and its reference value equated as follows:

$$e(res) = \begin{cases} \frac{c(res)-r(res)}{r(res)}, & 0 \le c(res) \le 2r(res); \\ 1 \quad , & c(res) > 2r(res); \end{cases} \quad (4)$$

c(res) stands for the current resource value, *r(res)* for its reference value and *e(res)* for the error value.

2) Regression Analysis

Regression module is activated after a configured number of tuning cycles to collect sufficient amount of data. Regression type - linear or nonlinear - can be configured by the DB admin. For linear regression equation (5) is utilized to calculate the next value for the performance parameter. For, nonlinear regression the polynomial regression[8] is utilized using the following equation:

$$y = a + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (5)$$

As, a is the intercept and β is the regression coefficient for variables $X_1 \ldots X_k$ and y is the value to predict. Regression module is used to predict next error value. The input error value is the average between the current and predicted error values to help the controller to deal with abrupt changes in the environment.

3) Input Normalization

The input error value is normalized to avoid overshooting resource allocation due to peaks or dynamic resources changes. The error *e(res)* and error difference $\Delta_{e(res)}$ are normalized using the following equations [17]:

$$N_e(i+1) = |(1-\gamma) * N_e(i) + \gamma * e(res)| \quad (6)$$

$$N_{\Delta e}(i+1) = |(1-\gamma) * N_{\Delta e}(i) - \gamma * \Delta e(res)| \quad (7)$$

N_e stands for the normalized error, $N_{\Delta e}$ for normalized error difference and γ for the constant weight which equals to 0.8. This normalization technique homogenizes the current error value with its past values while, giving more weight for the current one. Note that the actual input to the resource controller module is $N_e * e(res)$ and $N_{\Delta e} * \Delta e(res)$. The sign is positive in equation (6) as error values take different signs in fluctuations and it is negative in equation (7) as the values of error difference take the same sign in fluctuations.

B. Processing Stage

1) Fuzzification

The input values *e(res)* and $\Delta_{e(res)}$ are fuzzified using triangular membership functions[5]. In triangular membership function the membership is calculated according to equation (8).

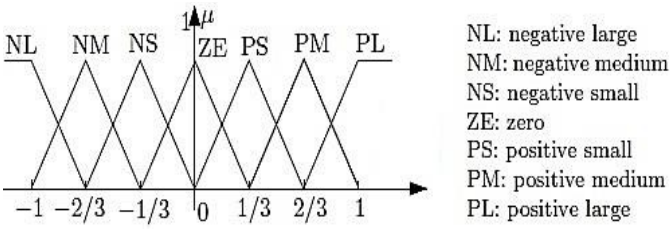


Fig. 3. Input Membership functions Graph

$$F(x) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (8)$$

Where a, b (center) and c are triangle membership vertices from left to right.

Table I. INPUT MEMBERSHIP FUNCTIONS

Membership Function	a	b	c
Negative Large (NL)	-∞	-1	-2/3
Negative Medium (NM)	-1	-2/3	-1/3
Negative Small (NS)	-2/3	-1/3	0
Zero (ZE)	-1/3	0	1/3
Positive Small (PS)	0	1/3	2/3
Positive Medium (PM)	1/3	2/3	1
Positive Large (PL)	2/3	1	∞

2) Inference Engine

The utilized inference mechanism is derived from [17] [18] and adjustment value is calculated using equation (9):

$$U_{(i+1)} = U_{(i)} + \alpha * Adj_{max} * \Delta U_{(i)} \quad (9)$$

As $U_{(i+1)}$ represents the adjustment for the current resource value, $U_{(i)}$ the current value of the resource in interval (i),

α , Adj_{max} are the output optimization factors and $\Delta U_{(i)}$ the inference engine output. α , Adj_{max} are illustrated in output optimization section.

The following algorithm specifies steps for calculating $\Delta U_{(i)}$

Input: e(res) , Δe(res).
Output: $\Delta U_{(i)}$
Steps:
for each mf ϵ Input_MembershipFunctions
x = TriangularMembership(e(res), mf)
y = TriangularMembership(Δe(res), mf)
if x > 0 THEN
add x to e(res)_MembershipFunctions
if y > 0 THEN
add y to Δe(res)_ MembershipFunctions
rules = combine(e(res)_MembershipFunctions, Δe(res)_ MembershipFunctions)
for each r ϵ rules
match = getMatrixMatch(r)
center = Center(match)
μ_r = MIN(r)
add center to Centers
add μ_r to WeightSet
Output = CenterOfGravity(Centers, WeightSet)

The following table represents the proposed rule matrix. Columns and rows represent membership functions of Δe(res) and e(res) respectively. Each element in the table construct a rule for example, (NL,NL) > PL

Table II. RULE MATRIX FOR $\Delta U_{(i)}$

ΔU(i)		Δe(res)						
		NL	**NM**	**NS**	**ZE**	**PS**	**PM**	**PL**
e(res)	**NL**	PL	PL	PL	PL	PM	PS	ZE
	NM	PL	PL	PM	PM	PS	ZE	NS
	NS	PL	PL	PM	PS	ZE	NS	NM
	ZE	PM	PM	PS	ZE	NS	NM	NM
	PS	PM	PS	ZE	NS	NM	NL	NL
	PM	PS	ZE	NS	NM	NL	NL	NL
	PL	ZE	NS	NM	NL	NL	NL	NL

Suppose for e(res) = 0.075 and Δe(res) = 0.3 the following membership functions have non zero membership

ZE (e(res)) = 0.6 , PS(e(res)) = 0.3

And for Δe(res) PS(Δe(res)) = 0.75

Then, combinations are generated from previous rules:

Rule1 (ZE,PS) with μ = MIN(0.6,0.75) = 0.6

Rule2 (PS,PS) with μ = MIN(0.3,0.75) = 0.3

Rule1 gives us the match NS from the rule matrix which has a center = $-\frac{1}{3}$

Rule2 gives us the match NM from the rule matrix which has a center = $-\frac{2}{3}$

3) Defuzzification

In defuzzification, the output $\Delta u(i)$ is calculated using the Center of Gravity equation (10) [5]:

$$\Delta U_{(i)} = \frac{\sum_{n}^{i=1} center(i) * \mu(i)}{\sum \mu(i)} \quad (10)$$

According to our example $\Delta U_{(i)}$ will equal $\frac{-4}{9}$

C. Post-processing Stage

1) Output Optimization

The output optimization factor (α) is used to handle process delay during resource allocation. It is summarized as the time between sending the new adjustment of a resource and the time the resource value is actually updated[19].

α is calculated with the same criteria as $\Delta U_{(i)}$ but with different membership functions and rule matrix.

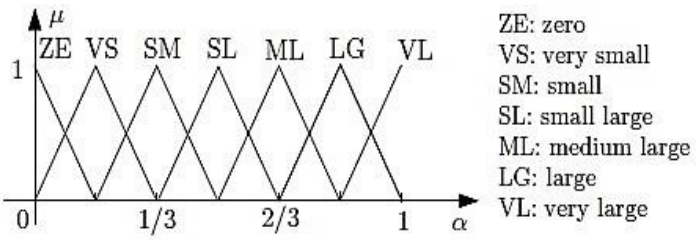


Fig. 4. α membership functions Graph

Table III. α MEMBERSHIP FUNCTIONS

Membership Function	a	b	c
Zero (ZE)	-∞	0	1/6
Very Small (VS)	0	1/6	1/3
Small Medium (SM)	1/6	1/3	1/2
Small Large (SL)	1/3	1/2	2/3
Medium Large (ML)	1/2	2/3	5/6
Large (LG)	2/3	5/6	1
Very Large (VL)	5/6	1	∞

The value of α is used to speed up or slow down the change magnitude of the resource value. In fluctuations periods it is set to small value to prevent overshooting in adjustment. When the current value is going away from the reference value it is set to relative large value to invert the change.

Table IV. RULE MATRIX FOR A

ΔU(i)		Δe(res)						
		NL	NM	NS	ZE	PS	PM	PL
e(res)	NL	VL	VL	VL	SM	VS	VS	ZE
	NM	VL	VL	LG	SL	SM	SM	SM
	NS	VL	VL	LG	ML	VS	SM	SL
	ZE	LG	ML	SL	ZE	SL	ML	LG
	PS	SL	SM	VS	ML	LG	LG	VL
	PM	SM	SM	SM	SL	LG	VL	VL
	PL	ZE	VS	VS	SM	VL	VL	VL

Adj_{max} is the max adjustment value that can be allocated. It is calculated as follows:

$$Adj_{max} = \left|\frac{c}{2} * e(res)\right| \qquad (11)$$

Where c is the current resource value. This equation is based on heuristic control rule[19] which states that the max resource adjustment shouldn't exceed half of the current resource value for stability of the system and to be proportional to the current error value for adaptability of the system.

2) Resource Allocator

This module concludes the work of the control cycle by sending the adjustment value(s) to the DBMS API for allocation using SQL commands.

V. EVALUATION AND EXPERIEMENTS

TPC-C and TPC-H benchmarks[20] are used to conduct the evaluation on data block buffer data structure; TPC-C is an online transaction processing (OLTP) benchmark. It involves a mix of five concurrent transactions of different types and complexity. TPC-H is a decision support benchmark. It consists of ad-hoc and concurrent queries. The operation environment for the experiments runs on Windows server 2008 with ORACLE 10g database server installed. Our proposed system is deployed as windows service. The user load is defined as 20 concurrent users that start with 2 concurrent users in the first transaction cycle and increase gradually by 2 until reaching 20 in the following run cycles. The tuning cycle period for the controller is set to 30 seconds (defined by trial and error to pose the minimum overhead on the DBMS performance while keeping track of workload changes). Regression is activated after 30 tuning cycles. It is configured to be nonlinear with variables; number of users per cycle and number of transactions per minute.

Figure 5 shows results for conducting TPC-C benchmark. The average response for the DBMS without tuning was 87 ms while the average response time after AHC tuning was 46 ms with 52% better than without tuning.

Figure 6 shows results for conducting TPC-H benchmark. The average response for the DBMS without tuning was 93.1 ms while the average response time after AHC tuning was 46.5 ms with 49% better than without tuning.

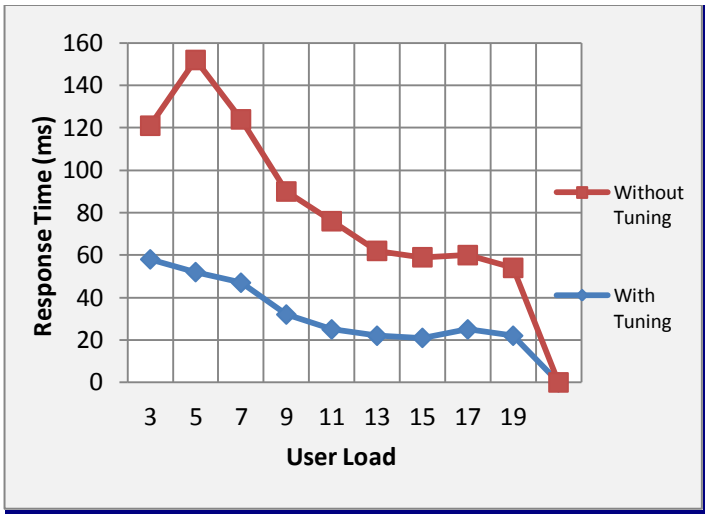


Fig. 5. TPC-C benchmark, comparing response time with and without tuning

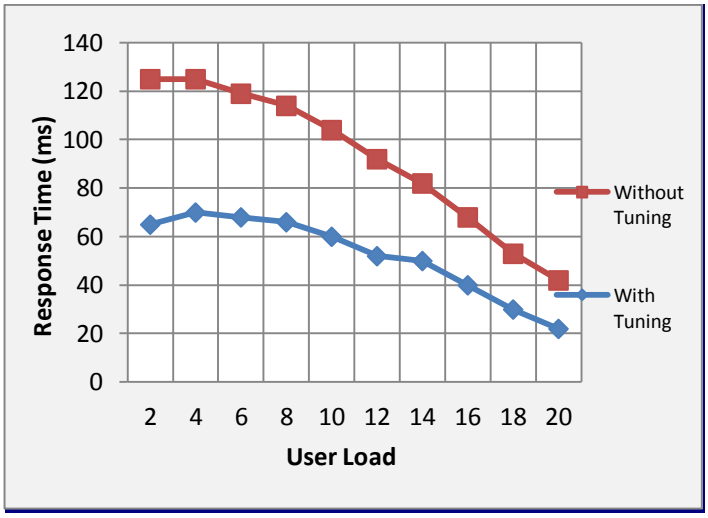


Fig. 6. TPC-H benchmark, comparing response time with and without tuning

VI. CONCLUSION AND FUTURE WORK

This paper proposes an adaptive hybrid controller (AHC) for tuning DBMS performance based on its resident data structures. AHC is featured by both generalization and adaptability. AHC generalization is achieved in three ways. First, operation profiler and resource allocator are developed to deal with any type of DBMS using Microsoft generic ADO API[21]. Second, the proposed rule matrices can be totally configured and adjusted using XML configuration file to adapt with different workload scenarios in a generic way. Third, regression analysis module allows the system to take former action with fluctuations and abrupt changes in the operation environment and workload. It can be configured by the DBA according to each resource case giving more flexibility to deal with different scenarios. AHC is adaptable in two ways. First, input normalization module normalizes inputs with its past values to reveal the fluctuations effect. Second, output optimization factors deal with process delay effect and scale the output to prevent down or over shooting in resources allocation.

Future work can include covering other memory-resident data structures of DBMS. Using machine learning techniques such as Neural Networks to equip our controller with the ability to learn the characteristics of its operation environment and to dynamically adjust its membership functions and rule matrices according to characteristics of work load on the operation environment.

References

[1] B. K. Debnath, D. J. Lilja And M. F. Mokbel, "Sard: A Statistical Approach For Ranking Database Tuning Parameters," Ieee 24th International Conference, 2008.

[2] R. V. Nehme, "Dtabase, Heal Thyself," Data Engg.Workshop, April 2008.

[3] S. S. Mittra, Database Performance Tuning And Optimization Using Oracle, Springer, 2003.

[4] J. H. Lilly, Fuzzy Control And Identification, Wiley, 2010.

[5] J. Jantzen, Foundations Of Fuzzy Control, John Wiley And Sons, 2007.

[6] A. O. Sykes, "An Introduction To Regression Analysis," [Online]. Available: Http://Www.Law.Uchicago.Edu/Files/Files/20.Sykes_.Regression.Pdf.

[7] A. Cottrel, "Regression Analysis: Basic Concepts," 2011. [Online]. Available: Http://Users.Wfu.Edu/Cottrell/Ecn215/Regress.Pdf.

[8] C. Ritz And J. C. Streibig, Nonlinear Regression With R, Springer, 2008.

[9] S. Agrawal, S. Chaudhuri And V. Narasayya, "Automated Selection Of Materialized Views And Indexes For Sql Databases," Microsoft Research, 2007.

[10] Surajit, Chaudhuri; Vivek, Narasayya;, "Self Tuning Database Systems : A Decade Progress," Microsoft Research, 2007.

[11] K. Philip, "Elements Of The Self-Healing System Problem Space," Ieee Data Engineering Bulletin, 2004.

[12] P. Liu, "Design And Implementation Of Self Healing Database System," Ieee Conference, 2005.

[13] R. V. Nehme, "Database, Heal Thyself," Data Engg. Workshop, 2008.

[14] B. G. L. A. A. Kitsopanidis, "Enhancing Database Retrieval Performance Using Record Clustering," Citeseerx, 2007.

[15] A. N. Chen, "Robust Optimization For Performance Tuning Of Modern Database Systems," European Journal Of Operational Research 171, 2006.

[16] U. P. K. S. F. Rodd, "Adaptive Tuning Algorithm For Performance Tuning Of Database Management System," International Journal Of Computer Science And Information Security, Vol. 8, 2010.

[17] J. Wei And C.-Z. Xu, "Eqos: Provisioning Of Client-Perceived End-To-End Qos Guarantees In Web Servers," Ieee Transaction On Computer, 2006.

[18] J. Rao, Y. Wei, J. Gong And C.-Z. Xu, "Dynaqos: Model-Free Self-Tuning Fuzzy Control Of Virtualized Resources For Qos Provisioning," Ieee Nineteenth Ieee International Workshop On Quality Of Service, 2011.

[19] F. G. Shinskey, Process Control Systems: Application, Design, And Tuning., Mcgraw-Hill, 1996.

[20] "Tpc," [Online]. Available: Http://Www.Tpc.Org/Tpcc/.

[21] Msdn, "Ado.Net," Microsoft, [Online]. Available: Http://Msdn.Microsoft.Com/En-Us/Library/E80y5yhx(V=Vs.110).Aspx.

[22] A. M. Brown, "A Step-By-Step Guide To Non-Linear Regression Analysis Of Experimental Data Using A Microsoft Excel Spreadsheet," Computer Methods And Programs In Biomedicine , Vol. 65, P. 191–200, (2001).

Designing a Multi Agent System Architecture for IT Governance Platform

S. ELHASNAOUI, H. MEDROMI, S.FARIS, H.IGUER, A. SAYOUTI
(EAS- LISER) Systems Architecture Team
ENSEM, Hassan II University
BP.8118, Oasis Casablanca

***Abstract*—This paper presents a multi-agents architecture which facilitates the integration of three major IT governance frameworks: COBIT5, ITIL V3 and ISO/IEC27002, to optimize the construction of a distributed system. This architecture proposes a new and easier method to develop a distributed multi agents system, where agents involved in this system can communicate in a distributed way thanks to functionalities offered by the system. It gives finally an overview of implementation of a prototype of the proposed solution limited for the moment to integration of processes most used in the majority of information systems.**

Keywords— IT Governance; Multi Agent System; COBIT 5; ITIL V3; ISO/IEC 27001/27001; Process; Information System Introduction

I. INTODUCTION

Governance ensures that enterprise objectives are achieved by evaluating stakeholder needs, conditions and options; setting direction through prioritization and decision making; and monitoring performance, compliance and progress against agreed-on direction and objectives [1].

There are several methodologies, standards, frameworks and good practices for Governance of information system. The most applicable and used today are ISO/IEC 27002, COBIT and ITIL [2]. Each has its positive aspects and its limitations. As a result, this paper aims to combine the three major frameworks for IT governance: ISO / IEC 27002, COBIT and ITIL in a comprehensive manner and propose an IT governance structure that covers broader aspects and supports all the needs of the organizations for more efficient IT management.

This paper describes a multi-agents architecture based on three major IT governance frameworks which are COBIT 5, ITIL V3 and ISO/IEC 27002. One of the most characteristics is the use of intelligent agents as the main components which focus on distributing the majority of the system's functionalities into processes [3].

We focused our multi-agents system architecture for IT governance by configuring our paper with the following parts: abstract of paper, brief introduction of technically requirements, fundamental aspects of state of the art which gives an idea about the theme of IT governance and the three major frameworks: COBIT 5, ITIL V3 and ISO/IEC27002, then we present an overview of multi agents system architecture and agents aspects, after that we present an IT Governance prototype which integrate some governance processes, finally a conclusion for this paper and references.

II. IT GOVERNANCE AND FRAMEWORKS OF GOOD PRACTICES

A. What is IT Governance?

Governance of information systems is an expression frequently used today in the world of management information technology [4] [5]. In fact, until we heard about the governance in administrative and policy areas but not in the computer world [6]. The word is obviously subject to multiple uses. In its broadest sense, governance can sign a way to direct; to manage or control [7].

IT governance corresponds to the implementation of tools which stakeholders consider it in the operation of the information system (IS) [8] [9]. IT governance aims to define the objectives of the information system in order to plan, define and implement processes related to the management of the lifecycle of the IS [10].

These activities are based on the control and performance measurement of these processes [11]. The famous organization ISACA (Information Systems Audit and Control Association) who gives a lot of interest in the governance field defines five main axis:

- Strategic Alignment
- Value Creation
- Risk Management
- Resource Management
- Performance measurement.

B. Frameworks

IT department has recourse to good practice's guidelines: Production with ITIL, governance with COBIT, security with ISO 27000, project management with PMBOK ... [12]. These standards promote broadcasting of good practices in the company, continuous improvement, homogeneity of the process and contributing to the professionalization of the services delivery [13] [14].

1) COBIT 5

COBIT (Control Objectives for Information Business year related Technology) is a methodology for evaluating IT

services in the company [15]. This approach is based on an association of 37 processes (best practices collected from experts SI) and on objective indicators (KGI) and (KPIs), to put the process under control in order to provide data for the company to achieve its objectives (alignment of technology on business strategy).

This is a control framework that aims to manage risks (security, reliability, compliance) and investment. [16] It does not provide guidance or recommendations to technical (technological choices, consolidation, crisis management ...). In other words, COBIT focuses on what the company needs to do, not how it should do. "

2) ITIL V3

ITIL [5] is an acronym for "Information Technology Infrastructure Library" (IT Infrastructure Library) .

ITIL Version 3 describes the service as an organization of human resources and IT (hardware and software) , whose objective is the provision of value for the company and the beneficiary of the service . [17]

ITIL provides an IT organization:

- Show that IT delivers value to the enterprise;
- Identify the business needs of the company and adapt IT to these needs and not the reverse ;
- Do not lose sight of the intake value that can have computers ;
- Focus on Quality of IT service;
- Improve the perception of IT services by customers / users ;
- Ensure the sustainability and development of the company.

ITIL Version 3 is interested in service throughout: the genesis of service to its end of life . [18] Five groups of activities have been identified:

- **Service Strategy:** Align IT strategy on business strategy , ensuring that the input value will enable the company to achieve its objectives .
- **Service Design**: Design Services from requirements collected by the Service Strategy .
- **Service Transition:** Ensuring the quality of the transition of a new service between studies and operations.
- **Service Operation**: Operate services effectively and efficiently .
- **Continual Service Improvement :** Creating conditions for continuous improvement of services .

3) ISO/IEC 27001/27002

ISO / IEC 27001 describes a process approach for establishing an ISMS (system safety management information system) . But if it sets the goal , it does not state specifically how it should achieve [19] . ISO 27002 presents a series of practical recommendations , addressing both technical and organizational aspects .

The standard defines a code of good practice for use by those responsible for implementing or maintaining a management system for information security . The information security is defined as " the preservation of confidentiality, integrity and availability of information" [20] .

The standard offers 11 major fields of security using 133 security objectives (controls) :

- Security Policy Information
- Organization of information security
- Asset Management
- Security related to human resources
- physical and environmental safeties
- Operation and Communications Management
- Access Control
- Acquisition , development and maintenance of information systems
- Incident Management
- Management Business Continuity
- Compliance.

III. Overview of Multi Agents System Architecture For IT Governance

A. Problematic

The responsible of information systems is facing a problem of IT governance due to blooming standards of good practice, there are those who want comprehensive, dealing with all areas of activity of IS without quite detailed, and those who treat a particular area in detail without a global view. Each share of particular concern: security, quality, customer services, auditing, project development, etc...

It is a necessary evil to recognize that each function in its own practices. Simultaneously raises the question of the establishment of a global, single framework for the IT department, which meets all expectations.

It's in this perspective that this work was done, the idea is to design a platform able to integrate three major IT governance frameworks which are COBIT 5, ITIL V3 and ISO/IEC 27002, using intelligent agents as the main components which focus on distributing the majority of the system's functionalities into processes.

B. The Multi -agent system

We conducted several studies to identify the best needs of the platform, and meet the expectations of users. The Modeling of the platform is based on the principle of SMA which is everyone must cooperate to achieve the same goal. For this reason, the architecture is composed of a set of different agents which communicate and cooperate with them in an intelligent manner.

1) What is an agent?

An agent is an entity (physical or abstract), characterized by the fact that it is autonomous in decision making, by his knowledge of itself and others, and its ability to act [21] [22] .

Experts in multi-agent systems have classified agents into three major categories according to an essential criteria that is the representation of its environment, these agents are: Reagents agents, Cognitive agents and Hybrid agents.

2) What is a multi agent system

A multi-agent system(MAS) is a distributed system consisting of a set of entities (programs)-relatively independent agents, each with their own thread, specific to fulfill goals, and ways to communicate and negotiate with other to accomplish their goal [23] [24].

MAS are ideally designed as a set of agents interacting in the cooperation, competition and coexistence manner.

Multi agents systems is a system composed of the following elements:

- An environment with a metric in general.
- A set of objects, which can associate a position in an environment in a given time. Agents can perceive, create, destroy and modify these objects.
- A set of agents, which represents the active entities of the system,
- A set of relationships between agents between them [25] [26].
- A set of operators that allow agents to perceive, produce, consume, transform and manipulate objects.

3) Contribution of MAS

The Multi-agent approach is justified by:

- Adaptation to reality
- Cooperation,
- The resolution of complex problems,
- Integration of incomplete expertise,
- Modularity,
- Efficiency,
- Reliability,
- Reuse.

C. IT Governance architecture based on MAS

The proposed model is a modular multi-agents architecture where all components are managed and controlled by different types of agents which are able to cooperate, propose solutions on very dynamic environments and face real problems.

There are different kinds of agents in the architecture, each one with specific roles, capabilities and characteristics. This fact facilitates the flexibility of the architecture in incorporating new agents.

As can be seen on Figure 3, the architecture defines three basic blocks which provide all functionalities of the architecture.

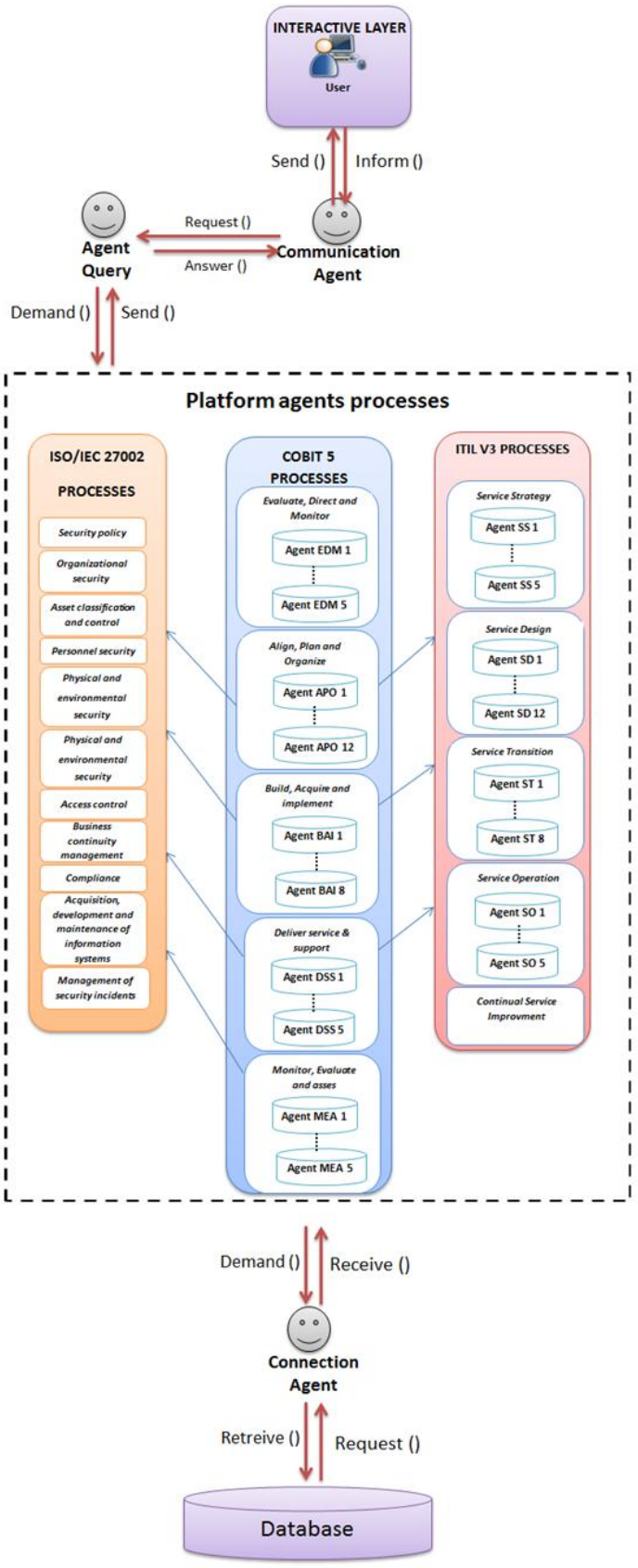


Fig. 1. The proposed multi-agent system model.

4) INTERFACE LAYER:

This represents all the users that can use to exploit the system functionalities.

5) AGENTS PLATFORM:

This is the core of this architecture, integrating a set of agents, each one with special characteristics and behavior. These are agents participating in the functioning of the system:

*a) **Communication Agent:** This agent is responsible for all communications between users and the platform. It manages the incoming requests from users to be managed by Agent Query. And send answers to Interface Layer.*

*b) **Agent Query:** This agent receives requests from communication agent and decides which agent process must be called by taking into account the request of users. Users can explicitly invoke a process, or can let the Agent Query decide which process from which framework is best to accomplish the requested task.*

*c) **Agent's Processes:** These agents manage the list of processes that can be used by the system; they represent processes of the three integrated frameworks: COBIT 5, ITIL V3 and ISO/IEC 27002.*

*d) **Connection Agent:** This agent manages interaction between the platform of agent's process and database of the system. It retrieves adequate data and sends it the agent process concerned.*

6) DATABASE:

This part of the system includes all the data and tables used by all components of the platform, including static data, indicators related to agents, and appropriate decisions to the various scenarios of behavior to be submitted to Query Agent depending on the state of the collaboration between process agents.

IV. PRESENTATION OF THE PROTOTYPE OF THE IT GOVERNANCE PLATFORM

Considering the large number of processes of the three governance frameworks integrated into our platform, we conducted research to develop a prototype that takes into account processes most important and used by information systems, in order to integrate them firstly. Here below the selected process for implementing this prototype:

ITIL: incident management, problem management, change management, request fulfillment.

COBIT: risk management, management of enterprise architecture, program management, portfolio management, project management.

ISO / IEC 27002: Managing Security Policies,

In addition, our platform takes into consideration both profile management and access rights and user management.

A. Software architecture

Since our platform is a web application, we chose to adopt the MVC architecture (Model-View-Controller) which is used for interactive web-applications.

This model minimizes the coupling between business logic and data presentation to web user.

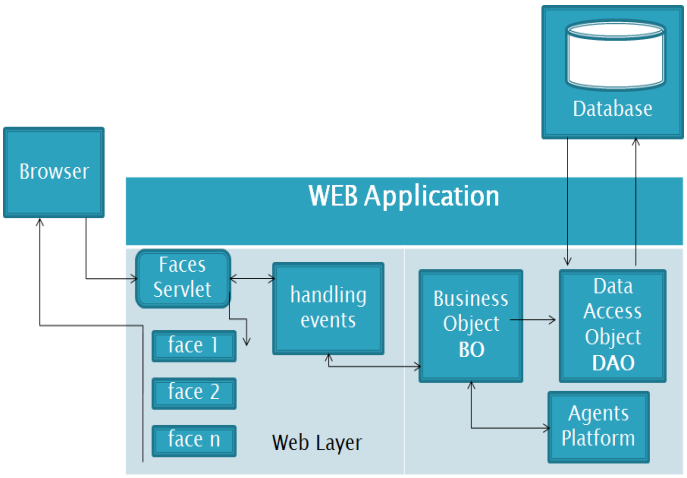


Fig. 2. Software architecture.

B. Use case diagram

Figure 3 shows use cases of the various features provided by various functionalities of the platform. All these functionalities can be managed by users after a successful authentication.

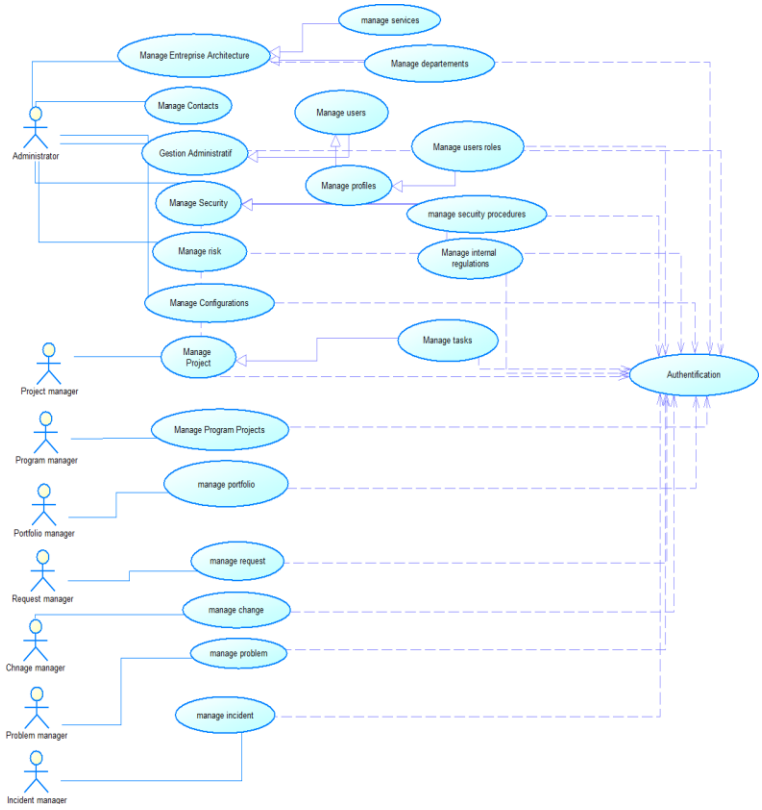


Fig. 3. Use case diagram of the prototype of IT Governance paltform.

V. PERSPECTIVES

Future work consists on extending our prototype by defining a detailed architecture of each agent and specifying

communication between these agents. Our objective is to validate the architecture that we propose in this paper by developing a distributed platform that will allow companies to establish effective governance of their information system.

VI. CONCLUSION

This article propose an architecture of IT Governance based on three major frameworks : COBIT 5, ITIL V3 and ISO/IEC 27002.

One of the most characteristics of proposed solution is the use of intelligent agents which communicate and cooperate in order to deliver answers of user's requests and to establish governance in the information system.

It briefly talked after that, about prototype of the IT Governance platform which integrate some processes from the integrated frameworks in order to validate the proposed architecture.

REFERENCES

[1] (AFAI, 2002) AFAI, ITGI, COBIT, Gouvernance, Contrôle et Audit de l'Information et des Technologies Associées, Troisième édition, 2002.

[2] (Agostinelli, 2009) S. Agostinelli, L'analyse du processus métier au coeur du système d'information oriente le projet d'intelligence économique. In, M., Ghenima, A., Ouhsel & S., Sidhom, (eds.).

[3] SIIE"2009 : 2e Conférence Internationale Systèmes d"Information et Intelligence Economique, (pp. 254-271). Nancy : IHE éditions, 2009.

[4] (Alonso, 1997) G. Alonso, D. El Abbadi, C. Mohan, Functionality and Limitation of Current Workflow Management Systems, IEEE Expert, 1997.

[5] (Ploesser, 2008) K. Ploesser, J. Recker and M. Rosemann, Towards a Classification and Lifecycle of Business Process Change, 9th Workshop on Business Process Modeling, Development, and Support (BPMDS'08), 16-17 June 2008, Montpellier, France.

[6] (Luftman, 2004b) J. Luftman, Assessing Business-IT Alignment Maturity, in Strategies for Information Technology Governance, Ed. by Van Grembergen, Idea Group, 2004, ISBN 1591401402.

[7] (Muehlen, 2008) M. zur Muehlen, D.T. Ho, Service Process Innovation: A Case Study of BPMN in Practice. In: Ralph Sprague, Jr. (Ed.): Proceedings of the 41st Hawai"i International Conference on System Sciences. Waikoloa, HI, January 7-10, 2008.

[8] Consilium-ICT, ITIL et la gouvernance des systèmes d'informations: vers une eadministration agile, Toulouse, Juin 2009.

[9] Maico Gehrmann, "Combining ITIL, COBIT and ISO/IEC 27002 for structuring comprehensive information technology for management in organizations", Navus – Revista de Gestão eTecnologia. Florianópolis, SC, v. 2, n. 2, p. 66 - 77, jul./dez. 2012

[10] S.Elhasnaoui, H. Medromi, A. Sayouti, « Etude de mise en œuvre des solutions conformes à ITIL et modélisation de la solution proposée », JDTIC 2012.

[11] S.Elhasnaoui, H. Medromi, A. Sayouti, Multi-agents modeling platform for IT governance based on ITIL" International Conference on Engineering Education and Research , ICEER 2013.

[12] S.Elhasnaoui, H. Medromi, A. Sayouti, "Une approche multiagents pour la gouvernance des systèmes d'information basée sur le référentiel ITIL », Journées Doctorales en Systèmes d'Information, Réseaux et Télécommunication (JDSIRT 2013).

[13] J.Skiti et H.Medromi, "La Gouverance des Technologies de l'Information à base du Système Multi-agent et le référentiel COBIT". Workshop on Information Technologies and Communication, WOTIC 2011, Casablanca, Morocco.

[14] Cigref, Gouvernance du système d'information, Problématiques et démarches. (Septembre 2002).

[15] Dario Forte, "Security standardization in incident management: the ITIL approach", Volume 2007, Issue 1, January 2007, Pages 14–16.

[16] M.N. Kooper, R. Maes, E.E.O. Roos Lindgreen "On the governance of information: Introducing a new concept of governance to support the management of information". International Journal of Information Management: The Journal for Information Professio- Professionals, Volume 31 Issue 3, June, 2011, Pages 195-200

[17] Eric LELEU, « Le COBIT : L'état de l'Art, Socle de la gouvernance des SI », 2009.

[18] Patrick Stachtchenko, « COBIT 5, ses apports pour le management et la gouvernance du SI », 25 Janvier 2013.

[19] Delbrayelle, Introduction à ITIL V3 et au cycle de vie des services, juillet 2011. ISO office ,"Information technology — Security techniques — Code of practice for information security management", 2005.

[20] ITGI and OGC, "Aligning CobiT® 4.1, ITIL® V3 and ISO/IEC 27002 for Business Benefit" 2008.

[21] Shoham, Y. Agent-oriented programming. Artificiel Intelligence, February 1992. Stanford,USA.

[22] Weiss, G., (Ed.), Multi-agent Systems. A Modern Approach to Distributed Artificial Intelligence, The MIT Press, Cambridge, Massachusetts, 1999.

[23] Dante I. Tapia, Sara Rodríguez, Javier Bajo, Juan M. Corchado, FUSION@, A SOABased Multi-Agent Architecture

[24] Sayouti, F. Moutaouakkil, H. Medromi. "The Interaction- Oriented Approach for Modeling and Implementing Multi-Agent Systems", International Review on Computers and Software (I.RE.CO.S), Vol. 5, N. 2, Mars 2010.

A.Sayouti, H. Medromi, Book Chapter in the book "Multi- AgentSystems - Modeling, Control, Programming, Simulations and Applications", ISBN 978-953-307-174-9, InTech, April4, 2011.

[25] A. SAYOUTI, H. MEDROMI –"Les Systèmes Multi-Agents : Application au Contrôle sur Internet" Auteurs Éditions universitaires européennes, Août 2012.].

[26] A.SAYOUTI & H. MEDROMI Autonomous and Intelligent Mobile Systems based on Multi-agent, Book Chapter in the book " Multi-agent Systems – Modeling Control , Programming, Simulations and Applications" .

Using Digital Image Processing to Make an Intelligent Gate

Sundus K. E.
College of Computer Sciences and Mathematics,
University of Mosul
Mosul, Iraq

AL_Mamare S. H.
College of Computer Sciences and Mathematics,
University of Mosul
Mosul, Iraq

***Abstract*—This paper presents an automatic system for controlling and dominating building gate based on digital image processing. The system begins with a digital camera, which captures a picture for that vehicle which intends to enter the building, then sends the picture to the computer. Image analyses performed to detect and recognize the vehicle, and matching the vehicle's image with the stored database of the permissible vehicles. Then, the computer sends a signal to the electro-mechanical part that controls gate to open and permits the vehicle to enter the building in case of the vehicle's image matches any image in the database, or sends an apology voice message in case of no identical image.**

The system is regarded as an empirical and it was applied on various types of vehicles. Results obtained were accurate and the system is successful for all vehicles used in the system test.

Keywords—image processing; color recognition; patch recognition

I. INTRODUCTION

Due to the technical difficulties and their complexity in some fields, researchers worked hard to find efficient algorithms that help to find alternative solutions appropriate to reach speedy solution of the problems[1]. Especially in using computer in the field of security and protection, the objects recognition by using computer becomes very necessary for making decisions.

Nowadays, there is a big scientific development in the informatics and programming techniques, in which the images took a very important role in various zone such as medicine media, education, design, industry, security etc [2]. The main reason for this development is the wide using of computer in all these zones. The digital image permits to get images instantly without needing for chemical treatments. The image processing is used in many computer vision applications [3].

It notice that the most researches nowadays, especially image processing researches, tend to use in security field, because the digital images become the main dependent component in most electronic devices used in this field. Where it can store a group of images in a computer database and then using images processing programs to recognize the captured pictures from monitoring camera can make the right decision depending on matching the recognized image with the stored database.

Due to increasing need for protecting the materials and the articles in the houses, directorates, etc, this paper used a digital images and microcontroller to control opening and closing a building gate.

In the section (III.A) of this paper, there is a declaration and illustration of the program used to recognize the vehicle and the main block diagram for decision making reinforced by figures of recognition stages. While section (III.B) talks about hardware part used as a model to represent the building gate controlled by Arduino. Section (IV) includes an application example. In addition, the conclusions are in section (V).

II. RELATED WORK

There are many recent ways to control entering building gates e.g. techniques for recognition such as fingerprint, iris print and face features by using remote control hardware or special identity card. The recognition applications on digital images are developed such as in case of fingerprint recognition or recognition of hidden text in the images[4].

In (2005) Yang proposed two algorithms for security system: the first was a system for humane face recognition and the other was a system for motion detection, the system based on Eigen edge technique [5]. In (2006), Kouma introduced real_time representation for security system based on real time principle to recognize the human face [6]. In (2010), Wael designed real time system to control the security of the building by using distribution mobile technique [7].

III. THE PROPOSED SYSTEM

The proposed system consists of a software part and a hardware part as follows:

A. Software

This paper used the patch of each vehicle to permit it to enter the building through the gate. The proposed algorithm for recognition (based on recognizing the patch of the vehicle) consists of two main stages: the first stage recognizes the border of the patch, and the second stage recognizes the patch contents.

Fig. 1 illustrates the general diagram of the algorithm, as presented by [8][9] with changes to be adequate with the patch image.

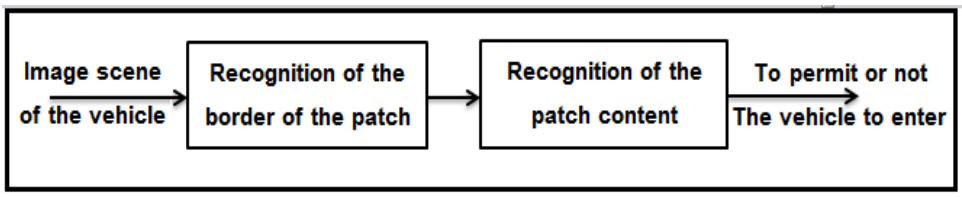


Fig. 1. Block Diagram of Recognition Process.

1) Patch's Border Recognition Steps:

Step1: A captured image enters to the computer and the program read it. Fig. 2 represents a scene for vehicle intended to enter the building.

Fig. 2. Scene for vehicle intended to enter the building.

Step2: Apply preprocessing step which involves the following:

a) Reduinge the image size to reduce the execution time.

b) Converting the color image into red, green, and blue layers.

c) Choosing threshold value for each layer by using trial and error method. The threshold used in this research was (95,130,88) for red, green and blue layers respectively for patch border in green color.

d) Converting the previous image into binary image depending on threshold values to ease data process.

e) Removing the noise outside the border of the patch from the produced image, see Fig. 3 .

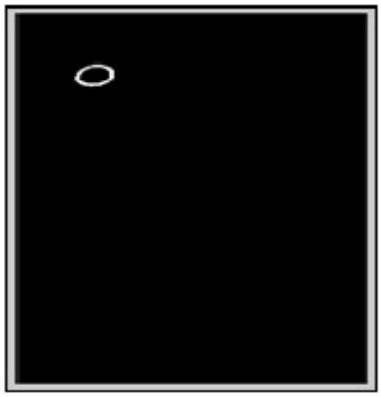


Fig. 3. The image after preprocessing.

Step3: Segmentation, in this step the patch segment is clipped from the original colored image by taking the maximum and minimum values of X and Y coordinates for the image. See Fig. 4.

Step4: Edge detection for the patch image is applied, see Fig. 5. Then, converting the border shape of the patch into a chain code and checking that the shape is circle, or stopping the processing because the patch is illegal and send an apology voice message.

Fig. 4. A clipped image of patch

Fig. 5. The border of patch

2) Recognizing The Inside Contents Of The Patch

This involves the following steps:

Step1: Converting the clipped segment of the colored image into a binary image.

Step2: Removing the border of the patch's image from the image, see Fig. 6.

Step3: Appling edge detection on the objects inside the patch by using chain code. In this step, the recognition depends on a number of objects included the patch's image and the chain code to achieve the objective.

Step4: Matching between the chain code of each object with the database to recognize the vehicle is permissible to enter or not.

Fig. 6. The inside patch contents.

B. Hardware Components

They include of the following parts:

1) Simulated Model

A small model used to simulate the real building. The model 1*1 m2 area represents 10*10 m2 real building i.e. the scale is 1:10 the height of the model is 28.5cm, the slide gate is 20 cm high and 28.5 width. See Fig. 7. Small electronic motor uses (12V) to move the gate by arm joining the motor with gate. Fig. 8 illustrates the mechanical components fixed to the model gate.

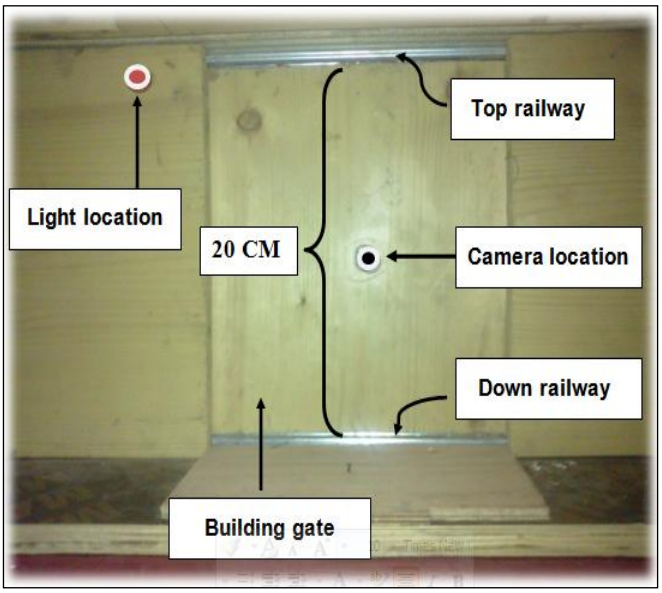


Fig. 7. Front view of the gate.

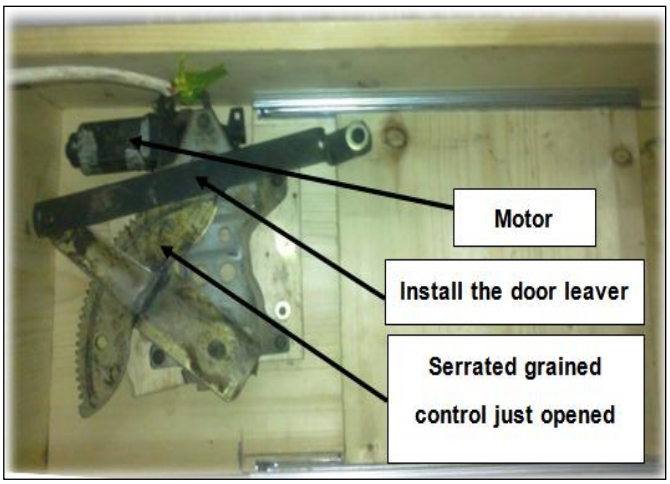


Fig. 8. Inside view of the gate

2) Arduino board

The Arduino board is a small microcontroller board. The Arduino board can write programs and create interface circuits to read switches and other sensors [10]. Arduino UNO is used in this paper.

3) driver

Driver is the join between the motor and the Arduino UNO. It consists of: relay, diode, copper pin, headers 18 pin, board, and capacitor, see Fig. 9.

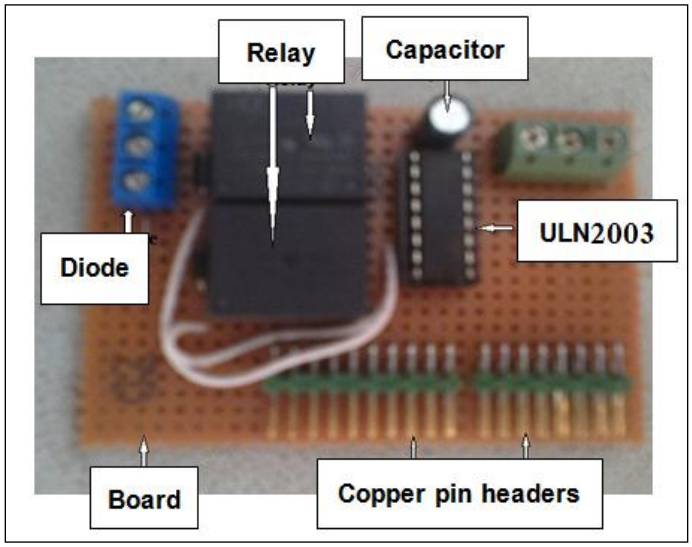


Fig. 9. Drivers Components

4) Electronic Circuit

The electronic circuit used in the model of proposed system to open, close, and stop the building gate is illustrated in Fig.10.

Fig. 11 shows the flowchart of the system procedure.

The system procedure begins with a picture captured by a digital camera for the vehicle that stopped before the building gate intending to enter the building. The camera sends a captured picture to the computer. The computer recognition program runs to recognize the vehicle depending on the color and shape of the patch and matching it with sorted database. In case of the matching occurred, the system will permit the vehicle to enter, by sending a signal (means open) to the electromechanical part (which is fixed to the gate to control opening and closing the gate), then the gate will open.

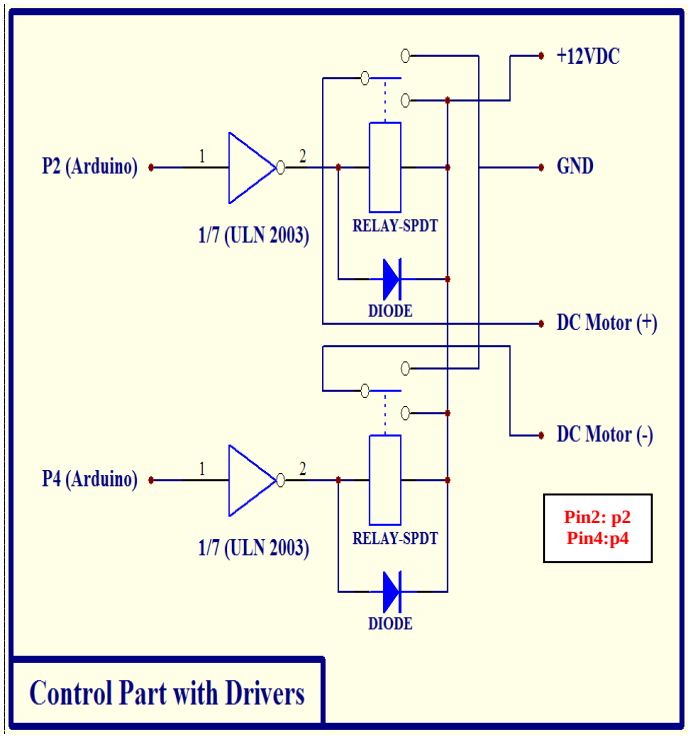


Fig. 10. Electronic circuit Hardware parts.The system procedure

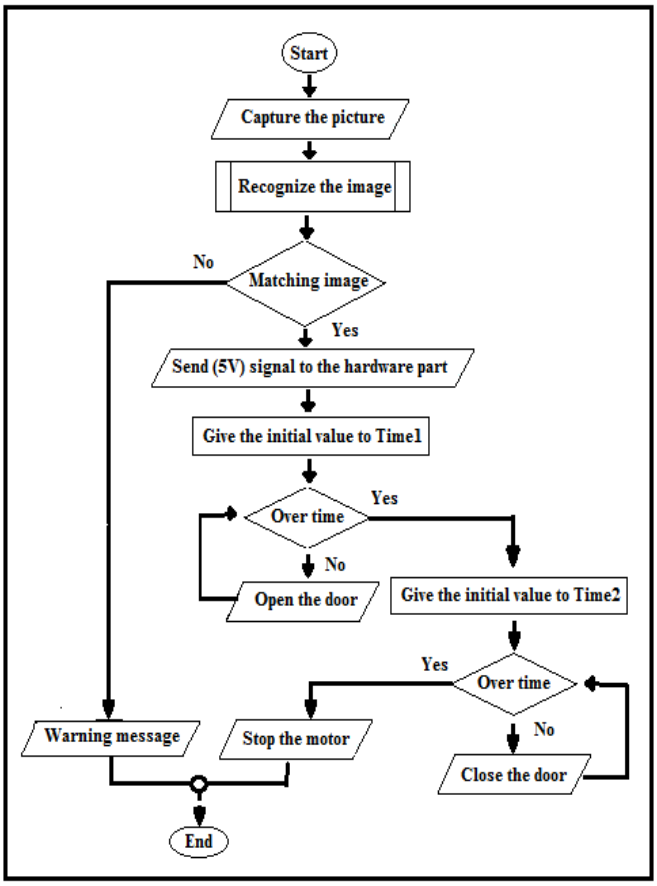


Fig. 11. The flowchart of the system procedure

While in case of the no matching occurs the system will send an apology voice message (means that the vehicle is not permissible to enter), and the system electronic circuit works as follows:

Opening the gate : In case of that the patch of the vehicle matches one in database, the computer will sends a signal(5V) to pin2 in the Arduino board which in its role sends a signal(5V) to the ULN2003 in the driver board, the driver sends(5V) to the fist relay, the first relay sends(12V) to the motor which will work to open the gate after getting the current. See Fig. 10.

Closing the gate: When the time of entering the vehicle to the building through the gate is out, the segment of code that related to closing the gate will be executed in the computer and sends a (5V) signal to the Arduino board which in its role sends a signal to the ULN2003 in the driver board.The ULN2003 sends a signal to the second relay. The second relay sends (12V) to the motor which begins to move to close the gate by switching the polarity of the (12V). See Fig. 10.

Stopping the motor: When opening time is out, the segment of code related to stop the motor will be execute in the computer and sends a (0V) signal to in the Arduino board which in its role sends these signals to the ULN2003 which sends signals to the first and second relays which in their role stop the motor current and the motor will stop, see Fig. 10.

IV. EXAMPLE

The proposed algorithm of recognition applied on various types of real vehicles and the sample of dummy vehicle used to apply the simulated model to open and close the gate. See Fig. 12-14, scene for the gate to permit the vehicle to enter into the building.

V. CONCLUSIONS

Automatic controlling and domination on the building gates are regarded as an important matter nowadays whether for vehicles or persons, that's due to the difficulty of the traditional systems because they are not accurate and required too much time, effort and cost. Even recent systems need for fixing special sensors in the building or needs mobile network connection to work. While the hardware and software in this proposed system are simple, inexpensive and available.

The matter of using computer to control and dominate the gate systems became very urgent to arrange, manage and control the security of the buildings. Thus, the proposed system is a computer system based on digital image processing and pattern recognition techniques. This system invests these techniques in the security field through recognizing the patch of the vehicle, which intends to enter the building. The results show 95% recognition percentage for the used samples of the real passenger's cars and the simulated dummy vehicles.

To increase security of system, it is recommended for future works to add recognition of license plate of vehicle in addition to patch recognition, which will increase the system authentication and reliability that mean increasing the efficiency of security work.

ACKNOWLEDGMENT

Our thanks to the journal staff for their patience and cooperation.

Fig. 12. Captured picture

Fig. 13. Scene before opening the gate

Fig. 14. Scene after opening the gate from inside the building model

REFERENCES

[1] E. Abbas, H. Farhan, "Face Recognition using DWT with HMM" , Eng. & Tech. Journal, vol.30, No.1, 2012, pp 142-154.

[2] M. Ahmed, A. Barzan, "Automatic Features Recognition for Symmetrical Shapes", Eng. & Tech. Journal, vol.30, No.12, 2012, pp 2027-2043.

[3] L. Crowley, H. Daniel and R. Emonet, " Autonomic Computer vision system", In International Conference on computer vision system, Icvs, vol. 17, 2007.

[4] M. Kadum, "Detection of Hidden Object In Speech Based on Fast Fourier Transform Algorithm", Eng. & Tech. Journal, vol.29, No.2, 2011, pp 375- 385.

[5] J. Yang, " Algorithms For An Integrated Vision-Based Home Security System", Multimedia University, Malaysia , 2005.

[6] J. Kouma, "Intelligent Home Security System", Applied Physics and Electronics, SE-901 87, Umea university, Sweden, 2006.

[7] W. Wael, "Design And Implementation Real Time System For Building Security By Using Distributed Mobile Technique Method", MSc. thesis, Computer Sciences and Mathematics, University of Mosul, 2010.

[8] R. Jasim, S. Khaleel, " Traffic Sign Recognition using Chain Code", AL-Rafiden Journal of Computer Sciences and Mathematics, No.9, vol.2, 2012.

[9] R. Jasim, S. Khaleel, "Recognition using ChainCode", Iraqi Journal of Statistical Sciences, No. 20, vol.11, 2011, pp 691-713.

[10] M. Banzi , Getting Started with Arduino, 2nd Edition,Copyright, ©Massimo Banzi. All rights reservedPrinted in the U.S.A, 2011.

An Algorithm for Summarization of Paragraph Up to One Third with the Help of Cue Words Comparison

Noopur Srivastava
Department of Computer Science
Shri RamSwaroop Memorial University
Barabank-225003, UP, India

Bineet Kumar Gupta
Department of Computer Science
Shri RamSwaroop Memorial University
Barabanki-225003, UP, India

***Abstract*—In the fast growing information era utility of technology are more precise than completing the assignment manually. The digital information technology creates a knowledge-based society with high-tech global economy which spreads over and influence the corporate and service sector to operate in more efficient and convenient way. Here an attempt was made on Extract Technology based on research. In this technology data could be refined and sourced with certainty and relevance. The application of artificial intelligence matched with the theories of machine learning would prove to be very effective. Sometime summarization of paragraph required rather than page or pages. So, Auto Summarization Model is an agnostic content summarization technology that automatically parses news, information, documents and many more into relevant and contextually accurate abbreviated summaries. This is a concept to convert a whole paragraph into one third. The Auto summarization technology reads a document, much better way than manually prepared, where, keywords and key phrases accurately weighted as they are found in the document, text or web page.**

***Keyword*—Data Mining; Data Warehouse; Artificial Intelligence**

I. Introduction

In present paper an attempt is made to introduce the essential research area of the data mining algorithm implementation and suggest important line on the basis of 'cue words'. Where, Auto Summarization uses a patented set of core algorithms to extract keywords and key phrases from any text-based document [24, 7]. In essence a machine learned method for reading or summarizing any text written in an electronic text format [27]. On the basis of cue word analysis one can select important lines from one paragraph [10, 12].

Auto Summarization is exceptionally good at content summarization incorporating its patented technology to summarize text, e-mail and html content into weighted lists of structuring of Web data and solve the problem about effectiveness in retrieval accordingly [7, 10].

Auto Summarization is exceptionally good in summarizing the text i.e. important part of the paragraph automatically without changing meaning of the paragraph. It will summarize text, emails, news, speech, etc into weighted lists of keywords and key phrases extracting the primary contextual sentence highlight of how the keyword / key phrase has been used [15, 29]. Uniquely positioned for web services, Auto Summarization is immediately capable of consuming documents of any length and subject matter, distilling the precise, contextual meaning of the content into keyword and key phrase summary formats. Extractor's unique patented technology delivers precise content summaries of any subject domain without retraining and without human intervention [29].

Auto Summarization is extremely effective for objectively distilling a document down to its key concepts providing users with highly focused keywords and key phrases including contextual examples of exactly how the keywords / key phrases have been used in the document - an extraction [20].

In contrast, a synopsis is a subjective interpretation of a document providing the end user with a high level statement of what that person believes the author intended the reader to comprehend. Such as an abstract Subjective is an important note - to date automated processes for generating a synopsis has not been perfected - and why they remain a human based process [12, 15].

II. The Aspects of Auto Summarization

Not just information but contextually accurate, relevant information is a critical tool for the success of business today. Rather than working through traditional, time consuming, and interactive search engine processes, incorporating Auto Summarization into Enterprise systems, empowers corporate information with relevant and meaningful representations meeting the needs of today's social workforce [7, 12]. Simple demonstration of auto summarization shows the summarization of notes from paragraph as shown in Fig. 1.

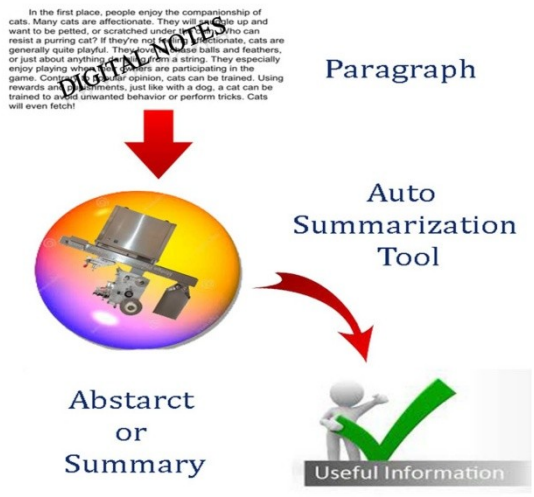


Fig. 1 Process of Auto Summarization

A. Artificial Intelligence Approach

It is based on fuzzy logic in which artificial intelligence tool also works. This technology is based on fuzzy logic and artificial intelligence [20, 9, 16]. Here artificial intelligence help to maintain the discipline in calculation of percentage through which one signify the actual subject percentage, while, fuzzy logic is to calculate the percentage of frequency.

With combination of both approaches that are fuzzy logic and artificial intelligence a unique formula prepared for the word, which help the user to identify the words of paragraph more accurately without wasting the time and money [20, 9].

In present research, summarization of cue words and maintaining frequency of each word and sentence by comparing it with cue word collections and stop word collections belongs under Artificial Intelligence approach.

B. Data Warehouse Approach

It is a central repository of data which is created by integrating data from one or more disparate sources [20]. Data warehouses store current as well as historical data and are used for creating trending reports for senior management reporting such as annual and quarterly comparisons [14, 25].

By this approach, proposed model collect the cue words means from the high frequency word store in respective subject at the time of adding paragraph of new subject. This process is going on for future reference when these words will match by auto arranger for final distribution of single note into different subject's collection [20]. Fig. 2 shows the collection or summarization of cue words and stop words, which maintain their frequency after comparison with each word, worked as Data Warehouse.

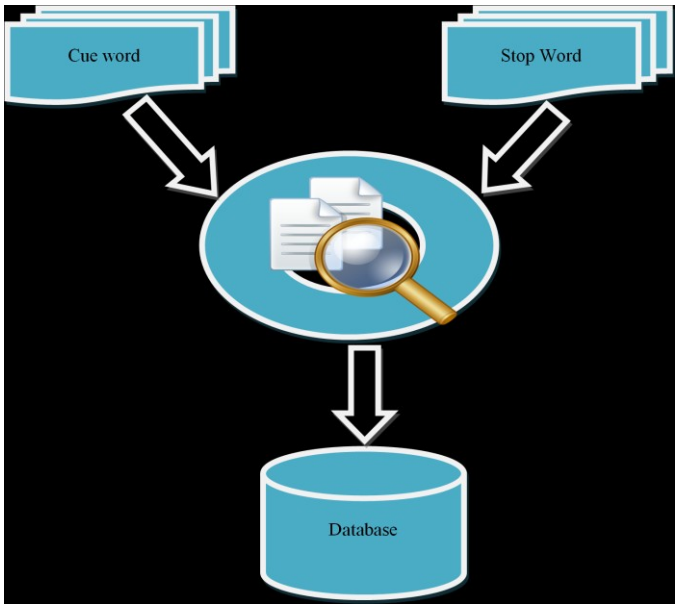


Fig. 2 Frequency of words in database

C. Data Mining Approach

Data mining is very helpful for extracting word or sentence after comparing each word from data warehouse. In present approach, comparison of cue words with new word in paragraph and make frequency table, is worked as Data Mining Fig. 3 [2, 28, 13]. Here the role of Data mining is to match the frequency of cue words to find out how many times the paragraph word is coming in the compared text. By this approach this model works more efficiently [5].

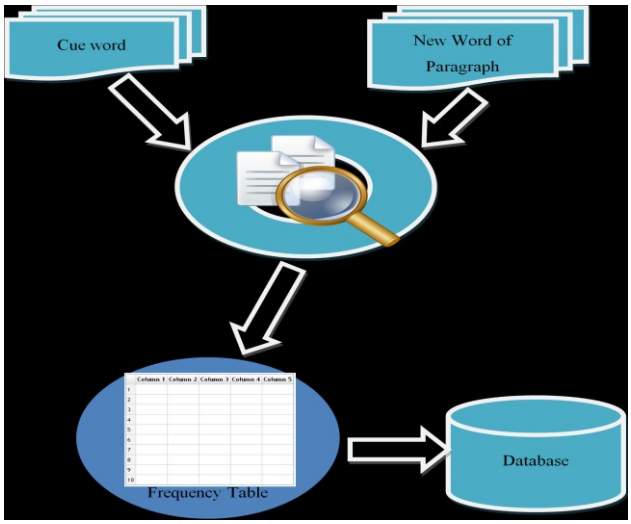


Fig. 3 Process of storing frequency table

D. Natural Processing Language (NLP)

A branch of Artificial Intelligence with analyzing understanding and generating the languages, which is used naturally in order to interface with computers in both written and spoken contexts using natural human languages instead of computer languages [17]. One of the challenges inherent in natural language processing is teaching computers to understand the way one can learn and use language. In the course of human communication, the meaning of the sentence depends on both the context in which it was communicated and each person understands the ambiguity in human languages [17, 4]. This sentence poses problems for software that must first be programmed to understand context and linguistic structures [6].

III. Proposed Technique

The proposed technique is based on NLP (Natural Language Processing) known as Gradual NLP algorithm. Automatic document summarization is an important research area in natural language processing (NLP). The technology of automatic document summarization is developing and may provide a solution to the information overload problem [17].

The process of summarization can be decomposed into three phases: analysis, transformation, and synthesis. The analysis phase analyzes the input text and selects a few salient features. The transformation phase transforms the results of the analysis into a summary representation. Finally, the synthesis phase takes the summary representation, and produces an appropriate summary corresponding to users' needs [2, 7]. In the overall process, compression rate, which is defined as the ratio between the length of the summary and that of the original, is an important factor that influences the quality of the summary. As the compression rate decreases, the summary will be more concise; however, more information is lost. While the compression rate increases, the summary will be larger; relatively, more insignificant information is contained. In fact, when the compression rate is 5–30%, the quality of the summary is acceptable as shown in Fig. 4.

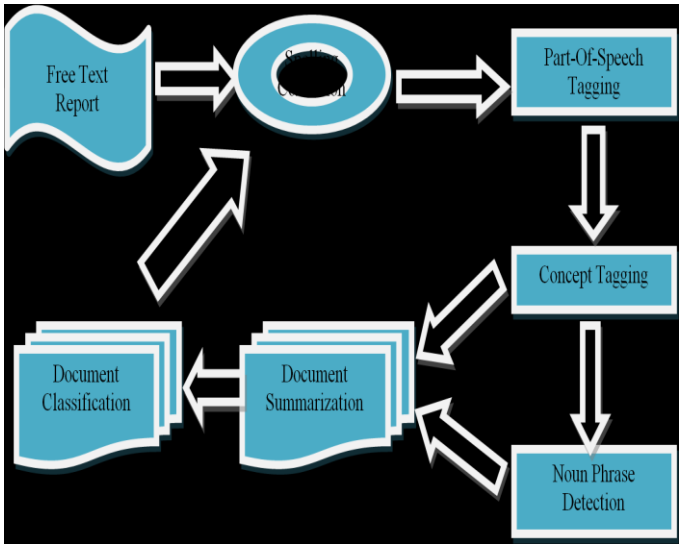


Fig. 4 Process NLP (Natural Language Processing)

A. Algorithm for Summarization

1) Select table word frequency (rs4), world list (rs5), stop word (rs3).

- Not stop word then frequency=frequency+1.
- Find total frequency and no. of words.
- Find average= total frequency/no. of words.
- If frequency used>average then store in rs5.

2) Open sentence (rs) and store each sentence with frequency 0.

3) Open cue word (rs6), basic id (rs7).

- If sentence contain number then freq= freq+1.
- If sentence contain cue word then freq=freq+1 for each word.
- If sentence contain wordlist then frequency=freq+1 for each word.
- If sentence not contain basic id then freq=freq+1 for each word.

4) Find total words in paragraph.

5) Find average length=total no. of words/no. in sentence

6) Find final score=score (average/length of sentence).*

7) Find total frequency=add all final score of the entire sentence.

8) Find average frequency= total frequency/no. of sentence.

9) If sentence frequency>avg frequency

- Extract that sentence.

10) Average frequency=total final score frequency/no. of sentence

11) Abstract summarized

B. Algorithm of summarization

Step 1: Implement Simple NLP algorithm in which first create 3 tables in the database with the names word frequency, word list, stop word, respectively. Then add a new paragraph in the summarization model. The words in the paragraph will be matched with the words available in the stop word. If the matched words are not stop word then the frequency will be increased by one. Then repeat the process for all the words in the paragraph and find the total frequency and count the number of words. In the last, find the average by using the formula given below:

Average = total frequency/no. of words.

Step 2: If the frequency used is greater than average then store and write the words in the word list table. Now, open the sentence from the paragraph that are added before and store in the newly created table sentence with zero frequency.

Now create a basic cue word table and store useful words. In "cue words" store the important words like noun, adjective and adverb. While, phrasing the sentence if found a number then increase the frequency by 1.

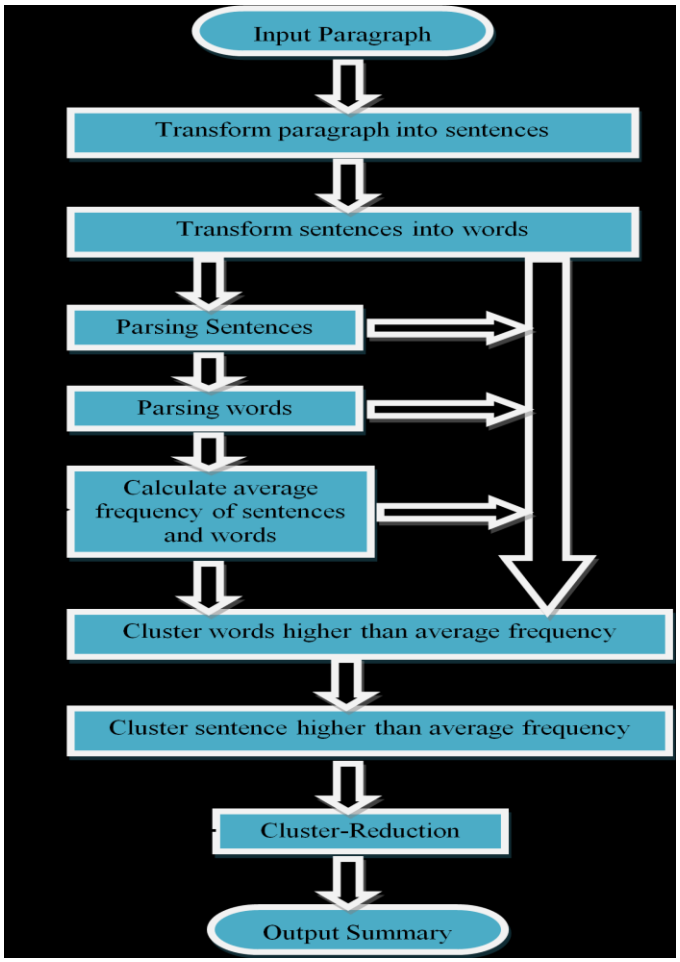


Fig. 5 Progressive representation of algorithm

If cue word is present in the sentence which is being phrased then increase the frequency by 1 of each word. If sentence contains wordlist then increase the frequency by 1 for each word. If basic is not found in the sentence, while phrasing then increases the frequency by 1.

Find the total number of words in the paragraph and average length of the sentence by using the formula below:

Average length = total no. of words/no. in sentence

Step 3: Find the Final score by using given formula:

Final score = score* (average/length of sentence)

Step 4: After receiving the final score of the sentence find the total frequency by using the given formula:

Total frequency = add all final score of the entire sentence

Step 5: Get the average frequency by using the given formula below:

Average frequency = total frequency/no. of sentence

Step 6: Compare the frequency of the sentence and average frequency. If frequency of the sentence is greater than average frequency, then extract that sentence.

Calculate the Average frequency again by using following formula:

Average frequency = total final score frequency/no. of sentence

Step 7: Check Final score if it is greater than average frequency then extract those sentences and show the final abstracted or summarized paragraph as in result form in front of user.

IV. Conclusion

In present work it is clear that not only information but contextually accurate, relevant information is a critical tool for the success of business today. Being able to source relevant information in context of the subject matter gives organizations an ultimate competitive advantage rather than working through traditional, time consuming, and iterative search engine processes, thus, incorporating extractor into enterprise systems. This empowers corporate information with relevant and meaningful representations, meeting the needs of today's social workforce.

In this context, paragraph break down into one third where one third part is abstract or summary for that whole paragraph. This helps to generate or convert whole paragraph into one third with highly important part as in the form of extract without violating the meaning of paragraph. Sometimes, it is required to extract a paragraph rather than whole page. So, this model is very effective and efficient to extract a paragraph. In this research we summarize only a paragraph but in future aspect we can summarize whole document into one-third. It is very much helpful for those students or people who cannot tell or express their knowledge. So, with the help of this technique they express their extract content of particular subject Auto Summarization is responsible for summarize the textual information approximately one-third valuable information for further decision support system or management information system. It can also be used for fetching important headline from the news. So, this model is very effective in retrieving the correct information and reduces the time complexity of the user.

Acknowledgment

This study is a part of my M.Tech final year dissertation conducted at the faculty of computer Science and Technology,

Shri Ramswroop Memorial University, Lucknow Deva road,Uttar Pradesh,India.

References

[1] A. K. Choudhary, J. A. Harding, and M. K. Tiwari, "Data Mining in Manufacturing: A Review Based on the Kind of Knowledge", Journal of Intelligent Manufacturing 20(5), pp. 501-521, 2009.

[2] B. Gupta and md. Hussain,"Algorithm to Evaluate the Rank of Research Papers Using Citation Graph in CiiT", International Journal of Data Mining Knowledge Engineering, ISSN NO 0974, Paper ID 3033, 2012.

[3] B. Park and H. Kargupta,"Distributed Data Mining: Algorithms, Systems, and Applications", 2002.

[4] B. Z. Manaris,"Natural Language Processing: A Human-Computer Interaction Perspective", University of Southwestern Louisiana, Academic Press, New York, vol. 47, pp. 1-66, 1998.

[5] C. Yu, K. L. Liu, W. Meng, Z. Wu, and N. Rishe, "A Methodology to Retrieve Text Documents from Multiple Databases. Knowledge and Data Engineering", IEEE Transactions, 14(6), pp. 1347-1361, 2002.

[6] D. Lewis, "Natural language processing for information retrieval, ACM, 39 (1), pp. 92-101, 1996.

[7] D. Das and A. F. T. Martins, "A survey on Automatic Text Summarization," Language Technologies Institute, Carnegie Mellon University, 2007.

[8] D. P. Ballou and G. K. Tayi, "Enhancing Data Quality in Data Warehouse Environments",Communications of the ACM, 42(1), 73-78, 2009.

[9] E. Charniak and D. McDermott, "Introduction to artificial intelligence", Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1985.

[10] Eduard Hovy and Chin Yew Lin, "Automated text summarization in SUMMARIST", MIT Press, pp. 81–94, 1999.

[11] G. Kundu, "Adapting Text Instead of the Model: An Open Domain Approach", Conference on Computational Natural Language Learning, 2011.

[12] I. Mani, "Automatic Summarization", John Benjamin's Publishing Co. pp.1-22, 2001.

[13] I. Taksa,"Research and Trends in Data Mining Technologies and Applications". Information Retrieval, 11(2), pp.165-167, 2008.

[14] J. M. Pe´rez, R. Berlanga, Marı´a Jose´ Aramburu, and Torben Bach Pedersen," Integrating Data Warehouses with Web Data: A Survey", Ieee Transactions On Knowledge And Data Engineering, vol. 20, no. 7, 2008.

[15] K. Jezek, and J. Steinberger, "Automatic Text Summarization (the state of the art 2007 and new challenges)", Znalosti , pp. 1-12, 2008.

[16] L. Suanmali, N. Salim and M. S. Binwahlan, "Fuzzy Logic Based Method for Improving Text Summarization", International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

[17] Md. M. Haque, S. Pervin, and Z. Begum, "Literature Review of Automatic Single Document Text Summarization Using NLP", International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 3 No. 3, pp. 857-865, 2013.

[18] M. J. Pe´rez, R. Berlanga, Marı´a Jose´ Aramburu, and B. T. Pedersen,"Integrating Data Warehouses with Web Data: A Survey", IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 7, 2008.

[19] M. Lease,"Natural Language Processing for Information Retrieval: the time is ripe (again)", ACM, 2007.

[20] N. Srivastava, B. K. Gupta and N. K. Tiwari, "An Approach to Develop a Framework to Enhance the Performance of Digital Notes Based on Auto Arranger", International Journal of Engineering Research and Development, Vol. 10, Issue 4, pp. 53-57, 2014.

[21] R. Arora, P. Pahwa and S. Bansal, "Alliance Rules of Data Warehouse Cleansing. IEEE, International Conference on Signal Processing Systems, Singapore", 743–747, 2009.

[22] R. Mihalcea, H. Liu, and H. Lieberman, "Natural processing language for natural processing programming", Springer-Verlag Berlin Heidelberg , pp. 319–330, 2006

[23] R. Studer, V.R. Benjamins and D. Fensel," Knowledge Engineering: Principles and Methods. Data & Knowledge Engineering", 25(1-2), 161-197, 1998.

[24] S. Gholamrezazadeh, Mohsen Amini Salehi, and Bahareh Gholamzadeh , "A Comprehensive Survey on Text Summarization Systems" , IEEE, 2009.

[25] S. Chaudhuri, and U. Dayal," An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD record, 26(1), pp. 65-74, 1997.

[26] S. E. Madnick, Y. W. Lee, R. Y. Wang and H. Zhu, "Overview and Framework for Data and Information Quality Research",ACM Journal of Data and Information Quality, 1(1), 2, 2009.

[27] S. Suneetha, "Automatic Text Summarization: The Current State of the art," International Journal of Science and Advanced Technology (ISSN 2221-8386), vol. 1, no. 9, pp. 283-293, 2011.

[28] T. L. Daniel, "Discovering Knowledge in Data. An Introduction to Data Mining", John Wiley & Sons, Inc., 0-471-66657-2, 2005.

[29] V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques", Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, 2010.

[30] X. Li, "A Supervised Clustering and Classification Algorithm for Mining Data with Mixed Variables". IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 6(2), 376-406, 2006.

A Secure Electronic Transaction Payment Protocol Design and Implementation

Houssam El Ismaili[1], Hanane Houmani[2], Hicham Madroumi[3]
Architecture of Systems Team - ENSEM, Hassan II University, 8118, Casablanca – Morocco

***Abstract*—Electronic payment is the very important step of the electronic business system, and its security must be ensured. SSL/TLS and SET are two widely discussed means of securing online credit card payments. Because of implementation issues, SET has not really been adopted by e-commerce participants, whereas, despite the fact that it does not address all security issues, SSL/TLS is commonly used for Internet e-commerce security. The three-domain (3D) security schemes, including 3-D Secure and 3D SET have recently been proposed as ways of improving ecommerce transaction security. Based on our research about SSL, SET, 3D security schemes and the requirements of electronic payment, we designed a secure and efficient E-Payment protocol. The new protocol offers an extra layer of protection for cardholders and merchants. Customers are asked to enter an additional password after checkout completion to verify they are truly the cardholder; the authentication is done directly between the cardholder and card issuer using the issuer security certificate and without involving the third party (Visa, MasterCard).**

Keywords*—*E-commerce; Secure Socket Layer (SSL); Secure Electronic Transaction (SET); 3D-Secure

I. Introduction

Electronic commerce or e-commerce provides participants, including consumers and merchants, with a number of benefits, such as convenience and time savings. E-commerce transactions can be categorized into business to business (B2B), business to consumer (B2C), consumer to consumer (C2C), and public/private sectors to government [1]; we focus on B2C transactions in this paper.

In B2C transactions, the credit card is the most widely used method of payment for Internet ecommerce transactions. According to an Internet shopping habits survey conducted by Survey.Net (http://www.survey.net), 36% of Internet users purchase goods by transmitting their credit card number via a secure form; the percentages for other payment methods are significantly lower. Given that the debit/credit card is the primary means for consumers to purchase products or services online, the possible compromise of credit card numbers is a serious threat to the consumer. The E-payment system brings users with higher efficiency, credibility and speeding-up transactions settlement, which reduce the pay risks caused by time lags in handling the bills. However, it also comes with new risks, i.e. security problem of transactions.

The research reported here builds on the electronic payment security; we study the security of e-commerce protocols and we propose a new efficient protocol to ensure a high security for electronic payment transactions.

The objective of our protocol is to provide issuers with the ability to authenticate cardholders during an online purchase without involving the third party VISA or MasterCard. We define a new transaction flow involving cardholder, merchant, payment gateway and card issuer, and allowed parties to identify themselves to each other and exchange information securely using digital certificate. For some implementation reasons, the cardholder is not requested to have his digital certificate, he use the password code to be authenticated by the card issuer.

II. Security Requirements Of E-Payment

It goes as follows [2]:

A. Information confidentiality

All information during the transactions has the request of being kept confidential. For instance, account number and user name may be embezzled by others who have access to them; business opportunity may be lost if order and payment information of your customer's are obtained by competitors. Thus, encryption is required in the E-C information transmission.

B. Data integrity

E-C should provide medium to identify data integration, ensuring the Web data do not be altered in transmission.

C. Authentication of participants

The parts involved may have never met each other. So to make the transaction successful, the first step is to identify the two parts, which is the essential prerequisite of transactions.

D. Non-repudiation

The transaction must have such services that enable one party to prevent another party denying having taken a particular action, e.g. sending order/payment information, confirmation of order/payment. Both consumer and merchant also require this service.

E. End-user implementation Requirements

We focus here on the major barriers causing implementation failures in SET and other protocols including usability, flexibility, affordability, speed of transaction, and interoperability.

- Usability – The system must be easy to implement, including installation. The consumer requires the card issuer and merchant to provide a secure system that is not complex, while the merchant requires the acquirer and security software developers to provide a simple application that meets the security requirements.
- Flexibility – The system must allow e-commerce consumers to order products or services from any location, and not just from one PC. Here, the consumer is the entity requiring theflexibility service, while the merchant is the entity providing the service.
- Affordability – The costs of implementing and using the system must be affordable for consumers and merchants, since these end-users are unlikely to be prepared to pay significantly extra to participate in Internet e-commerce transactions. For example, consumers are not willing to pay for a digital certificate in order to conduct e-commerce transactions although it is required in some e-payment scheme such a SET. Merchants will also not wish to invest significantly in engineering e-payment infrastructure.
- Reliability – The system must be reliable since it is used for the transmission and manipulation of sensitive information.
- Availability – The system must be available when needed.
- Speed of transaction – The transaction speed must be acceptable for e-commerce end-users.
- Interoperability – The system must be interoperable between different computing platforms, web browsers and server software packages in order to enable its use by the widest possible spectrum of e-commerce consumers and merchants.

E-C secure protocols are the widely recognizes logical operating standards for secure completion of information exchange, as well as the critical technique to ensure the confidentiality, integrity, authentication and non-repudiation of online transactions. Their completion serves as a key to provide online security. Internet E-C security protocol is the central research areas in E-C as the endeavors to promote the development of E-C, and guarantee its security. The prevalent protocols are Security Socket Layer (SSL), Secure Electronic Transactions (SET) and 3D-Secure.

These protocols allow using cryptography to send confidential information on the Internet without being readable to malicious individuals. However, it turned out that these protocols are not as secure as we thought they would be. Indeed, several errors were discovered in cryptographic protocols after some years of use. The consequences that can generate vulnerability in a cryptographic protocol can be costly and irreversible for companies and individuals.

In this paper, we consider how E-commerce security requirements are fulfilled by our new protocol based on payment gateway and digital signature.

III. Related Work

There have been many studies of E-commerce security. Security in E-commerce was described in the paper written by Dhilon [3] who introduce the stages to be provided for online purchase, the approach is based on encryption and compression for making information unreadable. However, E-commerce security has become a consistent and growing problem as new internet technologies and application are developed; it needs new architecture to adapt to many changes. Al-SLamy [4] described the role of Pretty Good Privacy (PGP) to provide confidentiality, authentication, compression and segmentation services for E-commerce security. Byung Lee [5] introduced The Advanced Secure electronic payment (ASEP) which use ECC (Elliptic Curve Cryptosystem), SHA (Secure Hash Algorithm) and 3BC (Block Byte Bit Cipher) instead of RSA and DES in order to improve the strength of encryption and the speed of processing. Xuan Zhang [6] designed and implemented a new payment process to guarantee goods atomicity, certified delivery atomicity and protects sensitive information of cardholder and merchant.

Secure Sockets Layer (SSL) is a commonly used protocol used to encrypt messages between web browsers and web servers [7]. It encrypts the datagrams of the Transport Layer protocols. SSL is also widely used by merchants to protect the consumer's information during transmission, such as credit card numbers and other sensitive information. SSL is used to provide security and data integrity over the Internet and thus plays an important role. SSL has now become part of Transport Layer Security (TLS), which is an overall security protocol. One of the major problems of SSL is that the merchant can store the sensitive information of the cardholder, and the protocol does not prevent the non repudiation because the client authentication is optional.

SET (Secure Electronic Transaction) come to resolve the weakness of SSL in authentication and protection of sensitive information, SET ensures payment integrity, confidentiality and authentication of merchants and cardholders [8]. But SET is characterized by the complexity and the cost supported by the merchant (compared to the alternative proposed by SSL) because of the logistics of certificates distributing and client software installation, also it's difficult to manage non-repudiation. To deal with it, VISA introduce 3D-secure [9], this protocol is based on the introduction of additional control when buying online in addition to the classic sensitive cardholder data. The customer validates the payment in new window by entering a secret data agreed with its own bank (password, date of birth, code received by SMS or generated by a personal drive).

IV. Secure Electronic Payment Protocol Design

Our main idea is to design a secure and efficient protocol to protect online payment transactions against the fraud without involving the third party, our protocol respond to the requirements of e-payment security: confidentiality, integrity, authentication and non-repudiation.

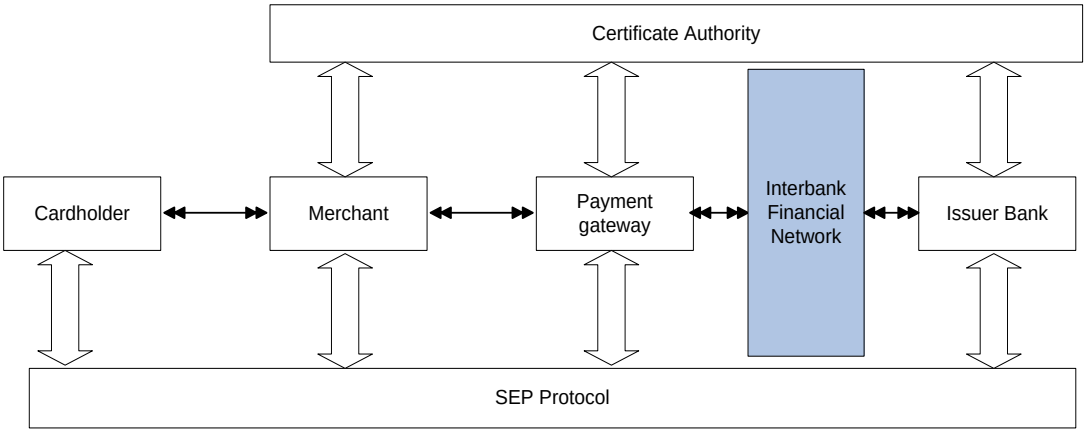


Fig. 1. Entities in SEP protocol

Our Secure Electronic Payment (SEP) protocol avoids the complexities relating to the implementation unlike SET and 3D-secure, integration and utilization are also easier than before.

For the convenience of written expression, we use the following notational conventions in this paper.

C : Cardholder
M : Merchant
PG : Payment Gateway
IB : Issuer Bank or Cardholder Bank
CA: Certificate authority
Vshop : Virtual Shopping Site
PAN: Card Number
CVV2: Card Verification Value or Crypto (three digits)
ExD: Expiry date of the card
OI: Order Information
PI: Payment Instructions
OIMD: OI Message Digest
PIMD: PI Message Digest
POMD: Payment Order Message Digest
K: Symmetric key generated randomly
Kum: Public key of merchant
Kupg: Public key of payment gateway
Kuis: Public key of issuer bank
Krm: Private key of merchant
Krpg: Private key of payment gateway
Kris: Private key of issuer bank
S: Sign
E: Encrypt
D: Decrypt
V: Verify signature
H: Hash
||: Concatenation
#: Disconnect
Eq: Equal

 : Certificate

Our SEP protocol includes the following entities (see figure 1).

The standard description of SEP is illustrated in figure 2.

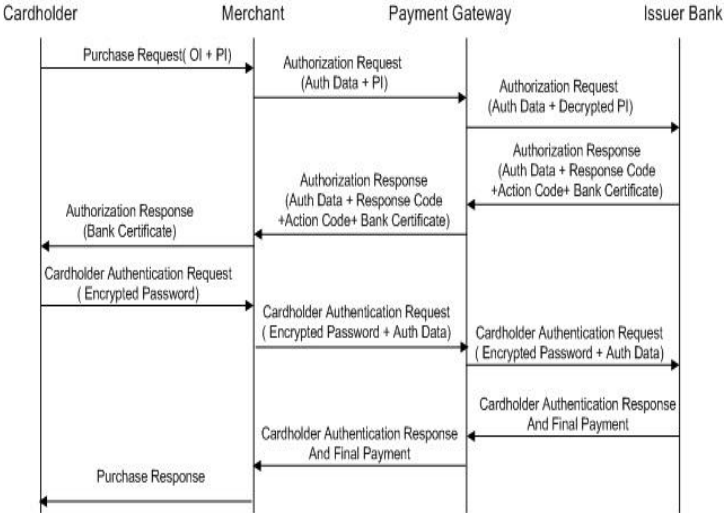


Fig. 2. Description of SEP protocol

A. *Registration process*

Merchant, payment gateway and issuer bank should register and obtain certificates from certificate authority (CA) before they involve in the SEP transaction.

Cardholder should register and obtain a password from his issuer bank (IB) before he involve in the SEP transaction. *Purchase Request*

Cardholder browses for items, select items to be purchased from the Vshop and get an order which contain the list of items to be purchased. Before stating purchase the cardholder and the merchant agree upon the order description amount. The cardholder then sends to the merchant his local ID and a fresh random challenge. The purpose of this is to give the cardholder with the merchant's signature certificate and the payment encryption certificate.

1) Cardholder generates OI, encrypted PI and dual signature. The dual signature is encrypted under a symmetric key generated randomly for the encryption; the cardholder is not requested to have his own certificate (see figure 3).

2) Cardholder prepares the purchase request and sends it to the merchant (see figure 4).

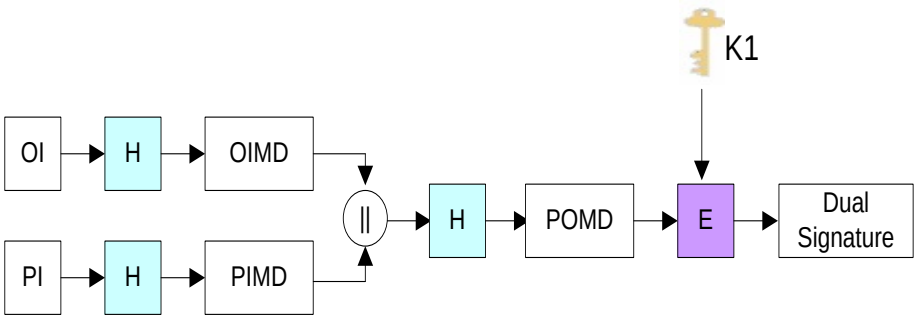


Fig. 3. Dual signature

3) The merchant extract the symmetric key, process the OI and transmit the encrypted PI to the payment gateway. (see figure 5)

B. Authorization Request

1) Merchant signs and sends authorization request to payment gateway, he sends the symmetric key K1 used for dual signature, the encrypted PI. The authorization request is encrypted under a symmetric key generated randomly. The payment gateway verifies the dual signature and gets PI. (see figure 5)

2) The payment gateway transmits the authorization with PI to the issuer bank through a secure and private interbank financial network. (See figure 6)

C. Authorization Response

1) The issuer bank verifies PI, verifies authorization request and run some issuer controls to check if the cardholder is allowed to make this transaction.

2) The issuer sends an authorization response and issuer bank certificate to the payment gateway through the secure interbank financial network (see figure 7). The authorization response contains the response code and the action code. The response code indicates if the authorization request is approved or no, the action code indicates if the cardholder is asked to be authenticated using his password. The purpose of this step is to give the cardholder with issuer bank encryption certificate.

3) The payment gateway signs and sends the authorization response and issuer certificate to the merchant. (see figure 8). The merchant check the action code, if the action code equals to 'Y' witch mean that the cardholder should be authenticated then, the merchant sends an authentication request to the cardholder containing the issuer certificate and some authorization data(see figure 9).

D. Cardholder Authentication Request

1) The cardholder verifies the issuer certificates and sends his personal password encrypted under the symmetric key. (see figure 10)

2) The merchant verifies the authorization data and transmit the encrypted password to the payment gateway. (See figure 11)

3) The payment gateway verifies authorization data, the hash of the encrypted password and transmits the encrypted password to the issuer for verification. The issuer decrypts the encrypted password and checks if is it the correct one for this cardholder. (see figure 12)

E. Cardholder Authentication Response and final payment

1) The issuer bank decrypts and verifies the password code, ensures the consistency between the authorization request and cardholder authentication request, debits the cardholder account and sends a payment response to the payment gateway. (see figure 13)

2) Finally the payment gateway transmits the payment response to the merchant (see figure 14). Merchant verifies the response and ships the good to the cardholder.

F. SEP and Information Confidentiality

For each step of transmission a symmetric key is generated randomly to encrypt electronic payment data. The encryption prevents the illegal information access and information stealing in transmission.

G. SEP and Authentication

- ✓ Cardholder authenticates merchant and issuer bank
- ✓ Merchant authenticates payment gateway and issuer bank
- ✓ Payment gateway authenticates merchant and issuer bank
- ✓ Issuer bank authenticates cardholder using the password code.

H. SEP and Information Integrity

Data integrity is ensured by using MACs (Message Authentication Code) based on hash functions MD5 (16 bytes) or SHA-1 (20 bytes). The MAC is sent for every message transmitted between ecommerce actors.

I. SEP and Non-Repudiation

The non-repudiation property is guarantee by using the password code during the cardholder authentication request.

The issuer bank authenticates the card and the cardholder, so the cardholder cannot deny the fact that he had sent information afterwards.

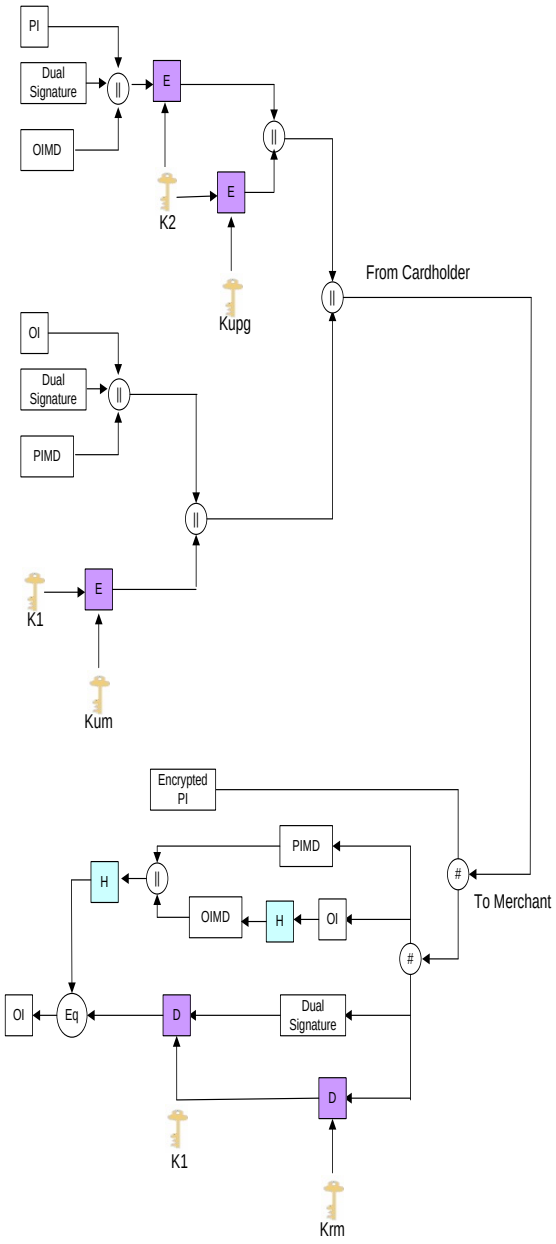


Fig. 4. Purchase request from cardholder to merchant

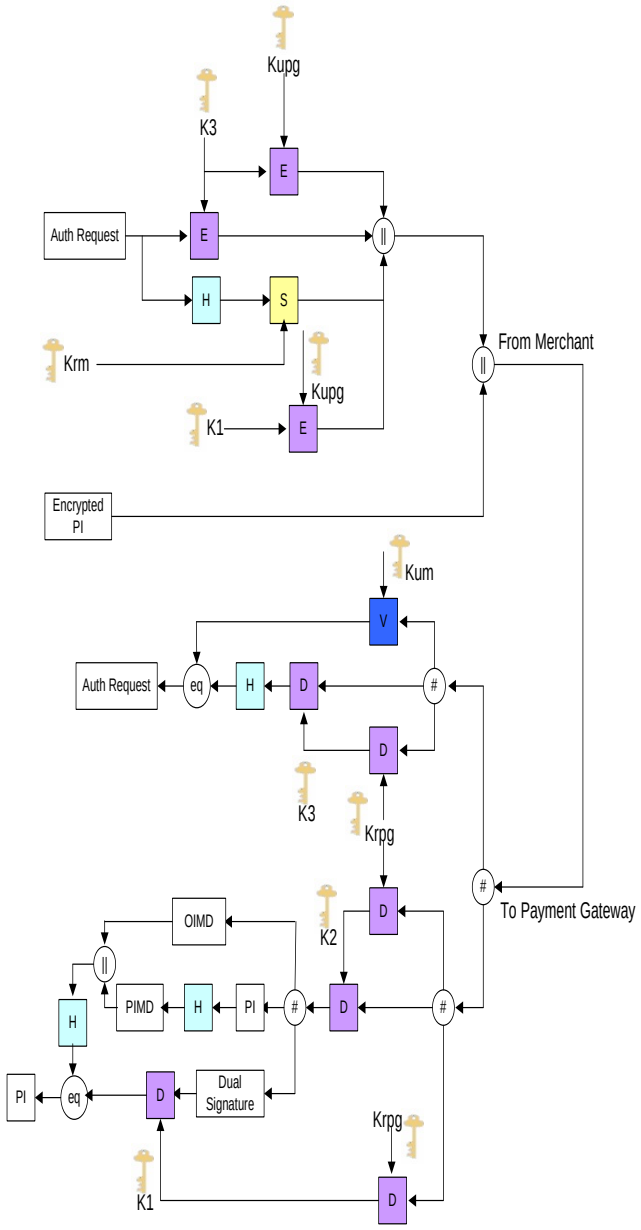


Fig. 5. Authorization request from merchant to payment gateway

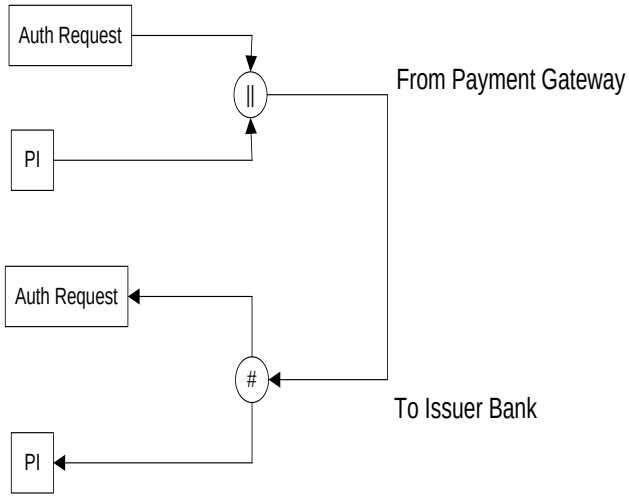


Fig. 6. Authorization request from payment gateway to issuer bank

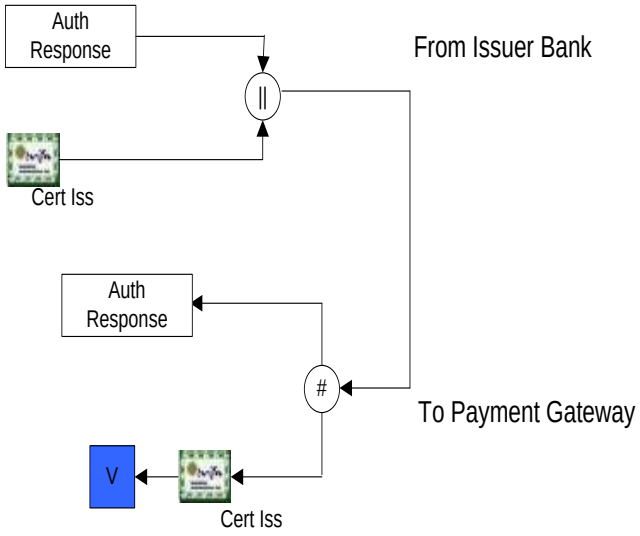


Fig. 7. Authorization response from issuer bank to payment gateway

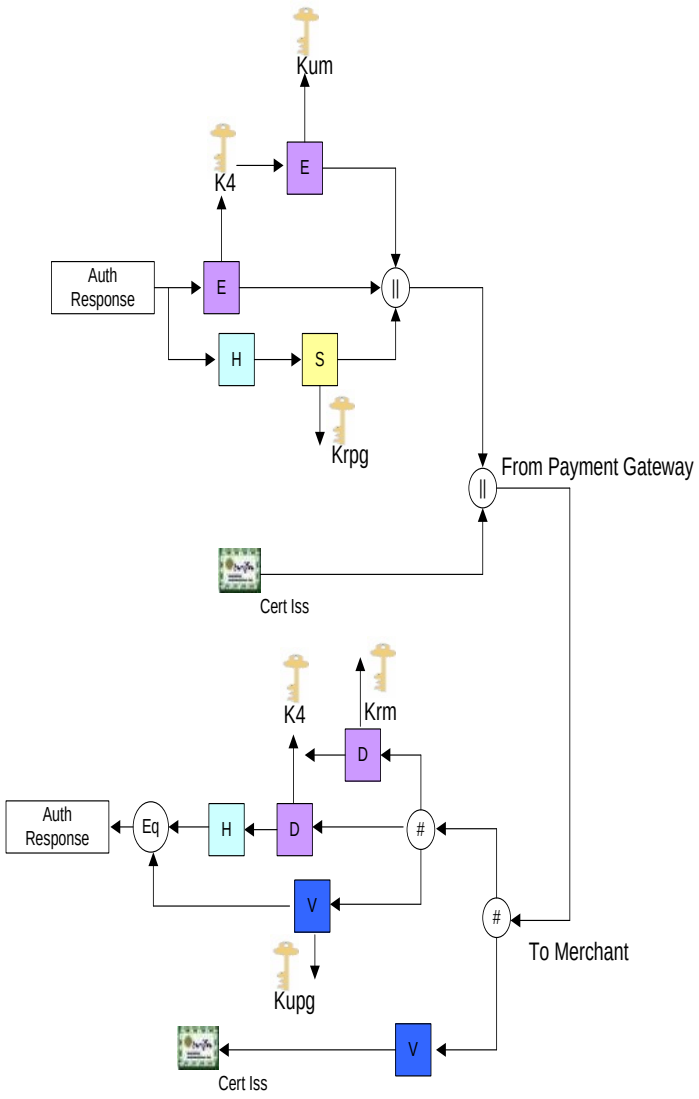


Fig. 8. Authorization response from payment gateway to merchant

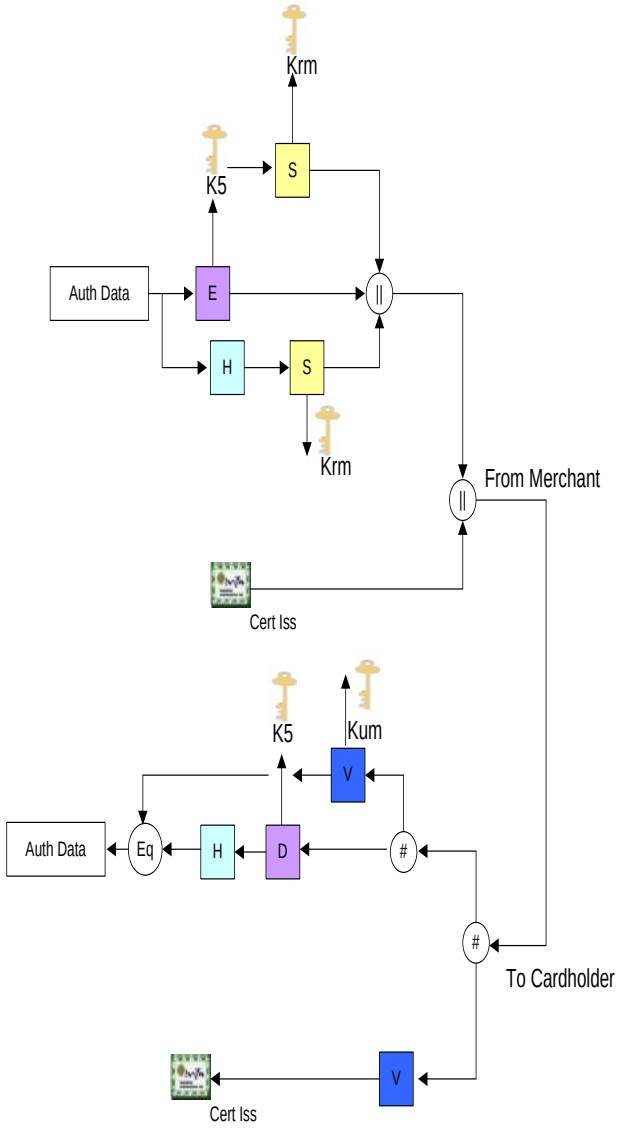


Fig. 9. Authorization response from merchant to cardholder

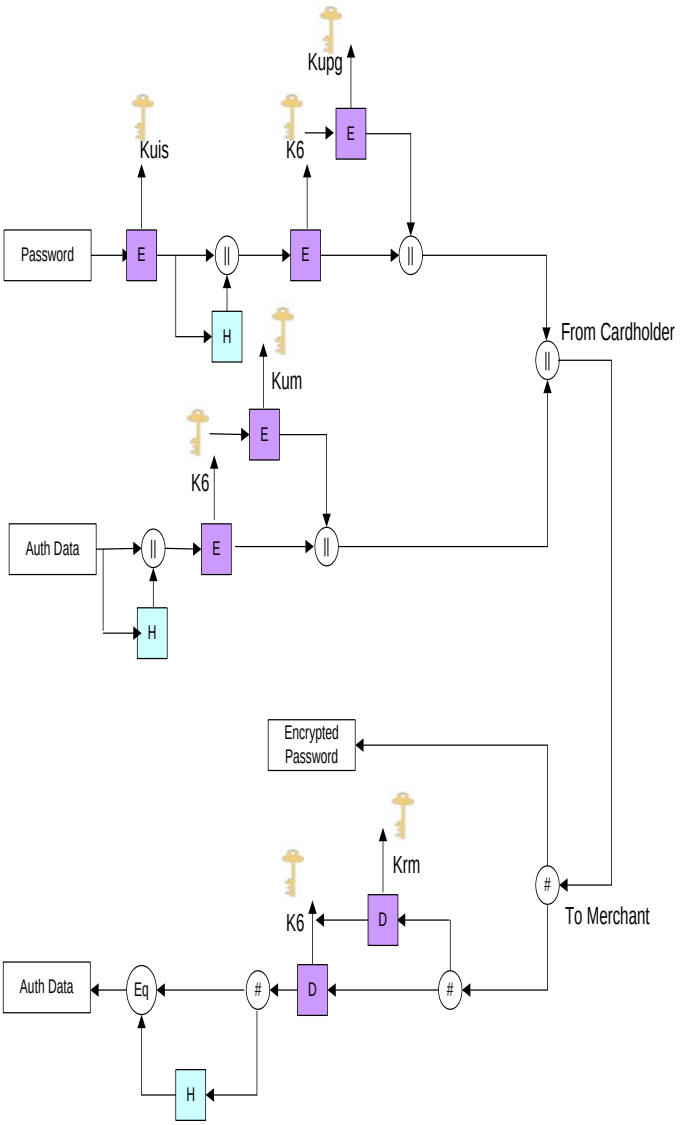


Fig. 10. Cardholder authentication request from cardholder to merchant

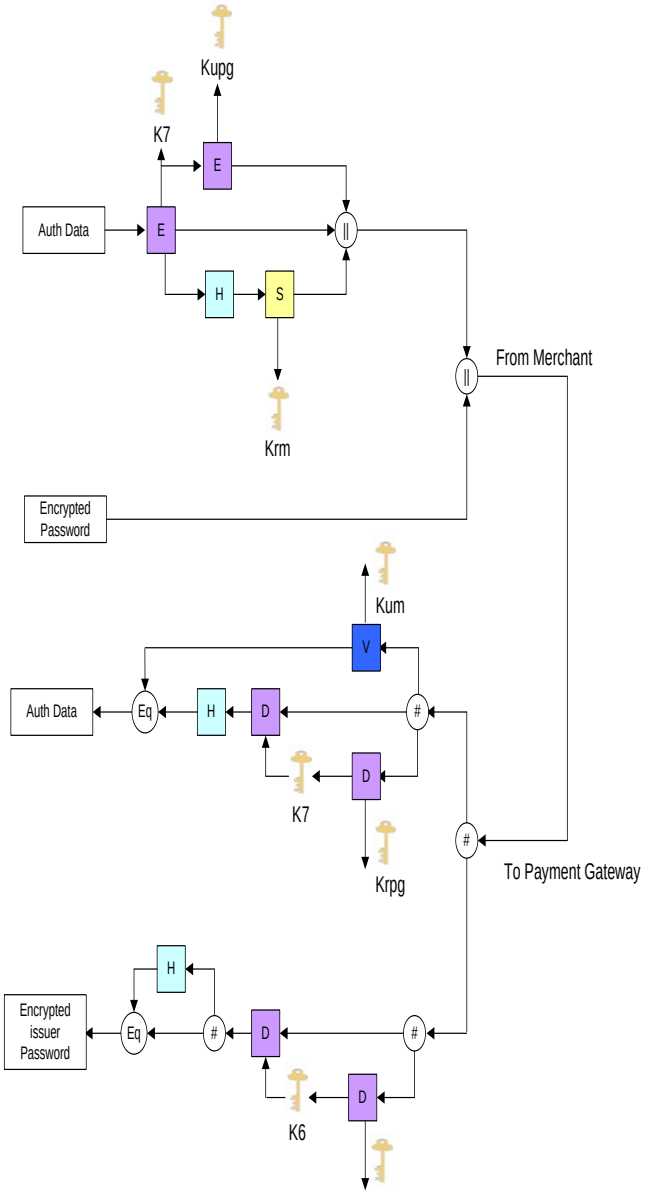


Fig. 11. Cardholder authentication request from merchant to payment gateway

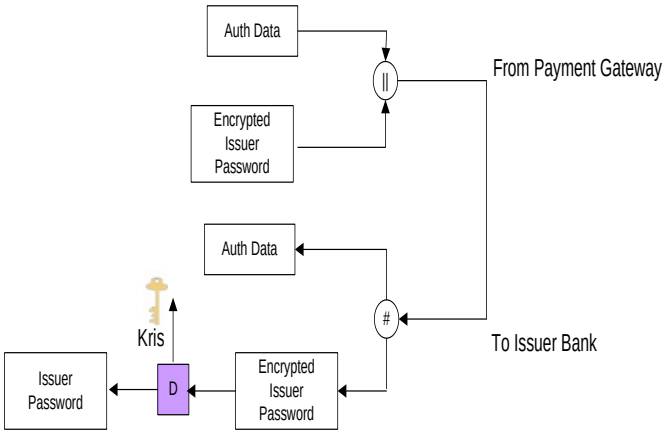


Fig. 12. Cardholder authentication request from payment gateway to issuer

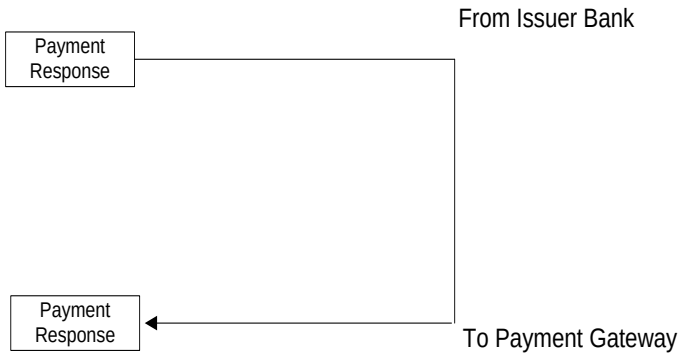


Fig. 13. Payment response from issuer to payment gateway

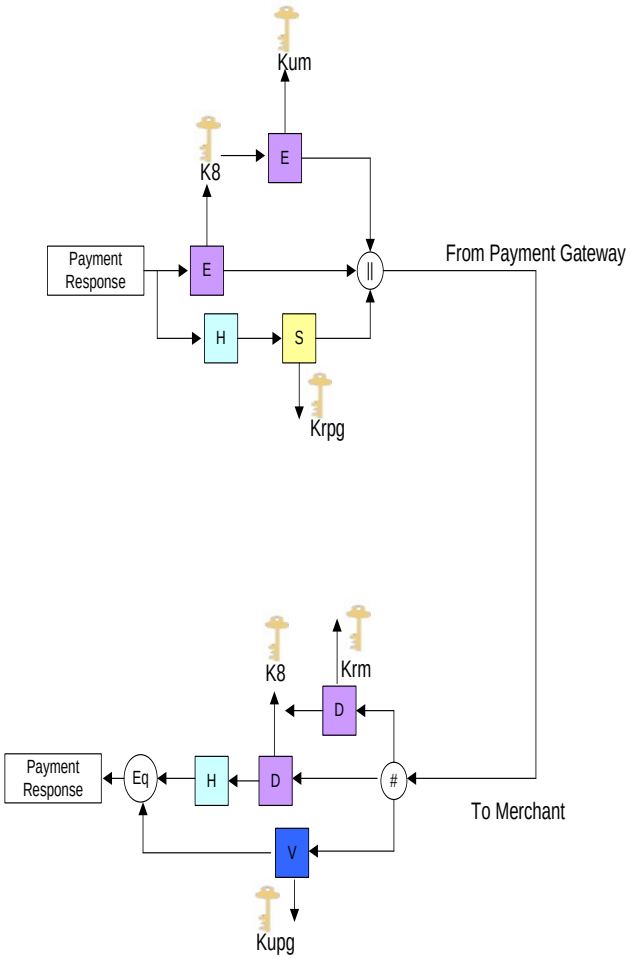


Fig. 14. Payment response from payment gateway to merchant

J. SEP and End-user Implementation Requirements

- ✓ Usability: cardholder, merchant needs to install a special plug. The initialization process is so simple, since the cardholder does not need to have his certificate.
- ✓ Flexibility: SEP protocol have the desirable property that it can be used from any PC, as is currently the case for e-commerce transactions relying simply on

SSL/TLS for cardholder- merchant communication security.

- ✓ Affordability: if we compare SEP with 3D-Secure, 3D-Secure needs more investment in term of connectivity with VISA and ACS setup cost, also the merchant should be able to manage the cardholder authentication redirection to VISA. SEP needs just the attribution of security certificates to merchant, payment gateway and issuer bank, and plug-in setup.
- ✓ Reliability: Of course, whilst the presence of incorrect functionality in security critical elements of SEP protocol is unlikely, there is still a significant possibility that accidental vulnerabilities will be present in implementation. Past experience indicates that it is very difficult to produce software which does not possess vulnerabilities exploitable by malicious software.
- ✓ Availability: Unlike 3D-Secure, for SEP protocol card issuers and acquirers are not required to implement any system with VISA. Once the issuer has the software, they can support SEP transactions. Equally, consumer will be happy to perform a simple registration process to get the password coder and install the plug-in, no security certificate is needed.
- ✓ Speed of transaction: SEP protocol employs DES for symmetric encryption and RSA for certificate verification. The issuer verification of cardholder identity is an important factor for transaction performance. The SEP protocol avoid the complexity of 3D-Secure related to Visa directory. It's difficult to decide about transaction speed because it's related also to networking speed and server's performance.
- ✓ Interoperability: SEP plug-ins can be installed on the consumer PC easily, so interoperability issues are less likely to arise.

V. CONCLUSION

SEP protocol is a good transaction protocol for credit card payment. In this paper we improved how well SEP protocol meets the e-payment security requirements and identified end-user implementation requirement. A future research topic is to analysis the security and the performance of our protocol.

REFERENCES

[1] HASSLER, V. (2001). SECURITY FUNDAMENTALS FOR E-COMMERCE. ARTECH HOUSE, MASSACHUSETTS

[2] Z. Jiemiao, Research on E-Payment Protocol, Information Management, Innovation Management and Industrial Engineering (ICIII), 2011, pages 121 – 123

[3] G. Dhillon, J. Ohri, Optimizing Security in E-commerce through Implementation of Hybrid Technologies, CSECS'06 Proceedings of the 5th WSEAS International Conference on Circuits, Systems, Electronics, Control & Signal Processing, Pages 165 – 170.

[4] A.A. Slamy, E-Commerce security, IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.5, May 2008

[5] B. Lee, T.Lee, An ASEP (Advanced Secure Electronic Payment) Protocol Design Using 3BC and ECC(F2m) Algorithm, e-Technology, e-Commerce and e-Service, 2004. EEE '04. 2004 IEEE International Conference on, pages 341 – 346

[6] X. Zhang, Implementation of a Suggested E-commerce Model Based on SET Protocol, Software Engineering Research, Management and Applications (SERA), 2010 Eighth ACIS International Conference on, pages 67 – 73

[7] A.Craft, T A and R. Kakar, E-Commere Security, Conference on information systems Applied Research 2009, v2 Washington.

[8] H. Houmani, M. Mejri, Formal Analysis of SET and NSL Protocols Using the Interpretation Functions-based Method, Journal of Computer Networks and Communications Volume 2012, Article ID 254942, page 18

[9] P. Jarupunphol, C. Mitchell, Measuring 3-D Secure and 3D SET against e-commerce end-user requirements, Proceedings of the 8th Collaborative electronic commerce technology and research conference (CollECTeR (Europe) 2003), National U

Opinion Mining and Analysis for Arabic Language

Mohammed N. Al-Kabi
Faculty of Sciences & IT
Zarqa University
Zarqa, Jordan

Amal H. Gigieh
Faculty of Sciences & IT
Ajloun- College
AL-Balqa' Applied University
Ajloun, Jordan

Izzat M. Alsmadi
Information Systems Department
College of Computer & Information Sciences
Prince Sultan University
Riyadh 11586, P. O. Box 66833, Saudi Arabia

Heider A. Wahsheh
Computer Science Department
College of Computer Science
King Khalid University
Abha, Saudi Arabia

Mohamad M. Haidar
Web development Department
Brandtologie Company

***Abstract*—Social media constitutes a major component of Web 2.0 and includes social networks, blogs, forum discussions, micro-blogs, etc. Users of social media generate a huge volume of reviews and comments on daily basis. These reviews and comments reflect the opinions of users about different issues, such as: products, news, entertainments, or sports. Therefore different establishments may need to analyze these reviews and comments. For examples: It is essential for companies to know the pros and cons of their products or services in the eyes of customers. Governments may want In addition to know the attitude of people towards certain decisions, services, etc. Although the manual analysis of textual reviews and comments can be more accurate than the automatic methods, nonetheless, it is time consuming, expensive, and can be In addition subjective. In addition, the huge amount of data contained in social networks can make it impractical to perform analysis manually. This paper focuses on evaluating social content in Arabic language and contexts. Currently, Middle East is an area rich of major political and social reforms. The social media can be a rich source of information to evaluate such contexts. In this research we developed an opinion mining and analysis tool to collect different forms of Arabic language (i.e. Standard or MSA, and colloquial). The tool accepts comments or opinions as input and generates polarity based outputs related to the comments. For example the output can be whether the comment or review is: (subjective or objective), (positive or negative), and (strong or weak). The evaluation of the performance of the developed tool showed that it yields more accurate results when it is applied on domain-based Arabic reviews relative to general-based Arabic reviews.**

Keywords—Sentiment Analysis; Arabic Sentiment Analysis; Opinion mining; Opinion Subjectivity; Opinion Polarity

I. INTRODUCTION

The use of Internet is changed significantly through decades, where in the beginning it was restricted to the connection of four major U.S. universities and a number of government agencies in 1969. Since then the number of servers start to increase rapidly. In 1989 a major event occurs when Tim Berners-Lee casts the term World Wide Web (WWW) which is based on hypertext, and changed the way of communication through the Internet. In 2004 the term Web 2.0 is used and a number of services and tools are released, which makes the WWW more cooperative and sharable. Therefore the key component of Web 2.0 is social media, which helps to serve different societies all around the world.

Web 2.0 is offering products and services that are different from its predecessor WWW. The number of Web 2.0 users increases on a daily base, where it is impossible for a single user to learn and use all these products and services. Web 2.0 helps to let it users to be more collaborative. YouTube, Facebook, Twitter, etc are examples of Web 2.0 services.

Arabic language is spoken by many people in many countries. Arabs constitute around 5% of World population and around 3.8% of Internet users [1]. Recent years witnessed an explosive increase in the volumes of social media data which is broadcasted and shared related to different daily activities. This may include data which use Modern Standard Arabic (MSA) and colloquial Arabic. The colloquial Arabic is greatly varied, and is classified into five main regional forms in the Middle East [2]:

1) Arabian Peninsula Arabic (Khaliji Arabic): Includes Gulf, Baharna, Najdi, Omani, Hejazi, Shihhi, Dhofari, and Yemeni Arabic slangs.

2) Mesopotamian Arabic: Includes Iraqi and North Mesopotamian Arabics.

3) Syro-Palestinian Arabic: Includes Levantine, Judeo, Mediterranean Sea or Cypriot, and Bedawi Arabic.

4) Egyptian Arabic: includes Chadic and Sudanese Arabic including: (Nubi, Juba and DarfuriArabics), Sa'idi and Egyptian Arabic.

This research is funded by the Deanship of Research and Graduate Studies in Zarqa Private University / Jordan.

5) Maghrebi Arabic: includes the Arabic Vernaculars used in North African coast of the Mediterranean Sea such as: Moroccan, Tunisian, Arabic, Arabic, Arabic, and Saharan Arabics.

The Arabs who used Arabian Peninsula Arabic (i.e. Khaliji Arabic) could not understand the accent used in Maghrebi Arabic, so in this case both have to use MSA to communicate with and understand each other.

Social media data include: News stories, opinions, current status, different activities, and comments and reviews about these items. Opinions are essential to people and before the Internet era when somebody needs an opinion he/she asks his/her family, relative or a friend. Customer opinions are In addition essential to companies; therefore they used to conduct surveys in different forms before the evolution of the Internet to evaluate people opinion on some issue or event.

Opinions are then very important. Whenever we need to make a decision we want to hear others' opinions. This is not only true for individuals who may use advices from the others, but In addition true for organizations and governments. Many tools were built and developed to analyze English opinions. The interest in opinion analysis and mining has grown due to different reasons. On one side it is due to the rapid evolution of the World Wide Web (WWW), which changed the view and the use of the Internet. It has changed the web into a collaborative framework where technological and social trends come together. On the other side, the huge use of the services has been accompanied with an increase in freely available online reviews and opinions about different topics, subjects or entities [3].

Opinion mining/sentiment analysis is a new emerging field of study and a very active research area since the year 2003. It is concerned with the analysis of people's sentiments, opinions, attitudes, evaluations, and emotions expressed in one of the known natural languages towards entities such as: persons, products, services, companies, events, issues, or topics. Studies in this field are conducted as part of computer science studies. However, it is In addition conducted in management and social sciences, since some of these studies are important to the business and society [4]. Sentiment analysis and Opinion mining were first explored in 2003 by [5, 6]. Although these two terms (Sentiment analysis and Opinion mining) are not exactly the same, but they used interchangeably by a number of authors, where the meaning of term opinion is broader than the meaning of the term sentiment.

Web-based social network services such as: Twitter, Facebook, and Google+ enable different users with common interests or real-life connections to connect with each other through those virtual networks to share their opinions, ideas, and information. These Web services are applied in different domains such as: Government, Business, Dating, Education, Finance, Medical/health, Social and political applications [7].

According to the leading free provider of Internet Web metrics, Alexa: www.alexa.com, [8], social network sites such as: Facebook were ranked second globally at the time of conducting this study [9]. In addition, YouTube is ranked third, and Twitter ranked tenth. Those social networks in the top ten showed that such websites and services are widely used by humans all over the world. In the Arab countries these Web metrics are similar to those presented on the global level. In Egypt the largest Arab country for example, Facebook is In addition ranked first, YouTube ranked third, etc. Same thing can be said to most of the other countries in the region.

Most opinion analysis and mining methods have been developed for English text and are difficult to generalize to other natural languages such as: Arabic which is highly inflectional. The number of studies in this field which are conducted on Arabic text whether it is expressed in MSA or colloquial Arabic is limited when it's compared to the studies conducted in English sentiments and opinions. Arabic is one of the Semitic languages which is written from right to left, and written in a cursive way. In addition Arabic language has 28 consonants, and has no upper and lower case consonants as in English.

Arabic is a challenging language for a number of reasons: It has a very complex morphology relative to the morphology of other languages such as: English. Arabic language is a highly inflectional and derivational language which makes monophonically analysis a very complex and difficult task [10]. Further Arabic opinions are highly subjective to context domains, where you may face words that have different polarity categories in different contexts. Arabic Internet users mostly used colloquial Arabic rather than using MSA, where colloquial Arabic resources are scarce. The percentage of spelling mistakes within these Arabic opinions is high, and this represents an additional challenge.

These few lines would not be sufficient to list the differences between Arabic and English languages. Therefore it is impossible to apply most of the opinion analysis and mining methods which are proposed and implemented on English sentiments and opinions directly on Arabic sentiments and opinions. Some of the studies related to Arabic opinions/sentiments analysis are using the analysis methods developed mainly, but not directly for English language. These studies use machine translation (MT) to automatically translate Arabic sentiments and opinions to English, in order to be able to use those analytical methods which are designed mainly for English opinions/sentiments. For example Bautinet et al. study [11] and Rushdi-Saleh et al. [12] study conclude that this approach is an attractive one. The use of MT will lead to degradation of the accuracy of final results of the opinion analysis and mining, as a result of the incapability of MT systems nowadays to accurately translate from one natural language into another, as accurately as professional human translators. Our intuition or idea is that such translation is not necessary and is not effective and does not yield more accurate results than methods that are applied directly on mining opinions and sentiments directly, without using machine translation.

In this research, we have developed a tool to analyze different Arabic opinions whether they are written in colloquial Arabic or Modern Standard Arabic (MSA) or both. This was an ambitious goal to develop a tool to deal with both standard and colloquial Arabic. In comparison with previous

tools such as those mentioned in Rushdi-Saleh et al. [12] study which is restricted to MSA, Almas and Ahmad [13] study which is restricted to MSA Arabic financial terms, and El-Halees [14] study which In addition uses MSA, our tool can be hence more comprehensive.

In this study different opinions written in MSA or/and colloquial Arabic are classified into predefined set of categories based on their contents. Classifying those different opinions is not a straight forward process, since the essential lexical resources are not there, especially those related to colloquial Arabic. Implicitly this study includes a manual building of two general purpose lexicons to discern the polarity of an opinion expression, whether the opinion uses MSA or/and colloquial Arabic. In addition, another sixteen domain-specific lexicons were built manually. Those domain-specific lexicons were built to decide automatically the polarity of a sentiment expression within the following eight domains: Technology, Books, Education, Movies, Places, Politics, Products, and Society. So the total number of lexicons built is 18, where nine of these polarity lexicons are dedicated to positive polarity, and the other nine lexicons are dedicated to negative polarity. An opinion is considered neutral, when its tokens are divided equally between positive and negative lexicons. The tool is In addition capable to determine whether Arabic social media reviews are (subjective or objective), (positive or negative), and (strong or weak).

The rest of this paper is organized as follows. Section 2 overviews related work. Section 3 describes the methodology followed with examples showing exactly how our tool works. In Section 4, we present the algorithms implemented in the opinion mining tool. Section 5 presents the results of the experimental analysis and evaluation. Finally in section 5, conclusions and possible future work are discussed.

II. Related Works

A review to previous studies conducted in this field shows that researchers proposed and used several approaches which provide variant solutions to automatic sentiment analysis and opinion mining. This section exhibits a number of these studies about this field, with an emphasis on studies related to automatic analysis of Arabic sentiments and reviews.

Sentiment analysis systems can be divided according to the scope of the input, therefore we have document-level (where the classification of opinions depend on the whole document), sentence-level, or phrasal-level which analyze part of the sentence. Sentence-level sentiment analysis classify sentiments after segmenting the document into several sentences and compute the polarity, while document-level sentiment analysis systems do not segment sentiment's document into several sentences. Pang et al. [15] used a document level polarity categorization to classify opinions. El-Halees [14] study evaluated three different methods to identify the polarity of documents. Yi et al. [16], Kim et al. [17], Elhawary and Elfeky [18], and Abdul-Mageed et al. [19] on the other hand dealt with sentence level polarity categorization attempts to classify positive and negative sentiments for each analyzed sentence. Phrase-level sentiment analysis is conducted by Wilson et al. [20], where they determine first whether the expression is neutral or has a polarity. Afterward if the expression under consideration is not neutral, the contextual polarity is determined.

Elhawary and Elfeky [18] study and similar our study discussed the lack of a standard Arabic dataset for business reviews and sentiments. For Arabic, the Internet lacks websites similar to www.yelp.com which has many English business reviews. Therefore their study started by collecting Arabic business reviews, and dedicating 80% of the collected business reviews to train their classifier which is used to identify review documents. They In addition constructed a number of Arabic lexicons used to analyze different Arabic reviews and sentiments. The polarity of each Arabic business review whether it is: positive, negative, neutral or mixed is judged based on the built lexicons.

A manually annotated corpus of Modern Standard Arabic (MSA) and a polarity lexicon are developed by [19]. The authors In addition developed high performance automatic Subjectivity and Sentiment Analysis (SSA) system which is based on manually annotated MSA corpus.

Different methods were used by El-Halees [14] to determine the polarity of a number of Arabic documents. The polarity of the whole Arabic documents is determined first using lexicon-based method, where the output from the first method (lexicon-based) is considered as a training set for maximum entropy method, which is In addition used to classify these documents. Author In addition used KNN method in her study to classify collected Arabic documents.

Sentiment analysis can be divided according to the type of output or the desired classification. Traditionally, sentiment analysis indicates whether a review or comment is positive, negative or neutral. Wilson et al. [21], Abbasi et al. [22], Elhawary and Elfeky [18] studies depend on lexicons containing positive and negative words/phrases ranked by their score, and classify opinions into positive, negative, neutral or mixed. In other classification category the opinions were determined as strong or weak. Some studies proposed a feature weighting schemes that can enhance classification accuracy. Paltoglou et al. [23] study assigns weights to features and applies weighting functions scale linearly related to the number of times a term occurs in a document. This was a significant factor to increase the accuracy of sentiment classification.

One of the earlier approaches adopted in a number of studies is based on translating the source Arabic document (opinions) into English and then use the same applicable techniques to analyze English sentiments. Almas and Ahmad in [13] used machine translation systems to translate the source document or review from (Arabic, Italian, French, Chinese, Korean, German, Japanese, and Spanish) to English language before passing them to an English based sentiment analysis system. The problem of this approach was the loss of nuance after translating the source to English. Rushdi-Saleh et al. [12] used different machine learning algorithms to classify the polarity of Arabic reviews extracted from specialized Web pages related to movies and films. Inui et al. [24] study adopts translating opinions from English to Japanese, followed by sentiment analysis. They applied sentiment-oriented sentence filtering method to mitigate many translation errors that occur

as a side-effect of translation to reduce the influence of translation errors in multilingual document-level review.

The use of machine translation followed by sentiment analysis is not restricted to Arabic comments and reviews, but it includes other languages. As a sample of these studies the Banea et al. [25] is presented in this section, which used machine translation to translate Romanian and Spanish reviews and comments to English, and then apply the sentiment analysis tools on the translated materials. In addition they improve their study and conduct another study Banea et al. [26], where in this study they added Arabic, French, and German reviews in addition to Romanian and Spanish reviews and comments used in their previous study.

Some of the studies in this field are domain based studies. Domain features should be collected for the domain under consideration, as exhibited in the study of Balahur et al. [27], where the term is used to describe special product classes. Afterward the polarity (i.e. positive or negative) is determined for each of the features attributes using an annotated corpus. Other researchers select domain specific features plus the topic of the opinion as a clue. Choi et al. [28] presents a framework for sentiment analysis, focus on the sentiment clue that is related to a sentiment topic (defined as a primary subject of sentiment expression in a sentence), such as: company, person or event. They use a domain-specific sentiment classifier for each domain with the newly aggregated clues (e.g. a subject or the topic of the opinion) based on a proposed semi-supervised method. Yi et al. [17], Kim et al. [18], Choi et al. [28] extract opinion about a subject focus on the sentiment clue that is related to a sentiment topic. This is defined as a primary subject of sentiment expression in a sentence such as: company, person or event.

Ortiz et al. [29] study views and evaluate a domain independent sentiment analysis system against a multiple-domain opinion corpus. The results showed that high accuracy can be achieved by relying entirely on high quality, manually acquired and linguistic knowledge.

Al-Subaihin et al. [30] study exhibits a design for a sentiment analysis tool for Modern Arabic which segments the reviews into sentences, then collect sentimental meaning of words in each sentence based on sentiment lexicons. The tool can get the pattern of words' role in the sentence and use that pattern to match from a set of the acquired annotated patterns that map the sentence to get the polarity. The whole polarity is deducted from the sentiments of sentences. Their tool focused on Modern Standard Arabic (MSA) only while in this paper we tried to enable the tool to deal with both (Modern Standard Arabic (MSA) and Colloquial Arabic).

Al-Kabi et al. [31] conducted a study to compare two free online sentiment analysis tools: SocialMention and Twendz using Arabic and English comments and reviews. To conduct their study they constructed three polarity dictionaries: English polarity dictionary, Arabic polarity dictionary, and Emoticon polarity dictionary. They conclude that SocialMention is more effective than Twendz. Another study compares two free online sentiment analysis tools (SocialMention and SentiStrength) that support Arabic language is conducted by Khasawneh et al. [32] and based on 1,000 Arabic comments and reviews collected from Twitter and Facebook. They conclude that SentiStrength tool is more effective than SocialMention.

Al-Kabi et al. [33] collected 4625 Arabic reviews and comments from Yahoo!-Maktoob Website. The collected reviews and comments are classified manually into four domains (Arts, Politics, Science and Technology, and Social). They analyze different aspects of the collected dataset such as the reviews' length, the numbers of likes/dislikes, the polarity distribution and the languages used.

III. THE METHODOLOGY

This section presents the adopted approach to automatically analyze large volumes of Arabic user's reviews using both Modern Standard Arabic (MSA) and colloquial Arabic, where the analysis includes adopting classification algorithms to determine: Subjectivity, Polarity, and Intensity.

We first developed a basic lexicon-based tool for Arabic opinion mining. This tool can process Arabic opinions collected from different social media resources, regardless of their domain. Therefore this proposed tool uses word/phrasal sentiment features to handle Arabic textual opinions whether they are using MSA or colloquial Arabic or both. The following steps have to be followed to identify subjectivity, polarity, and intensity.

A. Opinion Analysis Schema

Sentiment analysis is concerned with analyzing the attitude of the opinion holder (i.e. the person who presented the opinion) or in other words analyzing the subjective opinions text (i.e. text containing opinions, emotions or sentiments). This study presents an automatic tool to analyze Arabic opinions regardless of the Arabic language style used whether it is MSA or colloquial Arabic or both. The tool is capable to determine the subjectivity, polarity and intensity of the evaluated Arabic opinions, where specific syntactical features are used to determine the strength of the opinion. The schematic overview of our approach is exhibited in figure 1.

This study is composed of the following five phases:

1) *Dataset collection.*
2) *Text normalization.*
3) *Specific features extraction from the opinions text that was collected.*
4) *Lexicons creation.*
5) *Using classification algorithms to classify opinions into several categories based on the lexicon that was built.*

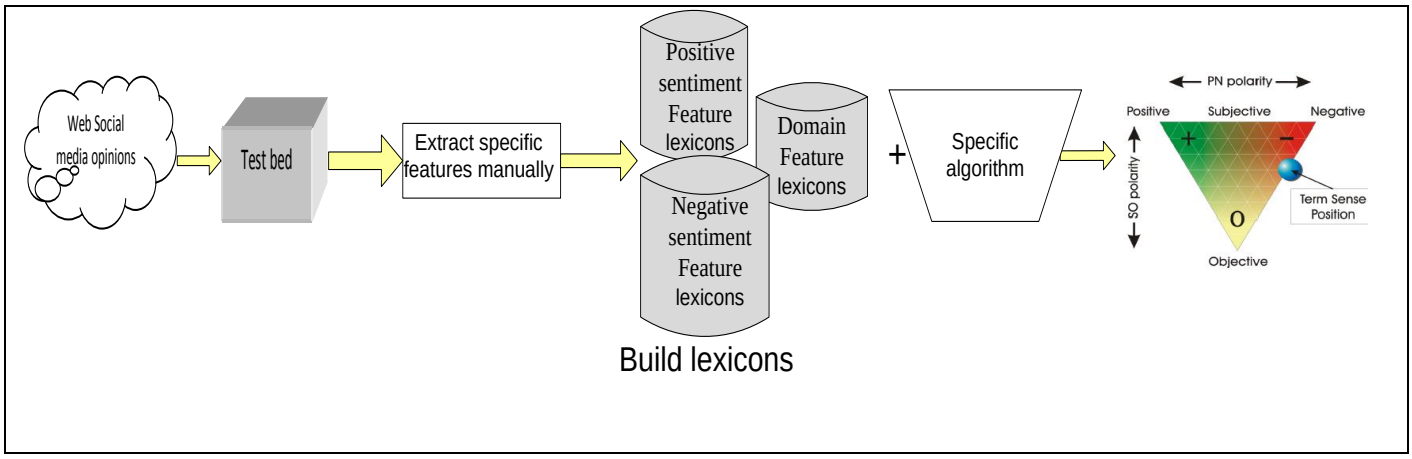


Fig. 1. Opinion Analysis Schema

B. Dataset collection

This study started by collecting Arabic reviews from 72 social media websites. The total number of the collected Arabic reviews was 1,080. These reviews use colloquial Arabic or Modern Standard Arabic (MSA) or both.

1) Dataset Characteristics

This section exhibits some of the characteristics of the collected Arabic reviews:

Some reviews consist of only one word, e.g. good,"جيد". Chat language is used to express some reviews such as:" 7ilwi awi" which means in English "very sweet".

Latin letters and English phonetics (transliteration) are used to express Arabic phrases such as "jamiljidannnnnnnnnnn" which means in English "very nice". An appropriate method has to handle repeated Latin and Arabic letters.

In addition some of the collected Arabic reviews use elongation, through the use of dash-like "kashida" character to stretch the Arabic word, i.e., حلـــو جـــــــــــدا which means very sweet. This extension or Kashida should be handled by the developed system as well.

Most of the collected Arabic reviews use a mixture of colloquial Arabic and MSA, such as (very nice, "وايد جميل"), where the Arabic word (very, "وايد ") is a colloquial Arabic word, and the Arabic word (nice, "جميل") is an MSA word.

Many of the collected Arabic reviews contain spelling errors, such as (very beautiful, " جمي جداء جميل جدا").

Some of the textual reviews are mixtures of Arabic and English. This mixture between Arabic and English is called Arabizi language, i.e." حبيت التصميم it was very nice", which means I like the design it was very nice.

Some of reviews weren't related to the topic of the review, so it is considered a spam review.

There is no exact or specific style or pattern the users have to follow to write their reviews. Therefore we are dealing with fully unstructured Arabic text.

Around 90% of the Arabic reviews in the dataset were opinions or subjective text, and around 10% were objective text (facts).

The above characteristics represent a summary of the Arabic opinions' analysis problems that should be usually handled by any proposed automatic solution or handling system.

C. Preprocessing Steps

The opinion mining tool performs a number of preprocessing steps to normalize and prepare the opinions for processing. First, the tool removes punctuations and non-Arabic letters. In addition, some Arabic letters are normalized such as: (أ ,إ, and آ) are converted to (bare Alif, ا), and (Yaa', "ئ ,ى") replaced by (Yaa', ي) and, (taa', haa', "ة,ه") replaced by (haa',ه). These steps are shown below:

1) Remove digits, punctuations, special symbols and non- letters.

2) Normalization:

a) Convert (Alif, "أ,إ,آ,ا,ى,ٱ) to (bare Alif, ا).

b) Convert (Yaa', "ئ ,ي") to (Yaa', ي).

c) Convert (Taa' Marbuuta , "ة", final haa', "ه") to (final haa', "ه").

d) Convert (Waaw "ؤ","و") to (Waaw, "و").

3) Filtering non-Arabic text

4) Tokenization

D. Taxonomy of Opinion Analysis

Here is the taxonomy for the major concepts and steps used to analyze different Arabic reviews. Table 1 presents different main taxonomies generated by the tool.

TABLE I. MAIN CLASSIFICATION CATEGORIES

Classification Category	Description
Subjectivity Classes	Opinion/Fact
Polarity Classes	Positive/Negative/Neutral Sentiments
Strength (Intensity) Classes	Strong/Weak/Normal Opinion Text
Domain Classes	Determine on what domain is the opinion

The tool uses the features shown in table 2 to generate the taxonomies shown in table 1. Features shown in table 2 are extracted manually from the collected Arabic reviews.

TABLE II. FEATURES CATEGORIES

Feature Category	Description
Domain Features	All words or bag of words which can distinguish domains from each other.
Polarity Features	All words/phrases yield (positive or negative) sentiment in opinion text.
Negation Features	All words that preclude the word or sentence.

Table 3 exhibits the main techniques adopted in this study to classify different Arabic reviews.

TABLE III. OPINION ANALYSIS TECHNIQUES

Classification Category	Description
Machine Learning	Naïve Bayes Technique.
Similarity Score	Word/Phrase Matching, frequency term counts, weight score.
Normalization and Tokenization	Prepare Arabic opinions before analysis.

This tool can handle Arabic general opinions collected from different social media recourses, and try to categorize them into specific domains. Table 4 shows the domains of different Arabic reviews covered in this study.

TABLE IV. CLASSIFICATION DOMAINS

Classification Category	Description
General Domain	Independent Base Domain.
Specific-Domain Arabic Opinion	Technology, Books, Education, Movies, Places, Politics, Products, and Society.
Web Media Corpus	Social media web pages e.g. (Facebook, blogs, online news, forums).

Our tool is based on more than one lexicon to classify different Arabic opinions. These lexicons contain the extracted features included in the dataset collection, where the content of each lexicon is shown in table 5.

TABLE V. LEXICON CATEGORIES

Category	Description
Polarity Lexicon	Contains the Positive and Negative Sentiment's features.
Domain Lexicon	Contains the features that discriminate specific domain from the others.
Strength Lexicon	Polarity lexicons with weight for each entry.
Negation Lexicon	Contains the negation words.

E. Feature Extraction

Opinion features are extracted manually. After collecting opinions' dataset, these features are used to construct different lexicons used in the analysis and classification steps. Figure 2 shows the essential steps to extract different types of features which are used in this study.

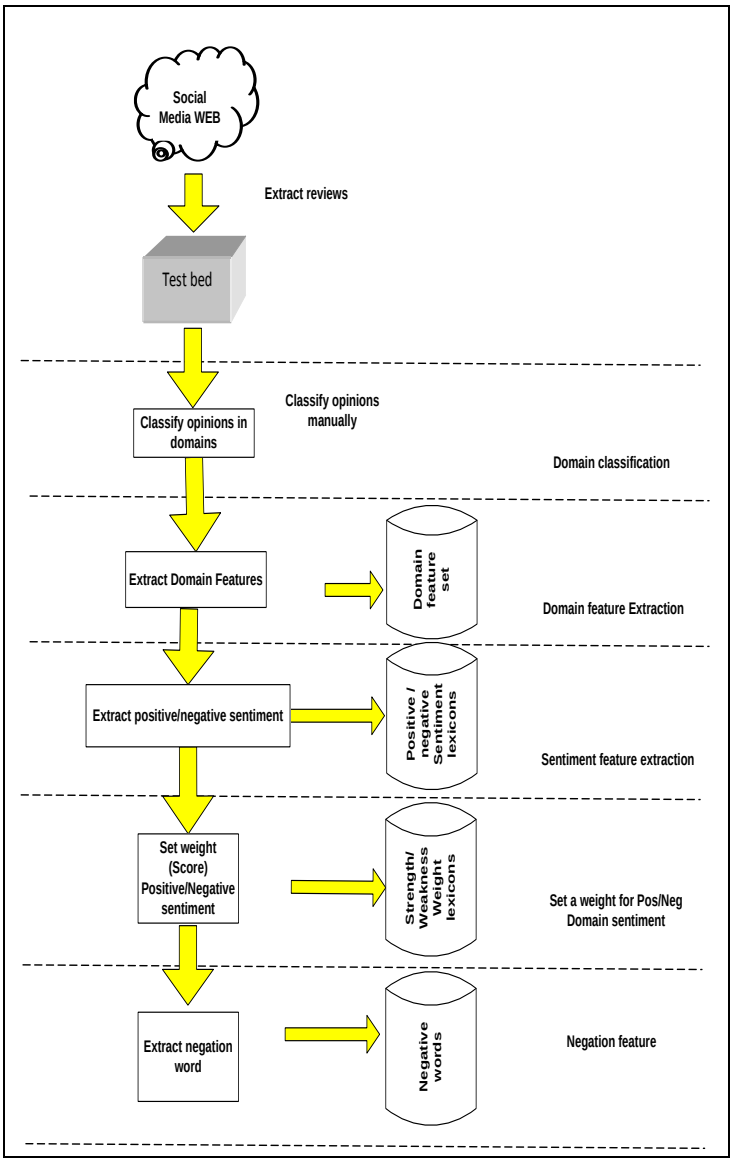


Fig. 2. Outline of Feature Extraction.

1) Domain Features

Domain features are used as clues to determine the domain to which the opinion may belong to using classification algorithms [34, 36]. These features are collected from the training dataset after classifying them manually into domains, to select the features that can discriminate one domain from another. In other words to use them as inputs (training data) to the classifier, to determine the instance reviews related to any domain automatically (domain adaptation).Our dataset is classified into eight domains: Technology, Books, Education, Movies, Places, Politics, Products, and Society.

To prepare the domain sentiment lexicons, we extract the domain features from the opinions text after classifying the dataset manually into the different domains.

2) Polarity Features

Polarity features are divided into positive and negative (sentiments). These features are extracted from the collected Arabic reviews to build the polarity lexicons. Arabic polarity features are Arabic words or phrases that express the positivity or the negativity of the user attitude related to a specific topic. These features are considered from syntactical point of view such as: "adjectives", "verbs", "nouns", "adverbs". They may also come as a mixture of a "group of words".

As mentioned before the main challenge to researchers in Arabic opinion analysis field is the lack of necessary resources, especially the lack of polarity sentiment lexicons. Therefore we have to create these lexicons which contain the positive and negative features already extracted manually from Arabic reviews. Two polarity examples are shown in this subsection. The first example shows how to extract a positive polarity feature, while the second example shows how to extract a negative polarity feature.

3) Negation Features

Arabic negation words represent all the words that negate Arabic words and sentences. Arabic negation keywords such as: (no, "لا") and (not, "لم") convert the sentiment polarity state to an opposite state.

4) Examples

The examples exhibited in this section show how to extract manually the essential features from the collected Arabic Opinions.

Example 1: Consider the following sample in the next excerpt, of the collected Arabic reviews (Movies domain) with its English translation:

Arabic Comment	فيلم رائع جدا وممتع استمتعوا بمشاهدة هذا الفيلم و هو يعتبر نقلة في صناعة هوليود الفيلم خليط من أفلام زي ماتريكس و سبيدرمان
English Translation	"A wonderful and enjoyable movie. Enjoy watching this movie, it is considered as a shift in the Hollywood industry. It is a mix between movies like Matrix and Spiderman "

Table 6 exhibits manually the extracted domain features from the above Arabic excerpt. This sample is taken from the domain of movies. Domain features are used to determine the domain of each Arabic review within the dataset. To be more specific, a classifier (e.g. NB) is used to determine the domain of each Arabic review and comment.

TABLE VI. MANUALLY EXTRACTED DOMAIN FEATURES

أفلام	الفيلم	فيلم	بمشاهدة	هوليود
Films (Movies)	The film (movie)	Film (movie)	watching	Hollywood

Table 7exhibits a sample of positive polarity features extracted from the above Arabic sample review. These features are stored in the polarity lexicon to be used later by the tool to determine the polarity of different Arabic reviews and comments.

TABLE VII. A SAMPLE OF POSITIVE POLARITY FEATURES

Arabic Sentiment Features	ممتع	رائع
English Translation	Enjoyable	Wonderful
Polarity	Pos	Pos

Example 2: Consider the following sample in the next excerpt, of the collected Arabic reviews (Movies domain) with its English translation:

Arabic Comment	فيلم طويل وبطيء وممل
English Translation	"The movie was long, slow and boring "

In this case the manually extracted domain features are restricted to one feature (Movie, "فيلم"), so the NB classifier is based on this feature to determine the domain of the above excerpt. Table 8 exhibits a sample of negative polarity features extracted from the above Arabic sample excerpt.

TABLE VIII. A SAMPLE OF NEGATIVE POLARITY FEATURES

Arabic Sentiment Features	بطيء	ممل	طويل
English Translation	Slow	Boring	Long
Polarity	Neg	Neg	Neg

In addition two other examples are presented in this study about political reviews in example 3 and 4. These two examples show how the tool can determine the polarity of any review, and how to determine whether the review is a fact or an opinion.

Example 3:

Consider the following sample in the next excerpt, of the collected Arabic reviews (Politics domain) with its English translations:

Arabic Comment	انا ضد هذا القرار السياسي، قرار صعب يشكل ازمه للمواطنين.
English Translation	I am against this political decision, it is a hard decision It may cause a crisis for citizens
Arabic Comment	انا ضد هذا القرار السياسي، قرار صعب يشكل ازمه للمواطنين.

Table **9** exhibits manually extracted polar features from the above Arabic political excerpt.

TABLE IX. MANUALLY EXTRACTED FEATURES FROM AN ARABIC POLITICAL REVIEW

Arabic Sentiment Features	ازمه	صعب	ضد
English Translation	Crisis	Hard	Against
Polarity	Neg	Neg	Neg

Therefore the tool considers the above Arabic political excerpt as a negative point of view. Next is another example that shows the review under consideration as a fact, where this review is free from any sentiment. This review expresses a fact and not an opinion.

Example 4:

Consider the following Arabic excerpt which is considered by the tool as a fact, since it is free from any polarity feature.

Fact:

Arabic Comment	هذا القرار جاء لتنفيذ احكام القانون فقط
English Translation	This decision was made to implement the provisions of the law only.

F. The Lexicons

Our tool depends on the lexicons already built and composed of the manually extracted features as it is a lexicon-based tool. So in the following subsections we will present a brief summary about each one of these lexicons.

1) Domain Lexicons

Domain features were extracted manually to be used by the classification process to identify automatically the domain of each evaluated Arabic review.

2) General Polarity Lexicons

Two lexicons were created to classify opinions. The first lexicon is for positive sentiment features which contain 2,404 positive features or sentiments. The second lexicon is for negative features or sentiments which contain 5,521 ones. These positive and negative features/terms sentiments were collected from the training dataset and there is a part added by translating an English sentiment lexicon presented in [36].

3) Domain-based Sentiment Lexicons

Two lexicons were built for every domain. One for Arabic positive opinions and the other for Arabic negative opinions.

4) Score (Weight) Lexicon

Polarity lexicons used in this study have a weighting score for each Arabic term/feature in these lexicons. Those weights were proposed by the authors of this study. The values of the weighting scores are in range of: (1 to 9) for both positive and negative features/terms, where 1 indicates that the feature/term is the weakest possible positive or negative feature/term, and 9 indicate that the feature/term is the strongest possible positive or negative term/feature.

G. The Classification Categories Set

This study is based on the syntactical features using the sentiment term frequencies to identify the subjectivity and the polarity of different Arabic reviews and comments. In addition the weight scores of sentiment features are used to determine the polarity and the strength of each Arabic review and comment.

The sentiment features used in this study are terms extracted manually from the collected Arabic comments and reviews which correspond to documents in this field. Where TF (term frequency) refers to the number of times a specific term Ti occurs in D (Arabic comment/review). The weight of each sentiment feature is determined manually.

The tool depends on the frequency of positive and negative features/terms to identify the polarity of evaluated Arabic review. The evaluated Arabic review is considered positive when the frequency of positive terms/features in it exceeds the frequency of negative terms/features in the same Arabic review. The tool considers the evaluated Arabic review as negative when the frequency of negative features/terms in it exceeds the frequency of positive terms/features in the same Arabic review. The evaluated Arabic review is considered by the tool as neutral if the frequency of positive terms/features in the evaluated Arabic review equals the frequency of negative features/terms. In addition, the tool depends on the scores in polarity lexicons which are used to determine the strength on each evaluated Arabic review.

The following paragraphs show the pseudo code used in our tool to identify different taxonomies of Arabic reviews evaluated to this tool. The tool considers any evaluated Arabic review free from any terms in the polarity lexicons a fact and not an opinion.

Opinion Determination:

If (No. of Positive terms in a review > 0) or (No. of Negative terms in a review>0) then Review is opinion

Fact Determination: If (No. of Positive terms in a sentiment = 0) && (No. of Negative terms in a sentiment = 0) then Review is Fact.

Let

P={*Pos*, *Neg*, *NU*,*U*}, where *Pos*: Positive review, *Neg*: Negative review, *NU*: Neutral review, and *U*: undetermined

$\sum_{i=1}^{n} Pos - TF_i$: Total number of positive terms in the evaluated review.

$\sum_{i=1}^{n} Neg - TF_i$: Total number of negative terms in the evaluated review.

Then the polarity is determined as shown in the following pseudo code:

Review Positive Polarity Determination:

$$If \left(\sum_{i=1}^{n} Pos - TF_i \right) > \left(\sum_{i=1}^{n} Neg - TF_i \right) then$$
$$P \leftarrow Pos$$

$$P_{pos\%} = \frac{\sum_{i=1}^{n} Pos - TF_i}{\left(\sum_{i=1}^{n} Pos - TF_i \right) + \left(\sum_{i=1}^{n} Neg - TF_i \right)} \times 100$$

Review Negative Polarity Determination:

$$If \left(\sum_{i=1}^{n} Pos - TF_i \right) < \left(\sum_{i=1}^{n} Neg - TF_i \right) then$$
$$P \leftarrow Neg$$

$$P_{neg\%} = \frac{\sum_{i=1}^{n} Neg - TF_i}{\left(\sum_{i=1}^{n} Pos - TF_i\right) + \left(\sum_{i=1}^{n} Neg - TF_i\right)} \times 100$$

Review Neutral Polarity Determination:

$$If \left(\sum_{i=1}^{n} Pos - TF_i\right) = \left(\sum_{i=1}^{n} Neg - TF_i\right) then$$
$$P \leftarrow NU$$

Undermined Polarity Determination:

If (*Pos-TF = Neg-TF*) *and* (*Pos-TF* ≠ 0) *then*
P←U

The tool second step is dedicated to compute the (strength/intensity) of each evaluated Arabic review with its polarity. The computation of the strength/intensity is based on emotions closeness to (sentiments/opinions) as shown in [3] and [37].The following pseudo code shows different types of strength/intensity and the formulas used to compute them:

Let

I={SP,SN,WP,WN, NU,U }, where SP: *strong positive, SN: strong negative WP: weak positive, WN: weak negative, NU: neutral, U: undetermined.*

Max_Pos_Score: Max of the set of Pos-Weights of the positive sentiments

Max_Neg_Score: Max of the set of Neg-Weights of the negative sentiments

Then the polarity is determined as shown in the following pseudo code:

Review's Strong Positive Polarity Determination:

If (Max_Pos_Score>Max_Neg_Score) and
(Max_Pos_Score≥5) then
I←SP

Review's Strong Negative Polarity Determination:

If (Max_Pos_Score<Max_Neg_Score) and
(Max_Neg_Score≥5) then
I←SN

If (Max_Pos_Score>Max_Neg_Score) and
(Max_Pos_Score<5) then
I←WP

Review's Weak Negative Polarity Determination:

If (Max_Pos_Score<Max_Neg_Score) and
(Max_Neg_Score<5) then
I←WN

If (Max_Pos_Score =0) and (Max_Neg_Score =0) then
I←NU

Review Undetermined Polarity Determination:

If (Max_Pos_Score = Max_Neg_Score) and
(Max_Pos_Score≠0) then
I←U

Max_Neg_Score) and
(Max_Neg_Score<5) then
I←WN

5) Examples:

Consider the following Arabic excerpts from the places domain with its English translations. This example includes four Arabic reviews considered by the tool as: Positive, Negative, Undetermined, or Neutral.

Example 1:

Positive Review:

Arabic Comment	الخدمه في هذا الفندق اكثر من رائعه، انبسطت كثير على التصاميم داخل المبنى، جد جميل هالمكان ولو انه بعيد.
English Translation	The service in this hotel is more than wonderful; I really enjoy many designs inside the building, very beautiful place even that far.

So the algorithm identifies the above Arabic review from the places domain as a positive review since it has three positive terms and one negative term shown in table 10.

TABLE X. MANUALLY EXTRACTED FEATURES FROM A POSITIVE ARABIC REVIEW RELATED TO PLACES

Arabic Sentiment Feature	بعيد	جميل	انبسطت	رائعه
English Translation	Far	Beautiful	Enjoyed	Wonderful
Polarity	Neg	Pos	Pos	Pos

Another example is provided in this section to show how he tool identifies an evaluated Arabic review as undetermined.

Example 2:

The Arabic comment shown in this example is selected from places domain. The tool identifies an Arabic comment/review as undetermined when the number of positive polarity features is equal to the number of negative polarity features.

Undetermined:

Arabic Comment	خدمتهم سيئه ولكن المكان جميل
English Translation	Their services are bad, but the place is beautiful

Table 11 shows the essential two extracted Arabic features with their polarities and English translations. This table shows equality in the number of positive and negative Arabic features/terms extracted from the above Arabic comment. The tool labeled the above Arabic comment as undetermined when the frequencies of opposite polarities are equal.

TABLE XI. MANUALLY EXTRACTED FEATURES FROM AN UNDETERMINED ARABIC REVIEW RELATED TO PLACES

Arabic Sentiment Feature	جميل	سيئه
English Translation	Beautiful	Bad
Polarity	Pos	Neg

The Arabic comment shown in example 9 is considered by the tool as a fact and not an opinion (subjectivity classification). In addition the tool identifies the same comment as neutral and not as a positive or negative (polarity classification).

Example 3:

The Arabic comment presented in this example is identified by the tool as a fact within subjectivity category, and as neutral within polarity category since it is free from any extracted feature.

Fact/Neutral:

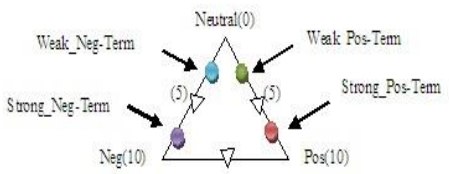
Arabic Comment	هذا الفندق يقع في شمال عمان
English Translation	This hotel is located north of Amman.

A number of Arabic reviews related to books domain are presented below to show how we can determine the strength and polarity using weight scores for the terms in the review.

Example 4:

Consider the following Arabic comment which belongs to domain of books. This review is characterized by using only positive features/terms, so it will be identified by the tool as a positive Arabic comment. This tool will search for the highest score (weight) in such cases, where the highest weight of the extracted features from the Arabic comment presented in this example is 10. If the value of the highest weight of features extracted from the Arabic comment exceeds 5 the tool will consider the strength of comment under consideration as strong, otherwise the strength of the comment will be considered weak.

In such cases, the tool will consider the above Arabic comment as a strong positive, since the highest weight shown in table 12 is 10, which implicitly means that this comment is strong, and since all comment features are positive, it considered a strong positive comment. This method to identify the strength of each Arabic review and comments is suggested by the authors of this study.



Strong Positive:

Arabic Comment	كتاب رائع ، محتواه جيد والمواضيع فيه مميزه
English Translation	A wonderful book, the content is good and the topics are distinctive

Three positive features are extracted from the above Arabic excerpt with their strength weights are shown in table 12.

TABLE XII. MANUALLY EXTRACTED FEATURES WITH THEIR STRENGTH WEIGHTS FROM AN ARABIC REVIEW RELATED TO BOOKS

Arabic Sentiment Feature	مميزه	جيد	رائع
English Translation	Distinctive	Good	Wonderful
Polarity (Weight)	Pos (w=10)	Pos (w=4)	Pos (w=10)

In addition consider the following Arabic review from the domain of books which considered by our tool as weak positive.

Example 5:

Consider the following Arabic review from the domain of books which considered as a weak positive by our tool, since it has only one weak positive feature/term so it will considered positive and it intensity it will be considered weak since the weight of this feature/term is 4 which is less than 5, so it will be identified as weak.

Weak Positive:

Arabic Comment	كتاب جيد الى حد ما.
English Translation	A good book to some extent.

Table 13 exhibits the extracted feature from the above Arabic review about a certain book.

TABLE XIII. MANUALLY EXTRACTED FEATURE WITH IT STRENGTH WEIGHT FROM THE ABOVE EXAMPLE 7

Arabic Sentiment Feature	جيد
English Translation	Good
Polarity (Weight)	Pos (w=4)

Example 6:

Consider another Arabic review related to books domain, and considered by the tool as strong negative review. The following Arabic comment has only negative extracted features, so it will be considered a negative comment by the tool.

In the case the tool has six negative weight, so the tool output is based on the highest score (weight) in such cases. Therefore in this example there are 3 weight values (4, 7, and 10), and since 10 is the highest our tool will identify the Arabic comment presented in this example as a strong negative.

Strong Negative:

Arabic Comment	كتاب ممل ولا يوجد ترابط بن المواضيع ، انا ما حبيته وما عجبني ،لا يستحق كل هالحكي لغته ضعيفه ،معلوماته سطحيه
English Translation	Boring book with no association between its topics. I do not like it, it is not worth mentioning, since its language is weak with superficial information.

TABLE XIV. MANUALLY EXTRACTED FEATURES WITH IT STRENGTH WEIGHTS FOR THE ABOVE EXAMPLE

Arabic Sentiment Features	سطحيه	ضعيفه	لا يستحق	ما عجبني	ما حبيته	ممل
English Translation	Superficial	Weak	Not worth	I dislike it	I do not like it	Boring
Polarity (Weight)	Neg(w =4)	Neg (w=7)	Neg (w=7)	Neg (w=10)	Neg (w=10)	Neg (w=7)

The tool uses the extracted features in table 14 to identify the above Arabic review as strong negative. Consider another Arabic review related to books domain, and considered by the tool as a weak negative review.

Example 7:

It is usual in Arabic and English to face sentences which have words that used before or after the extracted features/terms and leads to reduce the weights ere some words, so in the following Arabic comment the user uses the Arabic colloquial word (Somewhat, "شوي ") after the MSA word (Difficult, "صعبه") and this leads to a change in the strength of phrase, where the feature (Difficult, "صعبه") is saved in negative polarity lexicon and given a weight of 8, but these two terms are stored in negative polarity lexicon and the phrase (Somewhat difficult) given a weight equals to 4. This weight is considered by the tool as a weak negative review.

Weak Negative:

Arabic Comment	كتاب مفاهيمه صعبه شوي وتحتاج الى توضيح يا ريت الكاتب راعى مستويات القراء ، بس الصراحه محتاج تعديلات.
English Translation	A book with somewhat difficult concepts, and need to be clarified. It would be better if the writer took into account the levels of readers, but frankly it needs a revision.

The polarity and strength weight shown in table 15 are used by the tool to identify the above review as weak negative review.

TABLE XV. MANUALLY EXTRACTED FEATURE WITH IT STRENGTH WEIGHT

Arabic Sentiment Feature	صعبه شوي
English Translation	Somewhat difficult
Polarity (Weight)	Neg (w=4)

Example 8:

This example shows how the tool identifies an Arabic review as undetermined. Strength determining algorithm labeled Arabic review as undetermined when the values of high strength weights are equal as shown in the following sample review from books domain. This example is based on Algorithm 3 which is presented in section 4.3 of this study.

Undetermined:

Arabic Comment	كتاب مميز في طرح المواضيع ولكنه لا يعرض المواضيع بشكل مترابط وهذه سيئه في الكتاب.
English Translation	Distinguished book in presenting topics, but does not present topics coherently, and this is a disadvantage of the book.

The tool uses the polarity and strength in table 16 to identify the above review as undetermined.

TABLE XVI. MANUALLY EXTRACTED FEATURES WITH IT STRENGTH WEIGHTS FOR THE ABOVE EXAMPLE

Arabic sentiment	سيئه	مميز
English Translation	disadvantage	Distinguished
Polarity (Weight)	Neg(w=10)	Pos(w=10)

IV. ALGORITHMS

This section presents the pseudo code of the algorithms adopted in the opinion mining tool. The tool enables its users to input a single Arabic review/comment or a group of Arabic reviews/comments to identify their subjectivity (fact/opinion), polarity (Pos/Neg/Neut), and strength.

A. Subjectivity Algorithm

The following algorithm is adopted by the tool to identify different Arabic reviews evaluated as facts or opinions.

Algorithm 1: CNSA-MSA-SAT "determining the subjectivity"

Input:

R= Review/Document Text
T= the set of the Opinion tokens
PD= the set of Positive Sentiment Dictionary
ND= the set of Negative Sentiment Dictionary

Output*:*

R_1={*F,O*} *where F is a Fact and O is a Opinion*

Initialization:

Pos-TF =0 where Pos-TF: term frequency for Positive Sentiments
Neg-TF =0 where Neg-TF: term frequency for Negative Sentiments

```
Begin
1:  For  each ti Є T do
2:  Find ti Є PD where ti Є T
3:    If  ti Є PD then
4:  Pos-TF ←Pos-TF +1
5:  End if
6:  Find ti Є ND where ti Є T
7:    If  ti Є ND then
8:  Neg-TF ←Neg-TF +1
9:   End if
10:   end for
11:  If Pos-TF >1 or Neg-TF >1 then
12:    R1←O
13: Display  R1
14:  Else
15:   R1←F
16: Display  R1
17: end if
End
```

B. Polarity Algorithm

The following algorithm is adopted by the tool to determine the polarity of evaluated Arabic reviews regardless

whether they are using MSA or colloquial Arabic. Each evaluated Arabic review is considered either as positive, negative, neutral or undetermined review. This algorithm is In addition used in [38] and to build this tool.

Algorithm 2: CNSA-MSA-SAT" determining the polarity"
Input:
***R*:** Review/Document Text
***T*:** the set of the Opinion tokens
***PD*:** the set of Positive Sentiment Dictionary
***ND*:** the set of Negative Sentiment Dictionary
Output:
$P=\{Pos, Neg, NU, U\}$, *where* Pos: *Positive,* Neg: *Negative,* NU: *Neutral, U:Undetermined*

$$P_{pos\%} = \frac{\sum_{i=1}^{n} Pos - TF_i}{\left(\sum_{i=1}^{n} Pos - TF_i\right) + \left(\sum_{i=1}^{n} Neg - TF_i\right)} \times 100$$

$$P_{neg\%} = \frac{\sum_{i=1}^{n} Neg - TF_i}{\left(\sum_{i=1}^{n} Pos - TF_i\right) + \left(\sum_{i=1}^{n} Neg - TF_i\right)} \times 100$$

Initialization:
Pos-TF=0, *where Pos-TF is the term frequency for positive sentiments*
Neg-TF=0, where *Neg-TF* is the term frequency for negative sentiments

Begin
1: For each $t_i \in T$ do
2: Search for t_i in PD where $t_i \in T$
3: If $t_i \in$ PD then
4: Pos-TF←Pos-TF + 1
5: Else
6: Search for t_i in ND where $t_i \in T$
7: If $t_i \in$ ND then
8: Neg-TF ←Neg-TF +1
9: End For
10: If (Pos-TF ≥2) and (Pos-TF >Neg-TF) then
11: P←Pos
12: Return $P_{pos\%}$
13: End If
14: If (Neg-TF ≥2) and (Pos-TF <Neg-TF) then
15: P ←Neg
16: Return $P_{neg\%}$
17: End If
18: If (Pos-TF = Neg-TF) and (Pos-TF = 0)then
19: P←NU
20: Display P
21: If (Pos-TF = Neg-TF) and (Pos-TF ≠ 0)then
22: P←U
23: Display P
24: End If
End

C. Strength/Intensity Algorithm

This section presents the algorithm used to determine the Strength/Intensity of evaluated Arabic reviews this tool. Each evaluated Arabic review is considered either as: strong positive, strong negative, weak positive, weak negative, neutral or an undetermined review.

Algorithm 3: CNSA-MSA-SAT" determining the intensity and the polarity depend on the weight "
Input:
R = Review/Document Text
T = the set of the Review tokens
PD = the set of Positive Sentiment Dictionary
ND = the set of Negative Sentiment Dictionary
Pos-Weight =the value of Positive Weight
Neg-Weight = the value of Positive Weight
Output:
I *={SP,SN,WP,WN,NT,U} where SP: strong positive, SN: strong negative WP: weak positive, WN: weak negative, NT: neutral, U: undetermined*
Max_Pos_score *= Max of the set of Pos-Weight of the positive sentiments*
Max_Neg_Score *= Max of the set of Neg-Weight of the negative sentiments*

Begin:
1: For each ti Є T do
2: Find ti Є PD where ti Є T
3: Find Pos-Weight
4: End for
5: Find Max_Pos_score
6: For each ti Є T do
7: Find ti Є ND where ti Є T
8: Find Neg-Weight
9: End for
10: Find Max_neg_Score
11: Else
12: Return Max_Pos_score= Max_Neg_Score=0
13: If (Max_Pos_Score>Max_Neg_Score and Max_Pos_Score>= 5) then
14: I←SP
15: Return I
16: End If
17: If (Max_Neg_Score>Max_Pos_Score and Max_Neg_Score>= 5) then
18: I←SN
19: Return I
20: End If
21: If (Max_Pos_Score>Max_Neg_Score and Max_Pos_Score< 5) then
22: I←WP
23: Return I
24: End If
25: If (Max_Neg_Score>Max_Pos_Score and Max_Neg_Score< 5) then
26: I←WN
27: Return I
28: End If
29: If (Max_Pos_Score = 0 and Max_Neg_Score =0) then
30: I←NU
31: Return I
32: End If
33: If (Max_Neg_Score =Max_Pos_Score and Max_Pos_Score ≠0) then
34: I←U
35: Return I
36: End If
End

V. EXPERIMENTAL RESULTS

This section presents and discusses experimental results. The conducted tests aim to evaluate the effectiveness of the developed opinion mining tool to identify domains, subjectivity, polarity and strength of evaluated Arabic reviews. The results in this section are presented in the following three subsections.

The first subsection is presenting the results of the tests related to subjectivity classification, the second subsection is presenting the results related polarity, and the third subsection is presenting results related to the identification of the intensity of each evaluated Arabic review to the tool.

In the experiments for all classifiers, we used 66% of the dataset as a Training Dataset and 34% as a Testing Dataset.

We used the following four metrics to evaluate the quality of the tool in terms of opinion decision:

Accuracy: Is the degree of closeness that a measured value represents the correct value.

The Accuracy is defined by the formula (5.1):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{..... (5.1)}$$

Where TP is a true positive rate, FP is a false positive rate, TN is a true negative rate, and FN is a false negative rate [36].

Error: Is the degree of closeness that a measured value represents the incorrect value [39].

The formulas of the other two performance metrics (Recall and Precision) are shown next.

The Recall is defined by formula (5.2) [40]:

$$Recall_i = \frac{TP}{TP + FN} \quad \text{.....(5.2)}$$

The Precision is defined by the formula (5.3) [40]:

$$Precision_i = \frac{TP}{TP + FP} \quad \text{..... (5.3)}$$

where TP is the number of documents correctly classified as belonging to a class i ("true positive"), FP is the number of documents falsely classified as belonging to a class i ("false positive") and FN is the number of documents falsely classified as not belonging to a class i ("false negative") [37].

A. Subjectivity Results

This subsection presents the results of the tests conducted on the tool to evaluate its effectiveness to identify Arabic facts and opinions. A Naive Bayes Classifier proves it is more effective than others classification algorithms such as Decision Tree, K-NN, SVM to identify Arabic facts and opinions. Therefore it's adopted and used.

The overall accuracy shown in table 17 is 93.9%. In addition table 17 presents recall and precision values according to 5.2 and 5.3 formulas.

TABLE XVII. NAIVE BAYES SUBJECTIVITY RESULTS

Class	Accuracy	Error	Precision	Recall
Opinion	-	-	0.96	0.96
Fact	-	-	0.85	0.85
Dataset	93.9%	6.01%	0.93	0.93

B. Polarity Evaluation Result

This subsection presents an evaluation to accuracy of the tool to identify the polarity of each evaluated Arabic review. A K-NN Classifier proves it is more effective than others classification algorithms such as Decision Tree, Naive Bayes, and SVM to identify the polarities of different Arabic reviews. The overall accuracy shown in table 18 is 90%. In addition table 18 presents recall and precision values according to 5.2 and 5.3 formulas.

TABLE XVIII. K-NN POLARITY RESULTS

Class	Accuracy	Error	Precision	Recall
Positive	-	-	0.8	1
Negative	-	-	1	0.3
Neutral	-	-	1	1
Dataset	90%	10%	0.9	0.9

Table 18 shows that the effectiveness of tool to identify neutral Arabic reviews is optimum.

C. Intensity Evaluation Result

This subsection presents the results of the tests conducted on the tool to evaluate its effectiveness to identify the intensity of different Arabic reviews. Once again Naive Bayes Classifier proves it is more effective than others classification algorithms such as Decision Tree, K-NN, SVM to identify Arabic the strength of the evaluated Arabic review. Therefore Naive Bayes is adopted and used. The overall accuracy shown in table 19 is 96.6%. In addition table 19 presents recall and precision values according to 5.2 and 5.3 formulas.

TABLE XIX. NAÏVE BAYES INTENSITY RESULTS

Class	Accuracy	Error	Precision	Recall
Dataset	96.9%	3.1%	0.95	0.97

VI. CONCLUSION AND FUTURE WORK

This study presented a basic tool which can be used to analyze Arabic reviews and comments regardless of the type of the Arabic language (MSA or Colloquial) they used. In order to evaluate the proposed tool, we need a standard dataset to test its effectiveness.

We found that there is no standard dataset to be used. Therefore we collected Arabic reviews and comments. The collected Arabic reviews use only MSA and the first four Arabic Vernaculars presented in the section 1: Arabian Peninsula Arabic (Khaliji Arabic), Mesopotamian Arabic, Syro-Palestinian Arabic, and Egyptian Arabic. The proposed tool presented in this study is a lexicon-based tool. The collection of Arabic comments and review phase is followed by lexicon creation phase.

The lexicons used in this study are manually created, since they have manually extracted features, terms, and phrases from the collected reviews and comments. The tool is capable to identify the polarity, subjectivity, and strength/intensity of each evaluated Arabic review and comment. This study is based on 18 lexicons which built manually. Two general purpose lexicons were built to be used to identify polarity, and 16 domain-specific lexicons were built to be used to identify the polarity with eight different domains Technology, Books, Education, Movies, Places, Politics, Products, and Society. The last phase of this study includes an evaluation to the effectiveness of the tool.

The evaluation of this tool yields: a 93.9 % accuracy to classify the evaluated Arabic comments and reviews into their proper domains, a 90% accuracy to identify the real polarity of the evaluated Arabic comments and reviews, and a 96.9% accuracy to identify the strength/intensity of the evaluated Arabic comments and reviews. Tests on the tool reveal the reasons behind errors. The main reasons behind these errors are summarized by the use of spam reviews, spelling mistakes, and short comment length (One word).

We plan to enhance and extend this study by using a larger dataset which has more Arabic comments and reviews written in a wider range of Arabic Vernaculars. This tool is incapable to deal with Emoticons, chat language, Arabizi, so we plan to enhance this tool to be able to deal with these inputs. Future plans include adopting semantic techniques to identify polarity, subjectivity, and strength/intensity. In addition we plan to create lexicons automatically.

REFERENCES

[1] Arabic Speaking Internet Users and Population Statistics. http://www.internetworldstats.com/stats19.htm (2012, accessed January 2013).

[2] Varieties of Arabic. http://en.wikipedia.org/wiki/Varieties_of_Arabic (2013, accessed May 2013).

[3] B. Liu,“Sentiment Analysis and Subjectivity,” In Indurkhya N and Damerau F J (Eds) Handbook of Natural Language Processing, Chapman and Hall/CRC, Second Edition, 2010.

[4] B. Liu. “Sentiment Analysis and Opinion Mining. California,” US: Morgan & Claypool Publishers, 2012.

[5] T. Nasukawa, y. Yi,“Sentiment analysis: capturing favorability using natural language processin,” In: Proceedings of the 2nd international conference on Knowledge capture(K-CAP '03), New York, NY, USA. pp. 70-77, 2003.

[6] D. Kushal, S. Lawrence,D. Pennock,“Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” In: Proceedings of International Conference on World Wide Web (WWW-2003); 2003.

[7] Social networking service - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Social_networking_service (2013, accessed February 2013).

[8] Alexa - The Web Information Company. http://www.alexa.com (2013, accessed February 2013).

[9] Facebook Statistics and Metrics by Countries - Social bakers. http://www.socialbakers.com/facebook-statistics (2013, accessed February 2013).

[10] B. Hammo, H. Abu-Salem,S. Lytinen,“QARAB: A Question Answering System to Support the Arabic Language,” In: Annual Meeting of the ACL Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic, pp. 1-11. 2001.

[11] M. Bautin, L. Vijayarenu,S. Skiena,“International Sentiment Analysis for News and Blogs,” In: Proceedings of the International Conference on Weblogs and Social Media, Seattle, Washington, 2008.

[12] M. Rushdi-Saleh, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López,J. M. Perea-Ortega, “Bilingual Experiments with an Arabic-English Corpus for Opinion Mining. Language,” In: Proceedings of Recent Advances in Natural Language Processing, Hissar, Bulgaria. pp. 740-745, 2011.

[13] Y. Almas,K. Ahmad,“A note on extracting ‘sentiments’ in financial news in English, Arabic & Urdu,” In: The Second Workshop on Computation, al Approaches to Arabic Script-based Languages, Linguistic Society of America 2007 Linguistic Institute, Stanford University, Stanford, California., Linguistic Society of America, pp. 1–12, 2007.

[14] A. El-Halees,“Arabic Opinion Mining Using Combined Classification Approach,” In: Proceedings of the InternationalArab Conference on Information Technology,Zarqa, Jordan, 2011.

[15] B. Pang, L. Lee, S. Vaithyanathan,“Thumbs up? Sentiment classification using machine learning techniques,” In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, Penn. pp. 79-86, 2002.

[16] J. Yi, T. Nasukawa, R. Bunescu, W.Niblack,“Sentiment Analyzer: Extracting Sentiments about a given Topic using Natural Language Processing Techniques,” In: Proceedings of the Third IEEE International Conference on Data MiningIEEE Computer Society, pp. 427-434, 2003.

[17] S-K. Kim, E. Hovy,“Determining the sentiment of opinions,” In: Proceedings of the 20th international conference on computational linguistics(COLING 2004), Geneva, Switzerland. pp. 1367–1373, 2004.

[18] M. Elhawary,M. Elfeky,“Mining Arabic Business Reviews,” In: Proceedings of the2010 IEEE International Conference on Data Mining Workshops; pp. 1108-1113, 2010.

[19] M. Abdul-Mageed, M. Diab,M. Korayem,“Subjectivity and sentiment analysis of modern standard Arabic,” In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, Portland, Oregon, USA. pp. 587-591, 2011..

[20] T. Wilson, J. Wiebe, P.Hoffmann,“Recognizing contextual polarity in phrase-level sentiment analysis,” In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing(HLT '05), Stroudsburg, PA, USA, pp. 347-354, 2005.

[21] T. Wilson, J. Wiebe,R. Hwa,“Recognizing Strong and Weak Opinion Clauses,”Computational Intelligence Journal,vol. 22, pp. 73-99, 2006.

[22] A. Abbasi, H. Chen,A. Salem,“Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums,”ACM Transactions on Information Systems,vol. 26, no. 3 pp. 1-25, 2008.

[23] G. Paltoglou,M. Thelwall, “”A study of Information Retrieval weighting schemes for sentiment analysis,” In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10) Association for Computational Linguistics, Stroudsburg, PA, USA. pp. 1386-1395, 2010.

[24] T. Inui, M. Yamamoto ”Applying Sentiment-oriented Sentence Filtering to Multilingual Review Classification,” In: Proceedings of theWorkshop on Sentiment Analysis where AI meets Psychology(SAAIP), IJCNLP, Chiang Mai, Thailand, pp. 51–58, 2011.

[25] C. Banea, R. Mihalcea, J. Wiebe, S. Hassan,“Multilingual subjectivity analysis using machine translation,” In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 127–135, 2008.

[26] C. Banea,R. Mihalcea, J. Wiebe,“Multilingual subjectivity: Are more languages better?,” In: Proceedings of the 23rd International Conference on Computational Linguistics,pp.28–36, 2010.

[27] A. Balahur, A. Montoyo, “A feature dependent method for opinion mining and classification,” In: Proceedings of the IEEE international conference on Natural Language Processing and Knowledge Engineering(NLP-KE '08), Beijing, China, pp. 1-7, 2008.

[28] Y. Choi, Y. Kim,S-H. Myaeng,“Domain-specific Sentiment Analysis using Contextual Feature Generation,” In:Proceedings of the1st

International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement(TSA'09), Hong Kong – China. pp. 37-44, 2009.

[29] A. Ortiz, C. Hernández,R. García, "Domain-neutral, Linguistically-motivated Sentiment Analysis: A Performance Evaluation, Evaluación de unsistema de análisis de sentimiento basado en conocimiento,"Machine Learning, pp. 361-369, 2007.

[30] A. Al-Subaihin, H. Al-Khalifa,A. Al-Salman,"A Proposed Sentiment Analysis Tool for Modern Arabic Using Human-Based Computing," In: Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services(iiWAS '11), New York, NY, USA. pp. 543-546, 2011.

[31] M. Al-Kabi, N. Al-Qudah, I. Alsmadi, M. Dabour M., H. Wahsheh, (2013). "Arabic / English Sentiment Analysis: An Empirical Study," The 4th International Conference on Information and Communication Systems (ICICS 2013), Irbid, Jordan, pp. 1-6, 2013.

[32] R. Khasawneh, H. Wahsheh, M. AL-Kabi, I. Alsmadi,"Sentiment Analysis of Arabic Social Media Content: A Comparative Study," The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), London, UK., pp. 101-106 , 2013.

[33] M. AL-Kabi, N. Abdulla, M. Al-Ayyoub,"An Analytical study of Arabic Sentiments: Maktoob Case Study," The 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), London, UK, pp. 89-94 2013.

[34] Y. Fang, S. Parthasarathy,F. Schwartz,"Using Clustering to Boost Text Classification," In: Proceedings Of the IEEE International Conference on Data Mining, California, USA. pp. 123-127, 2001.

[35] J. Han,M. Kamber,"Data Mining: Concepts and Techniques, second edition,"Morgan Kaufmann Publishers, CA, San Francisco, 2006.

[36] Opinion Mining, Sentiment Analysis, and Opinion Spam Detection. http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html (2013, accessed January 2013).

[37] A. Chaudhuri, "Emotion and Reason in Consumer Behavior. Amsterdam," Elsevier Butterworth-Heinemann,2006.

[38] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh,M. Haidar,"An Opinion Analysis Tool for Colloquial and Standard Arabic," In: Proceedings of the fourth International Conference on Information and Communication Systems (ICICS 2013), Irbid, Jordan, pp. 1-5, 2013.

[39] I. Witten,E.Frank, "Data Mining: Practica Machine Learning Tools and Techniques,"Morgan Kaufmann Series in Data Management Systems, second edition, Morgan Kaufmann (MK) 2005.

[40] G. Paltoglou,M. Thelwall,"Twitter, MySpace, Digg: Unsupervised Sentiment Analysis in Social Media,"ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 4, pp. 1-19, 2012.

AUTHORS PROFILE

Mohammed Naji Al-Kabi obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his master's degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq (1981). Mohammed Naji AL-Kabi is an assistant Professor in the Faculty of Sciences and IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, Nahrain University and Mustanserya University in Iraq. He also worked as a part time lecturer at Jordan University of Science and Technology, and Sunderland University. AL-Kabi's research interests include Information Retrieval, Web search engines, Data Mining, Social media, Natural Language Processing and Software Engineering. He is the author of more than 66 peer reviewed articles in these topics. His teaching interests focus on information retrieval, Web programming, data mining, DBMS (ORACLE & MS Access).

Amal H. Gigieh got a Master Degree in Computer Information System CIS (2012) from Yarmouk University-Jordan, the bachelor degree in Information Technology IT (2003) from the AL-Balqa' Applied University-Jordan, worked as a computer lab supervisor and a computer practical courses lecturer in AL-Balqa' applied university. Gigieh also worked as online exams manager in the same univesity. Her research in Opinion Analysis and Mining

Izzat Alsmadi. An associate professor in software engineering. Born in Jordan 1972, Izzat Alsmadi has his master and PhD in software engineering from North Dakota State University (NDSU), Fargo, USA in the years 2006 and 2008 respectively. His main areas of research include: software engineering, testing, metrics, and information retrieval.

HeiderWahsheh. Born in Jordan, in August 1987, he obtained his Master degree in Computer Information Systems (CIS) from Yarmouk University, Jordan, 2012. Since 2013 Mr. Wahsheh starts working as a lecturer in the college of Computer Science at King Khalid University, Saudi Arabia. His research interests include: Information Retrieval, Data Mining, and Mobile Agent Systems.

Mohamad Haidar. Born in Jordan, in February 23 1991. Mr. Haidar obtained his bachelor degree in Computer Information Systems (CIS) from Yarmouk University, Jordan in 2012. Now, he works in Brandtologie company (Amman – Jordan) as a web developer.

Cloud and Web Technologies: Technical Improvements and Their Implications on E-Governance

Danish Manzoor
Department of E-Services,
Deanship of InformationTechnology
Northern Border University, Arar,
Kingdom of Saudi Arabia

Ashraf Ali
Information Technology Unit,
Foundation Program, Dhofar University,
Salalah Sultanate of Oman,

Dr. Ateeq Ahmad
Department of Computer Science,
Faculty of Science,
Northern Border University, Arar
Kingdom of Saudi Arabia

***Abstract*---Cloud computing technology helps to improve ICT based services like e-governance execution and create new business opportunities and their implementation. Cloud computing is an evolution of web based internet application and describes an advance consumption, supplement and delivery model for Information Technology and ICT services based on the global network. This enables allocation of resources and costs across a large pool of users while providing on-demand services with dynamic scalability. So we can say that a technology that has the capability and potential to offer solutions for e-governance is cloud computing. Cloud computing provide service-oriented access to users least compromising on security. In today's era software and their services are biggest cost concern for the implementation of IT environment in an organization. Cloud has the capability to reduce the cost in dramatic way for the all kind of the organization even it is small scale Industry or a big corporate organization. This makes Cloud an excellent platform to host e-governance services and application. The basic intention of this paper is whatever improvement happening in cloud technology and in web technology mention sub sequentially. If we apply them in existing e-governance application running under various department then we can minimize the some of the most basic affected components of application software like cost of the software in its execution, optimal time for usage and running of application software, storage capacity for storing of the data and network infrastructure used for the functioning of the application software.**

Keywords*---Cloud; Web Semantics; Burst; SaaS; PaaS; IaaS; G-Cloud; M-Cloud; TCO; ICT, E-Governance; ITIL; Cloud legacy; Legacy cloud system silos; Lock-In; RDF; XML; URI

I. INTRODUCTION

Cloud computing is a new business paradigm in a service oriented model, delivering business applications and other IT resources as services over the internet. Cloud computing is shared environment, multitenant and over-the-internet based service delivery model [29]. Cloud provides an attractive alternative to the common man as well as for organizations to transact and do business online. Another variant of cloud computing service model is based on the services offering. Cloud provides infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (Saas). A suitable combinational model for e-governance can be selected based on the desire. On the other hand, technology service providers are competing to provide cost effective and innovative solutions to host services on cloud. There are many providers partnering with governments to build and operate e-governance platforms on cloud. There are many vendors like Google, Rackspace and Amazon in US and Europe who are providing readymade platform for hosting such initiatives [4] [20][21] [29][30].

TABLE 1. PERCENTAGE OF STATE AND FEDERAL CIOS REPORTING STATUS OF CLOUD COMPUTING IN THEIR GOVERNMENT

No	Status of Cloud Computing	Percentage	
		State	*Federal*
1	Investigating the use of cloud computing but as yet not taken action	54	22
2	Running an active project to move portions of our computing infrastructure to a cloud computing environment	21	54
3	Undertaking a cloud computing pilot for a portion of our computing requirements	13	16
4	No formal plans to use cloud computing	5	8
5	Have been using cloud computing for years but under another name	5	NA
6	Other	3	NA

Status of Cloud computing in their governments [21]

Similarly there are few more vendors like Microsoft, IBM and HP are some key players engaged in developing SaaS based applications for governance initiatives. E-governance architecture calls for scalability and interoperability given the various interconnects that need to deployed [26][29]. Typically a cloud infrastructure is built with virtualization as key foundation component at all possible levels like server, storage, application, etc. This feature supports all non-functional requirements of e-governance applications seamlessly [4] [12][20] [21] [30].

II. CLOUD SERVICE MODELS

Cloud services models are as follows [4] [20][21] [30].

- Cloud Software as a Service (SaaS)
- Cloud Platform as a Service (PaaS)
- Cloud Infrastructure as a Service (IaaS)

If we see below diagram then we can easily understand. What are the actual differences between these models of the cloud?

Cloud Services Models		
Software as a Service	***Platform as a Service***	***Infrastructure as a Service***
Business operations over a network	Deploy customer-created applications to a Cloud	Rent processing, storage, network, other computing resources
"SaaS"	"PaaS"	"IaaS"

A. . Software AS A Service (SaaS)

SaaS is a complete online IT solution built and maintained by the SaaS provider.

Examples: Sales & Customer Database Services in the Cloud asSalesforce.com and Collaboration Services as Yammer.

There are many mobile Apps don't store data on the phone but are in fact SaaS. For example: Dropbox in a cloud solution that stores your data in another cloud (Amazon). There are many web applications that store our data like Prezi and Google Apps are also SaaS. SaaS can be utilized for various E-Governance services (G2B, G2G, G2E and G2C) [2]. The basic long-term vision of SaaS is centered around separating maintenance, software control and ownership from its actual use. SaaS providers will often not only provide one service of a certain software suite, but might typically provide a bundle of services, while models are possible in which SaaS services are provided by the company who has developed the software. By delivering services for set of software, the idea is that many of the present limitations constraining its utilization, deployment and evolution can be minimize. The shift in offering services also has implications for the SaaS providers' revenue models. Traditionally, the customer buys a license to use an application and installs it on their owned or controlled hardware [31].

Over time new updates can be including security patches installed other control and update activities. By buying a license, the customer gets unlimited usage of the software. Per user or per installed system additional licenses might be required. In contrast in the SaaS delivery models the user does not buy a lifetime license. The user need to pay only a certain amount for the software running on a third-party server and loses access when he ceases payment. This can be achieved as pay per use or pay per duration basis. The payment can either be charged as a pre-paid subscription or on a pay-as-you-go basis.

B. Platform AS A Service (PaaS)

PaaS is a platform on which you are expected to develop your own solution. Imagine a blank web site or empty database built and maintained by the PaaS provider.

Examples: Database – Oracle Cloud DB and Web site – Amazon AWS.

PaaS providers are able to manage the lower layers and relieve PaaS subscribers of the responsibility to select, install, maintain and operating the platform components. Infrastructure charges are implicitly available in PaaS offerings because PaaS consumes infrastructure resources in some form and the infrastructure charges are bundled in the rates charged for the PaaS execution environment resources (e.g., CPU, bandwidth, storage) [28] [31].

C. Infrastructure AS A Service (IaaS)

IaaS is when the service only provides raw components of a solution. Imagine a company that only provides you data storage and nothing else means you just taking services as for storing the data.

Examples: Laptop backup Mozy and Storage Amazon S3.

The capability provided to the consumer is to provision of processing, storage, networks and other fundamental computing resources where the consumer is able to deploy and run arbitrary software that can include operating systems and applications. Consumer does not manage or control the core cloud infrastructure but has control over operating systems, storage, deployed applications and possibly partial control of select networking components (e.g., host firewalls)[28][31].

III. CLOUD DEPLOYMENT MODELS

Independent of the specific cloud services model used, NIST defined four deployment models [4] [6] [20][21].

What are those models we can see in below pictorial diagram? It is classifying the different deployment model run under the cloud.

Cloud Deployment Models	
Private Operated for a single organizations	***Community*** Shared by several organizations supporting a specific community
Public Available to the general public or large industry group owned by an organization selling Cloud services	***Hybrid*** Two or more Clouds that remain unique but are bound by technology that enables data and application portability

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

There are obvious co relation between the features of the cloud computing and their deployment models. See in below diagram how features of the cloud computing varies and demonstrate the different behavior under their deployment models.

Actually it is basic comparison between the legacy system which is non cloud under running in different deployment model and cloud system running under the cloud deployment model. We can see features like cost, liability and assurance of availability is more mature compare to old and exiting legacy system in "Fig. 1".

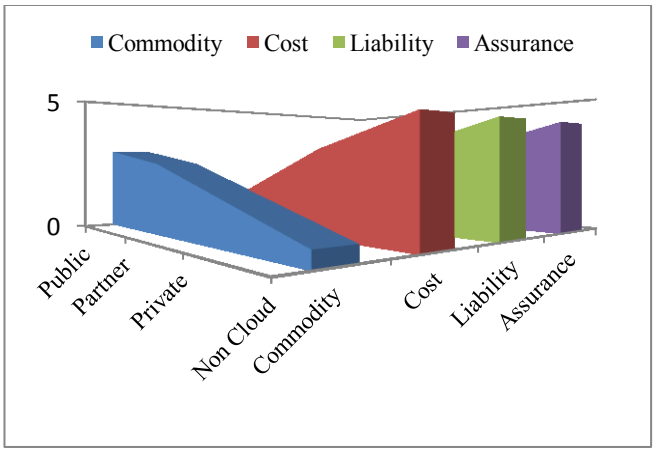


Fig. 1. Features of Public, Community, Private Clouds and Non-Cloud

Government continues to explore innovations for optimizing and harmonizing IT services across state government specially government business services as well. Multiple cloud computing deployment models ranging from internal private clouds to government community clouds to government line of business community clouds to public clouds [30]. As government explores its options, and the various scenarios for engaging cloud computing services, consideration must be given to a number of issues. These are including but cannot limit to the following.

- Actual total cost of a service
- User fees / access fees
- Exit strategy / switching strategy
- Potential for data breach
- Legal liability – assigned and assumed
- Access and use of government data – including emails and the content of email attachments to Service provider
- Provider's economic model for pricing which may include reselling government data, or reselling analytics about the data
- Location of the data and applications
- Jurisdictional issues related to the physical, virtual, and legal location of data and applications
- Proximity and threats from other data tenants – intended and unintended
- Risks related with the multi-tenant or multiplexing physical infrastructure environment
- The list continues to grow thus the need for agile, dynamic enterprise architectures

IV. Related Work

There are several organizations in US and India working for the implementation of the latest cloud and web technology for the e-governance model. There are several hurdles and blocks at operation level and technology levels. Some of them are like the structure of the data used for the e-governance operation if the structure of the data is in manageable format then it is easy to make it more relational. Government of India is also working for the proper and more logical structure of the data they are designing data in more appropriate and suitable format so it can be easy to manageable and can be integrated smoothly. Similarly in the area of the shared services governments are trying to make it more identical so it can be used by the multiple organizations. Suppose we have passport number then if a service is implemented based on the passport number then it can be benefitted by multiple departments like if visa department want to verify the visa details then by entering the passport number they can get all the related information of the visa. In the same way immigration department can check the authenticity and validity of the passport and visa by using the shared services provided by passport department. In more broad way suppose police department want to check criminal verification provided on the particular passport then they can verify it just entering the passport number under the shared services provided by passport department. So data management and shared services are the some of the major area where most of the government are working for e-governance program.

V. Technological Improvement In Cloud and Web

A. Improvement in Cloud

The Basic idea behind the cloud model is that anything that could be done in computing whether on an individual PC or in a corporate data center from storing data to collaborating on documents or crunching numbers on large data sets can be shifted to the cloud. Certainly cloud computing facilitates a new platform and location-independent perspective on how we are communicating, collaborating and working. So long as you can access the web you are able to work when and where you wish.

By using fast reliable Internet connectivity and computer power it does not matter where the document and e-mail or the data comes from for user sees on the screen. Cloud computing enables providers to use distant data centers for cloud computing. While some have predicted the end of the PC era with the rise of the cloud computing model many believe that most organizations and even individuals will continue to make use of traditional PCs and Laptops even if more and more of their use will be to access the cloud. For individuals, cloud computing means photo sharing, accessing web-based email and productivity software much of it for free [28]. For organizations shifting to the cloud means having the ability to contract for computing services on-demand rather than having to invest to host all the necessary software, hardware and support personnel necessary to provide a given level of services [28]. For governments, the value proposition of the cloud is mainly appealing, given both changing demands for Information Technology and challenging economic conditions [4][20] [21] [30] [31].

The essence of rolling out e-governance services is to ensure reach and high availability of the service and both these tenets have to be provided by the technology platform [29]. Surely, cloud computing provides both. Reach is ensured as it is hosted on internet. With the basic construct of the infrastructure, availability of the platform is ensured by

virtualization technologies with these underlying technologies and storage. The threat of having data read during transmission can be mitigated through encryption. Encryption in transit protects data as it is being transmitted to and from the cloud service. Encryption protects data that is stored at the service provider. Encrypting data in an on-premises cloud service on-ramp system can provide both kinds of encryption protection [17] [20] [21] [22].

B. Improvement in Web

The social work profession has acceptance of Web 2.0 technologies. One of the first references to Social Work 2.0 was made in "The New Social Worker" magazine which was started by Linda May Grobman in spring of 1994.Online publication continues to explore the application of web technology within the social work community. The first article of an ongoing social work 2.0 series was entitled "Caring Bridge: A Valuable Tool for Social Workers and those with Critical Illness" written by Karen Zgoda. It was followed by a column entitled Social Work There's a Blog for that by Karen Zgoda. The article noted that blogging was rapidly becoming a phenomenon within the social work community. Students and professionals had begun chronicling their career development as well as sharing information from their respective practice areas [17].In 2007, Jonathan Singer, started The Social Work Podcast which provides information about all the things in social work followed by a more formalized outline of the meaning of Social Work 2.0 in 2009 [15].

Web 3.0 or Semantic Web – The definitions of Web 3.0 vary significantly. Some believe its most important features are the Semantic Web and personalization. Concentrating on the computer elements Conrad Wolfram has argued that Web 3.0 is where "the computer is generating new information" rather than humans [32].

Andrew Keen, author of The Cult of the Amateur considers the Semantic Web an "unrealizable abstraction" and observes Web 3.0 as the return of experts and authorities to the Web. As example he points to Bertelsmann's deal with the German Wikipedia to produce an edited print version of that encyclopedia. CNN Money's Jessi Hempel looks forward Web 3.0 to emerge from new and innovative Web 2.0 services with a beneficial business model [33].

We can see how much improvement and advancement happened in web technology. By defining these technologies according to their versions, we can describe them briefly below.

a) Web 1.0

That Geocities & Hotmail era was all about read-only content and static HTML websites. People preferred navigating the web through link directories of Yahoo! and dmoz [27][28] [31].

b) Web 2.0

This is about user-generated content and the read-write web. People are overriding as well as contributing information through blogs or sites like Flickr, YouTube, Diggetc. The line dividing a consumer and content publisher is increasingly getting blurred in the Web 2.0 era[27][28][31].

c) Web 3.0

This will be about semantic web (or the meaning of data) personalization (example iGoogle) intelligent search and behavioral advertising among other things [27][28][31].

The semantic web is the next Big Thing. The Semantic Web is "a web of data". In Semantic Web basically HTML describes documents and RDF describes things. Why talk about a "page" in terms of style and links when you know a book has chapters and a CD has tracks?

d) Challenges

- Human error and system abuse
- Selfishness, why take the time to teach the machine how to teach me?

e) Opportunities

- Young people love their RSS feeds
- Data portability and sharing standards are 2008
- If we were all so selfish, Web 2.0 would have never worked (etc Wikipedia)

f) Bottom line (why we should care)

It's not just for academics anymore. People will use the tools that they find most useful. Where the people are advertisers they must follow.

C. Introducing Web 3.0

Web 3.0 creates a big collection of databases which can be connected on demand. The semantic web is about the meaning of data. What could it mean? It means as following

- How will be our information in organized?
- Will we still do the "surfing" or will the machine surf for us?
- The look of web will be the same for me as it does for everyone else?
- What are the technologies that will become usual?

VI. FUTURE WORK

This is the era of the service, when anybody taking any kind of the services then they do not want to take pain of the managing the services that's why people like the concept of the cloud technology and since there are very rapid improvement in web technology that why it is making it more sharpen because by using these web technology like semantic web technology we can get better result in term of personalization and customization of the interface. We can get our data result in more structured and order format. By using java script and json we can design and manage a complete application and we can manage a huge amount of the data in very optimize timing. In the legacy system for using services we need to manage the many areas. Suppose we want to create the profile of the employees in an organization then we need to deploy a complete application for the profile creation of the employees. Now for this application we need resources in term of the software, hardware, storage and network infrastructure. So it is costing a huge operational and maintenance cost for

the just creation of the profile of the some employees in an organization. Suppose we take this service from some cloud provider company who is providing "SaaS" services then just by paying some amount for this service we can use the service without investing the money in other related area of the infrastructure like software, hardware, storage and network.

VII. CONCLUSION

The attractive features of Cloud Computing like on demand self-service, network entrance, sites self-governing reserves of pool components, quick enhancement and calculated pre-planned services for the effective e-governance. As various researchers estimated that till year 2015 50% Cloud Computing is used in e-governance for its proper execution [29].

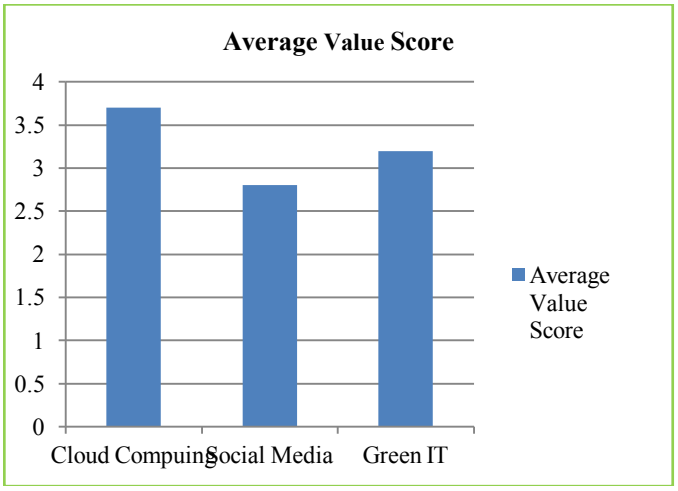


Fig. 2. Average value score given by CIOs to emerging technologies

The advanced development in the ICT field point out that government all over the world will make use of cloud technology. Cloud Computing can ensure the changes in the government scenarios and their policies for the proper working as well as the execution of their e – services. E-Governance system requires entities like hardware, software, service, network, management, business, policy, security etc to survive and function properly. This framework is created to suit the needs of any organization which is keen to develop an enterprise architecture framework to promote reusability agility and flexibility in its IT systems. Framework contains business architecture and functional (software) architecture built on the basis of business requirements. Technical (Software) architecture encompasses application data and technology architecture. The framework also contains management disciplines such as Change Management, Strategic Control and Program (Project) Management. The overall aim is in building and delivering application to bring effectiveness, transparency, efficiency and accountability. The above research is that we can get the better services than traditional computing with reduced cost with the help of cloud computing. The cloud model will ultimately serve to transform - in a big way - not just government Information Technology, but Information Technology in the corporate world as well. The transition still will take time But cloud computing is one of the best option to implement or enhance the Government services in education, healthcare and social up liftmen of the citizens of the developing countries.

REFERENCES:

[1] A. Joshi (Pacific University (Paher), Udaipur, Rajasthan) "Integration of Virtualization with Cloud Computing: Challenges for Government & IT Industry" Oct, 2011.

[2] A. R. Roberts "The Role and Regulations for Technology in Social Work Practice and E-Therapy: Social Work 2.0" New York, U.S.A.: Oxford University Press. ISBN 978-0-19-536937-3.

[3] Cloud Computing Use Cases Group (Google group). This group is devoted to defining common use cases for cloud computing - http://groups.google.com/group/cloud-computing-use-cases.

[4] CIO - NASCIO Cloud Computing Series "Capitals in the Clouds The Case for Cloud Computing in State Government Part I: Definitions and Principles" - National Institute of Standards and Technology (NIST) http://www.nist.gov/itl/cloud/index.cfm June, 2011.

[5] CIO - NASCIO Cloud Computing Series "Capitals in the Clouds The Case for Cloud Computing in State Government Part II: Challenges and Opportunities to Get Your Data Right" - National Institute of Standards and Technology (NIST) http://www.nist.gov/itl/cloud/index.cfm June, 2011.

[6] CIO - NASCIO Cloud Computing Series "Capitals in the Clouds Part III – Recommendations for Mitigating Risks: Jurisdictional, Contracting and Service Levels" - National Institute of Standards and Technology (NIST) http://www.nist.gov/itl/cloud/index.cfm June, 2011.

[7] CIO - NASCIO Cloud Computing Series "Capitals in the Clouds Part IV – Cloud Security: On Mission and Means" - National Institute of Standards and Technology (NIST) http://www.nist.gov/itl/cloud/index.cfm June, 2012.

[8] CIO survey report by TechAmerica and Grant Thornton LLP titled "Transparency and Transformation through Technology: TechAmerica's Twentieth Annual Survey of Federal Chief Information Officers," http://www.grantthornton.com/ Feb, 2010.

[9] Cloud storage from Wikipedia, the free encyclopedia - http://en.wikipedia.org/wiki/Cloud_storage.

[10] Dr. N. Bhatt (M.L. Sukhadia University Udaipur), Prof. A Aggarwal (Vice Chancellor, Gujarat Technological University, Ahmedabad, India) "E-Governance Techno-Behavioural Implications: Government Transformation: Agenda for E-Government 2.0" Oct, 2011.

[11] Dr. R.S.R Prasad, V.R.R Atukuri, "Cloud Computing Technology for Effective e-Governance", Professor, International Business Management, AcharyaNagarjuna University. Associate Professor, Computer Science and Engineering Department, MalineniLaksmaiah Women's Engg College Vol. 3 (1), 2012, 3241-3244.

[12] "Interoperability Framework for e-Governance (IFEG) in India" Draft Version 0.5 Government of India Department of Electronics and Information Technology Ministry of Communications and Information Technology Nov, 2012.

[13] J. Bhakdi "A comprehensive web 3.0 overview" - Focus on user generated business http://www.sophotec.com Feb, 2009.

[14] J. Bughin and M. Chui "The rise of the networked enterprise: Web 2.0" http://www.mckinsey.com/insights Dec, 2010.

[15] J. Hempel "Web 2.0 is so over. Welcome to Web 3.0" http://money.cnn.com/2009/01/07/technology/hempel_threepointo.fortune/index.htm Jan, 2009.

[16] J. Marijn, J. Anton - Challenges for adopting Cloud-based "Software As A Service (SaaS) in the public sector Delft University of Technology, Jaffalaan 5, 2628 BX Delft, the Netherlands, 150 Minories, London, EC3N 1LS, United Kingdom April, 2011.

[17] K. Andrew "Web 1.0 + Web 2.0 = Web 3.0" http://www.ajkeen.com/2008/04/25/web-10-web-20-w/ April, 2008.

[18] K. Zgoda, MSW, LCSW "SW 2.0: Social Work? There's a Blog for That". The New Social Worker Retrieved. http://www.4dgraphix.com/web-2-in-social-work.html June 16, 2012.

[19] L. Badger, T. Grance, R. Patt and C. Jeff Voas "Cloud Computing Synopsis and Recommendations" Recommendations of the National Institute of Standards and Technology, NIST Special Publication 800-

146, U.S. Department of Commerce Computer Security Division Information Technology Laboratory NIST Gaithersburg, MD 20899-8930 May, 2011.

[20] NASCIO on Data Governance "Data Governance - Managing Information As An Enterprise Asset: Part I - An Introduction" http://www.nascio.org/publications April, 2008.

[21] NASCIO on Data Governance "Data Governance Part II: Maturity Models – A Path to Progress" http://www.nascio.org/publications March, 2009.

[22] NASCIO on Data Governance "Data Governance Part III: Frameworks – Structure for Organizing Complexity" http://www.nascio.org/publications May, 2009.

[23] N. VijaykumarSETLabs Briefings Impact of Cloud Computing in Driving New Initiatives Vol. 9(2) 2011.

[24] R. A. Reddy, V. Varma IIIT- Hyderabad Imaginea "White Paper - Cloud and E-Governance" 2009.

[25] R.K. Das (NIC, Berhampur), M. Brahma (IBM India Systems & Technology Group, Bangalore) and A.K. Misro (Berhampur University, Orissa) "A Case Study - Cloud Computing for Economic Optimization in e-Governance" Oct 2011.

[26] R. Sharma, A. Sharma and R. R. Singh E-Governance & Cloud Computing: Technology Oriented Government Policies Vol. 2(2) Feb 2012.

[27] Semantic Web -http://www.w3.org/standards/semanticweb. U. Sivarajah and Z. Irani "Exploring The Application Of Web 2.0 In E-Government: A United Kingdom Context" May 8th – 9th 2012, Brunel University, University Kingdom.

[28] V. Gaur(Birla Institute of Technology, Mesra Ranchi, Jaipur, Rajasthan), P. Dhyani and O.P. Rishi (Rajasthan Central University, Rajasthan) "Cloud Services for Consumer in Federated Environment– Simple View of Discovery and Monitoring" Oct, 2011.

[29] V. Kundra U.S. Chief Information Officer "Federal Cloud Computing Strategy" February 8, 2011

[30] Web 2.0 from Wikipedia http://en.wikipedia.org/wiki/Web_2.0.

Efficient Eye Blink Detection Method for disabled-helping domain

Assit. Prof. Aree A. Mohammed
Computer Science Department
School of Science, Univ. of Sulaimani
Sulaimani, Iraq

MSc. Student Shereen A. Anwer
Computer Science Department
College of Science, Univ. of Salahaddin
Erbil, Iraq

***Abstract*—In this paper, we present a real time method based on some video and image processing algorithms for eye blink detection. The motivation of this research is the need of disabling who cannot control the calls with human mobile interaction directly without the need of hands. A Haar Cascade Classifier is applied for face and eye detection for getting eye and facial axis information. In addition, the same classifier is used based on Haar- like features to find out the relationship between the eyes and the facial axis for positioning the eyes. An efficient eye tracking method is proposed which uses the position of detected face. Finally, an eye blinking detection based on eyelids state (close or open) is used for controlling android mobile phones. The method is used with and without smoothing filter to show the improvement of detection accuracy. The application is used in real time for studying the effect of light and distance between the eyes and the mobile device in order to evaluate the accuracy detection and overall accuracy of the system. Test results show that our proposed method provides a 98% overall accuracy and 100% detection accuracy for a distance of 35 cm and an artificial light.**

Keywords—eye detection; eye tracking; eye blinking; smoothing filter; detection accuracy

I. INTRODUCTION

In the recent years due to the rapid advancement in the technology there has been a great demand of human computer or mobile interaction (HCI or HMI). Eye blink is a quick action of closing and opening of the eyelids. Blink detection is an important enabling component in various domains such as human computer interaction, mobile interaction, health care, and driving safety. For example, blink has been used as an input modality for people with disabilities to interact with computers and mobile phones [1].

In Viola [2] the chain of single-feature filters, Haar Cascade Classifier for identifying sub-region image is used. With the fast calculation of integral image technique, it can work in real time.

Eye tracking provides an almost seamless form of interaction with the modern graphical user interface, representing the fastest non-invasive method of measuring user interest and attention. While the mouse, keyboard, and other touch-based interfaces have long reigned as the primary input mediums associated with the field of human computer interaction, as advances continue to improve the cost and accuracy of eye tracking systems they stand poised to contend for this role [3]. An open and close eye template for blink pattern decisions based on correlation measurement is used in [4]. The method was specifically useful for people with severely paralyzed. A real-time eye blinking detection was proposed based on SIFT feature tracking with GPU based implementation [5].

An efficient method is proposed in [6]. A method is based on image processing techniques for detecting human eye blinks and generating inter-eye-blink intervals. A Haar Cascade Classifier and Camshift algorithms for face tracking and consequently are applied for getting facial axis information. Adaptive Haar Cascade Classifier from a cascade of boosted classifiers based on Haar-like features using the relationship between the eyes and the facial axis applied for positioning the eyes. The algorithm results show that the proposed method can work efficiently in real-time applications.

An EyePhone application which is developed in [7] is a system that capable of driving mobile applications/functions using only the user's eyes movement and actions (e.g., wink). EyePhone tracks the user's eye movement across the phone's display using the camera mounted on the front of the phone. The results indicate that EyePhone is a promising approach to driving mobile applications in a hand-free manner.

An efficient eye tracking system is presented in [1, 8] having a feature of blink detection for controlling an interface that provides an alternative way of Communication for the people who are suffering from some kind of severe physical disabilities the proposed system uses pupil portion for tracking the movement of eyes.

The outline of the paper is as follows. In section II, the proposed methods are presented. Section III studies the test results of a real time application for two cases: normal light and artificial light condition with and without using filter. Conclusions and future remarks are described in section IV.

II. PROPOSED EYE BLINK DETECTION

In figure 1 the major steps of the proposed Eye Blink to Control Mobile Phones EBCM are shown.

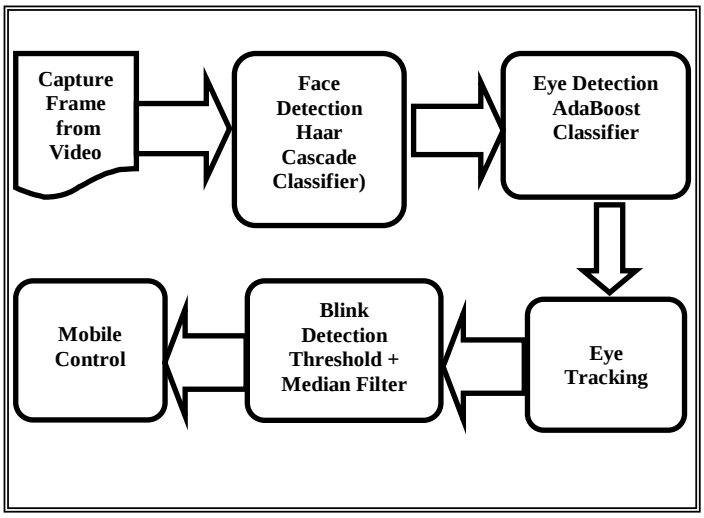


Fig. 1. EBCM general block diagram.

A. Frame Capturing

The first step of the proposed EBCM application is the initialization. After taking a short video of the participant's face using the front camera of the Samsung mobile. A *process Frame* method will be used to create the frames from the captured video. Afterwards the colored frames will be converted to gray scale frames by extracting only the luminance component as shown in figure 2.

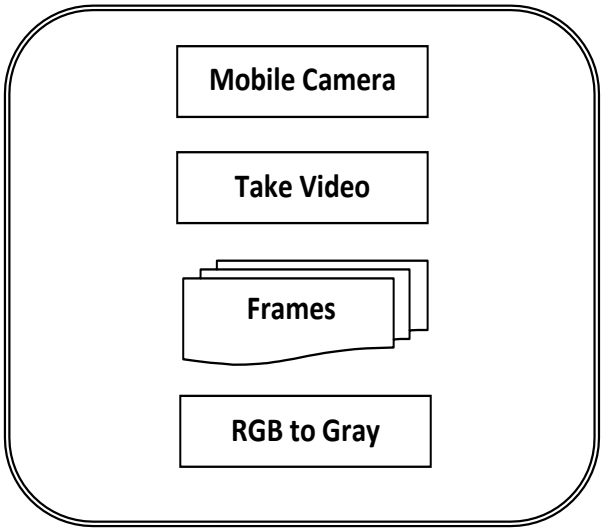


Fig. 2. Gray frames capturing.

B. Face Detection

The Haar classifier is used in EBCM algorithm for face detection. Haar classifier rapidly detects any object, based on detected feature not pixels, like facial feature. However, the area of the image being analyzed for a facial feature needs to be regionalized to the location with the highest probability of containing the feature. By regionalizing the detection area, false positives are eliminated. As the result, the face is detected and marked with color rectangle and will be used later to approximate an axis of the eyes for eye detection step.

C. Eye Detection

To detect the eye, first, the Haar cascade classifier should be trained, in order to train the classifiers, the AdaBoost algorithm and Haar feature algorithms must be implemented, two set of images are needed. One set contains an image or scene that does not contain the object.

The EBCM used all detected elements from the Haar Cascade Classifier, and the result show the detected eye in color rectangle as shown in figure 3.

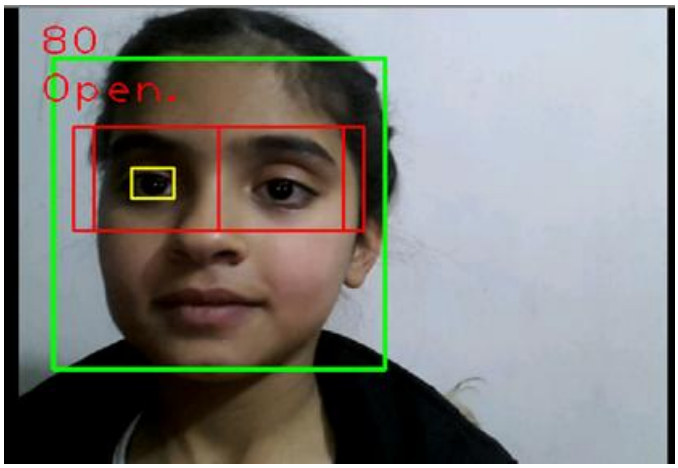


Fig. 3. Face and eye detection using Haar – like features

D. Eye Tracking

The corneal-reflection and pupil-center are the two eye's parts that are the most important parts to extract the features that will be used in EBCM method. These features help us in tracking the eyes movement. By identifying the center of the pupil and the location of the corneal reflection, the vector between them is measured. Besides, with further trigonometric calculations, point-of-regard can be found. The EBCM method succeeded in making the face and the eye's pupil moved together in the same direction synchronously and with the same direction. Let suppose that X is the human face which has been detected, P1 and P2 are two points related to the left eye, and they are moving synchronously with the movement of X as shown in figure 4.

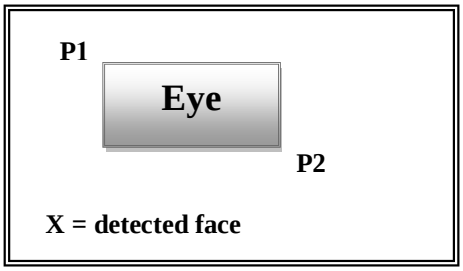


Fig. 4. Movement of face and the Eye's points

E. Eye Blinking

Eye blinking and movement can be detected with relatively high reliability by unobtrusive techniques. Though, there are few techniques discovered for the active scene where the face and the camera device move independently and the eye moves freely in every direction independently of the face. Although,

care must be taken, that eye-gaze tracking data is used in a sensible way, since the nature of human eye movements is a combination of several voluntary and involuntary cognitive processes.

Case 1: Without Filtering

The frames that have been detected earlier will be used in this step to find the status of the eye, if it is open or close. The algorithm gets 15 frames to identify the correct position of the eye. To determine the frame's pixels threshold, a binary threshold using the following equation has been applied. The threshold is initialized to 70 after experimentation we found that 70 is the best number to use.

If the intensity of the pixel src(x,y) in the frame is higher than threshold, then the new pixel intensity is set to a maxVal. Otherwise, the pixels are set to zero. Figure 5 depicts the flowchart of the binarization process of the given frames.

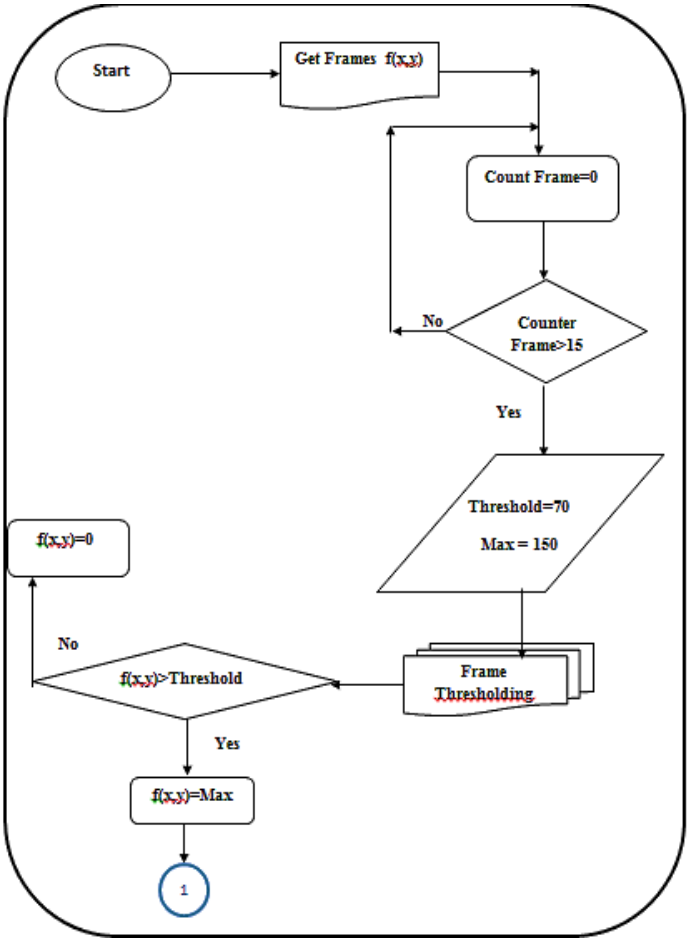


Fig. 5. Frame thresholding flowchart.

$$dst(x,y) = \begin{cases} \max Val & if\ src(x,y) > threshold \\ 0 & otherwise \end{cases} \quad (1)$$

In the binary frames, 0 represent the black color and 1 represent the white color for each pixel. These frames will go through a series of operations to convert all points of black and gray to zeroes and determine the length and width of the part under the eyebrows. If the index is not equal to zero, the points gray will increase. Otherwise, the number of black points will increase and this process help to discover whether the eye is open or closed .Consequently if the black points greater than 3, it means the case is open, otherwise the eye is close as shown in figure 6.

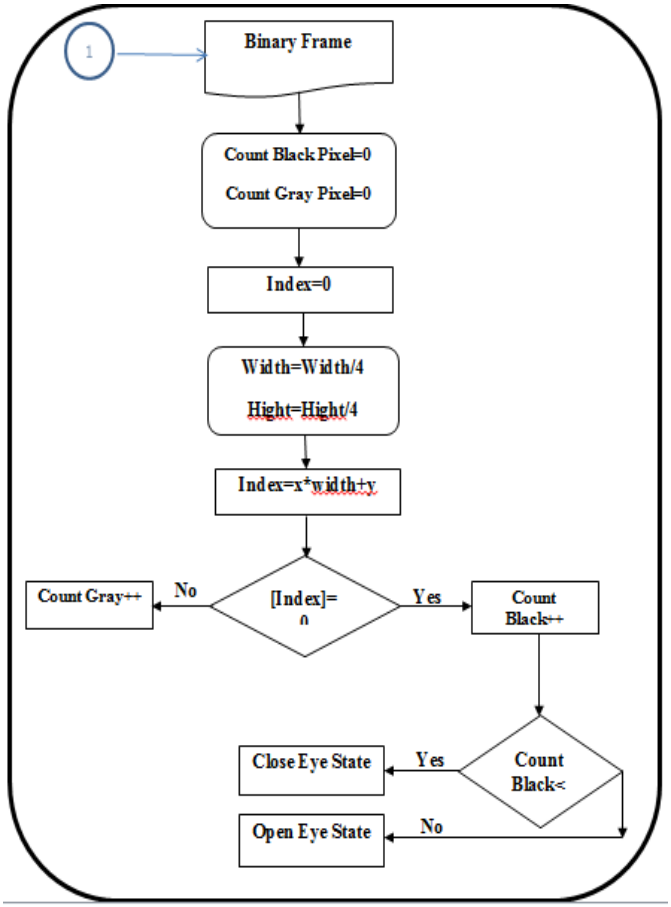


Fig. 6. Eye blinking flowchart.

Case 2: With Filtering

After detecting the eye blink successfully, the same steps that have been applied before will be used but with applying the Medium Blur Filter on the binary frames. The major objective of smoothing or blurring image is to decrease the noise. Such noise reduction is a typical image pre-processing method which will improve the accuracy of detection.

After applying the median filter on the frames, the algorithm will check if at least one black pixel appears. If there is no black pixel, the threshold value will increase and follow the same sequence, but if there is more than one black pixel, the process will terminate and get the state of eye value as that threshold.

F. Mobile Phone Controlling with Eye Blinking

The last step of the EBCM algorithm is controlling the mobile activity by making a phone call depending on the result of previous step. If the state of the eye is close which means the human eye blinked, a call phone to specific phone number will be made, otherwise no phone call will be made as shown in figure 7. The EBCM application used ACTION_CALL action to trigger built-in phone call functionality available in Android device.

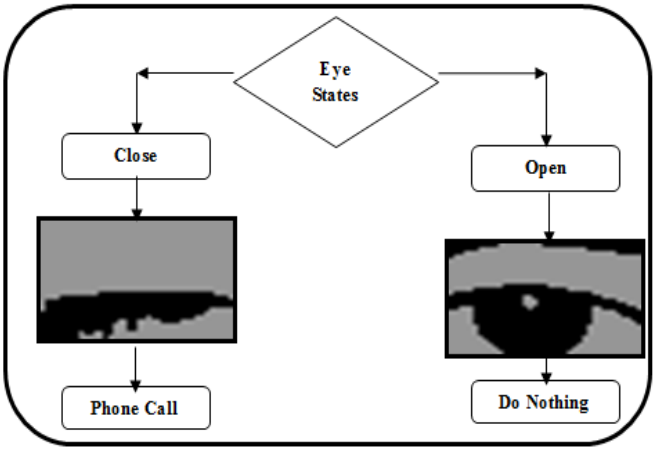


Fig. 7. The eyelid status

G. Performance Parameters

The main parameters that affect the proposed EBCM are the distance and the lighting. If the distance between the user's eye and the mobile phone is long the process of detecting will be difficult or impossible in some cases. The light also affects the process of detecting the eye either in a normal light condition or in an artificial light.

To find the accuracy of the eye detection in the proposed algorithm the following equations has been used [6].

Over all Accuracy =

((TP+TN)/(TP+FP+FN+TN))*100% (2)

Detection Accuracy =

(TP/(TP+FN))*100% (3)

Where TP is the number of frames that are correctly detected eye blinks (true positive); FN is the number of frames that show eye blinks but the program is not detected (false negative); FP is the number of frames that are reported as eye blinks but they are not (false positive); and TN is the number of frames that are correctly reported as no blinks (true negative).

III. Experimental Results

The test are conducted by using Intel(R) Core(TM) 2 Due CPU P8400 ,2.10 GHZ, 32 Bit processer with 4.00 GB of RAM.

The simulation program is compiled by using Eclipse Platform Juno (4.2) that supports Java code, OpenCV library [9, 10] (as an image processing and computer vision tools), and windows applications, which benefits from its features to design a user interface for android mobile application. The mobile that held the tests is Samsung Galaxy S3 that has Android version 4.3 with model number GT-19300. Finally, the software executed under Windows7 as the operating system.

The Performance parameters that are used in order to evaluate the proposed (EBCM) method are:

- The distance between the volunteer's eye and the mobile camera are taken into account, ranging (15, 20, 25, 30, 35, 40, 45) cm.
- The light condition
  - Normal room light or,
  - High quality artificial room light.

Figure 8 present a screen shot of our application.

Fig. 8. Disable volunteer during application test

A. Normal Light Condition (without Filtering)

Table I presents the test results of the case that the normal light is used without applying a filter.

TABLE I. Overall and Detection Accuracy Vs. Distance

Distance (cm)	TP	FP	FN	TN	Overall %	Detection %
15	0	0	0	0	0	0
20	0	0	18	53	75	0
25	1	0	99	2489	96	1
30	3	0	265	2300	90	1
35	221	77	4	1355	95	98
40	142	16	30	2671	98	83
45	2	0	10	26	74	17

B. Normal Light Condition (with Filtering)

Figure 9 shows the test results of overall and detection accuracy versus distance under the normal light using a median filter.

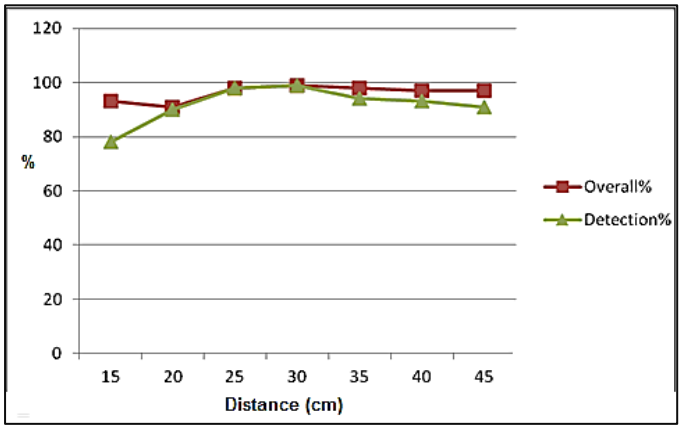


Fig. 9. Overall and detection accuracy vs. distance (with filter)

C. Best Light Condition (with Filtering)

In this case the detection accuracy rate is very optimal because of the effect of the filter as shown in figure 10.

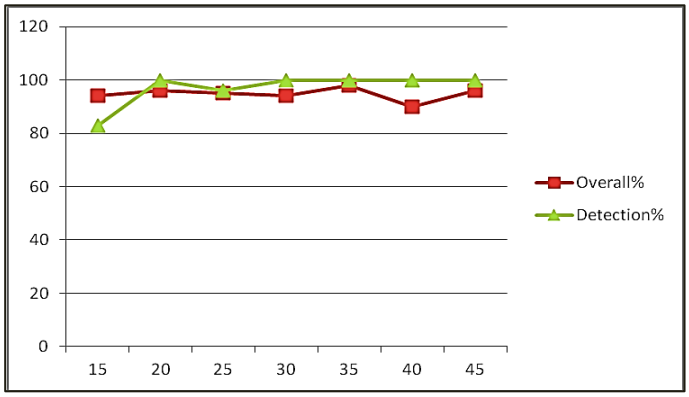


Fig. 10. Optimal detection accuracy

IV. Conclusions and Future Works

Eye blink detection is a very challenge problem for controlling mobile phones in a real time application. This is due to the movement of eyes and the variation of light for different distances from the mobile camera. The proposed method provides 8% of accuracy improvement for eye detection and blinking. When an artificial light is used the overall and detection accuracy are 98% and 100% respectively for a distance equal to 35 cm. Each frame takes an average of 71 ms for time execution which is very efficient for real time application.

The future work will be improving the security level of the proposed eye tracking system using voice recognition algorithm and adapting the application for a non-frontal face.

Acknowledgment

This research was supported by the center of disabling in Erbil city. The volunteers have been participated for conducting the experimental tests. Special thanks for the college of education/ Computer Department/ University of Salahaddin and School of Science/ Computer Department/ University of Sulaimani.

References

[1] S. Naveed, B. Sikander and M. Khyal, “Eye Tracking System with Blink Detection”, Journal of Computing Vol 4, Issue 3, 2012, pp.51-60.

[2] P. Viola and M Jones, “Rapid Object Detection using a Boosted Cascade of Simple Features”, Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, December 9-14, Vol. 1, 2001, pp. 511-518.

[3] C. Holland and O. Komogortsev, “Eye Tracking on Unmodified Common Tablets: Challenges and Solutions”, ACM Proceeding of Symposium on Eye Tracking Research & Applications. USA, 2012, pp. 277-280.

[4] Grauman, K., James, Betke, M., Gips, J. and Bradski, G. R., "Communication via Eye Blinks – Detection and Duration Analysis in Real Time," Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA, December 9-14, 2001, Vol. 1, pp. 1010-1017.

[5] M. Lalonde, D. Byrns, L. Gagnon, N. Teasdale, and D. Laurendeau, “Real-time eye blink detection with GPU based SIFT tracking,” Proc. of the 4th Canadian Conf. on Computer and Robot Vision (CRV), Montreal, Quebec, May 28-30, 2007, pp. 481-487.

[6] C. Devahasdin and T. Srinark, “A Method for Real time Eye Blink detection And Its Application”, The 6th International Joint Conference on Computer Science and Software Engineering, Thailand, 2009.

[7] E. Miluzzo, T. Wang, and T. Campell, “EyePhone: Activating Mobile Phones With Your Eyes“, MobiHeld 2010, August 30, New Delhi, India, 2010.

[8] S.Saravanakumar, and N.Selvaraju, “Eye Tracking and Blink Detection for Human Computer Interface”, International Journal of Computer Applications, Vol 2 – No.2, May 2010, pp. 7-9.

[9] Hewitt, R., “Seeing with OpenCV A Computer-Vision Library,” SERVO, January 2007, pp. 62-66.

[10] Hewitt, R., “Seeing with OpenCV Finding Faces in Images,” SERVO, February 2007, pp. 48-52.

A Fast Cryptosystem Using Reversible Cellular Automata

Said BOUCHKAREN
Department of Mathematics and Computer Science / LABTIC
National School of Applied Sciences of Tangier
AbdelMalek Essaadi University
B.P. 1818, Tangier Morocco

Saiida LAZAAR
Department of Mathematics and Computer Science / LABTIC
National School of Applied Sciences of Tangier
AbdelMalek Essaadi University
B.P. 1818,Tangier Morocco

***Abstract*—This article defines a new algorithm for a secret key cryptosystem using cellular automata which is a promising approach to cryptography. Our algorithm is based on cellular automata built on a set of reversible rules which have the ability to construct unpredictable secret keys using MARGOLUS neighborhood. To prove the feasibility of the algorithm, we present some tests of encryption, decryption and diffusion; a CPU time comparison with an encryption algorithm by blocks as for instance AES-256 is established. On the other hand, the security of the algorithm is proved and the implemented algorithm resists against a brute force attack.**

Keywords—AES; Cellular automata; Diffusion; Cryptosystem; MARGOLUS neighborhood

I. INTRODUCTION

Cryptographic algorithms are used to secure computer networks, electronic transactions or information exchanges. They are implemented in security protocols, electronic chips, etc. Otherwise, and since its appearance, the field of cryptography has experienced a great evolution, mainly in the design of many methods of encryption based on public or secret keys.

The methods of modern cryptography can be divided into two main categories: symmetric and asymmetric cryptography. Each cryptosystem uses keys to generate from the clear text a cipher text. The most known symmetric cryptography systems are DES, AES, RC4 and RC5, [1]. They are used in a secure communication protocols as TLS. Unlike symmetric cryptography, the asymmetric cryptography uses public and secret keys for encryption and decryption. The best known algorithm is RSA [1]; it is implemented in the SSL protocol.

Our work is part of a new approach to cryptosystems based on cellular automata (CA) which presents a promising approach to cryptography. CA gives a secret key for the encryption which cannot be predicted since it evolves a chaotic and complex system starting from an initial state [8-10]. A Brief history of CA can be found in [4].

In one dimension, many encryption concepts based on CA was studied, the most known ones belong to Wolfram [13-14]. In two dimension, some cryptosystems based on CA have been constructed for public and secret keys [9],[11],[15]; for example, CA were used to construct cryptosystem based on Vernam cipher and generated keys with a pseudo-random numbers sequence; CA were also used for block cipher. Concerning our work, we aim to design and to implement a novel two dimensional secret key cryptosystem based on reversible CA using MARGOLUS neighborhood [2]. We remind however that reversibility concept has been used for block encryption in one dimension; for more details, see [8].

This paper is organized as follows. The first section consists on a brief review of symmetric and asymmetric cryptographic methods based on secret and public keys. Section 2 outlines CA and section 3 focuses on MARGOLUS neighborhood. Section 4 describes our cryptosystem. To prove its feasibility, a simulation will be presented in section 5; it holds on the diffusion and performance tests where a comparison with the AES algorithm is established. Section 6 concludes the paper.

II. CELLULAR AUTOMATA AND REVERSIBILITY

A cellular automaton (CA) is a dynamic system defined by the following 4-tuple: dimension, set of finite states, neighborhood and set of rules. Dimension defines number of cells. Cells are updated accordingly to some rule. Such rule is based on the state of the cell and the neighborhood.

More precisely, let A be a cellular automaton defined by $A = \{S, Z^d, f, V\}$ where:

- S is a finite set of states,
- Z is the set of integers,
- d is the size of the automaton,
- Z^d is the space of the automaton,
- $f : S^n \to S$ is the rule (transition function), $n = card(V)\ where\ V = \{v_1, v_2, \dots, v_n\} \subseteq Z^d$ is the set of neighborhood, $(v_i \in Z^d, 1 \leq i \leq n)$.

We call configuration the allocation of the state of each automaton cell. To illustrate, we present the following test:

$A = \{S, Z^1, f, V\}$

- $S = \{0, 1\}$, two states,

- $V = \{-1, 0, 1\}$ (the neighbors of the cell i are i-1, i+0, i+1),
- f is defined by the following box:

V(∗)	7	6	5	4	3	2	1	0
$f(V)$	0	0	1	0	0	1	0	1

(*) Expressed in the basis 10 (5 ➔ s[i]=0, s[i-1]=1, s[i+1]=1).

Let the configuration at time$t$:

1	0	1	0	0	**1**	0	0	0	1

At time t +1 the configuration will be:

0	1	1	0	0	**1**	0	1	0	0

According to the definition table of f , we have in binary 00100101 which are worth 37 in decimal, we are talking about the rule 37. In total, there are $2^8 = 256$ rules. For an automaton of dimension d where the set of states is $S = \{s_1, s_2, \dots, s_n\}$, we can use n^d rules.

In this work, we use a reversible cellular automaton according to this definition: the automaton A with the evolution function f is reversible if there is an automaton B with the evolution function g such that for each configuration, $f(g(c)) = c$. For more details, we can refer to the works published in [3-6].-For the one-dimensional CA, it is possible to check if the automaton is reversible [7]. For the CA of other dimensions, the reversibility is undecidable [6].

III. MARGOLUS NEIGHBORHOODS

We construct two-dimensional CA by blocks based on MARGOLUS neighborhood. We remind that if we consider two-dimensional grid then it is possible to define different kinds of neighborhoods; the most common ones are Von Neumann (related to four neighborhoods), Moore (associated to nine neighborhoods) or MARGOLUS [4],[9],[12]. To justify our choice, MARGOLUS neighborhood allows creation of reversible CA and this reversibility gives more security to the encryption procedure.

For MARGOLUS neighborhood, the automaton cell is divided into blocks of n cells (four cells for example), we applied the transition function to each block. To create reversibility, we use the reversible transition function that can be built by creating a mapping between the set of states.

To illustrate, let A be a cellular automaton of dimension 2, with two states S = {0, 1}, we construct a block of four cells as indicated in Figure 1. The arrow indicates the conventional sense.

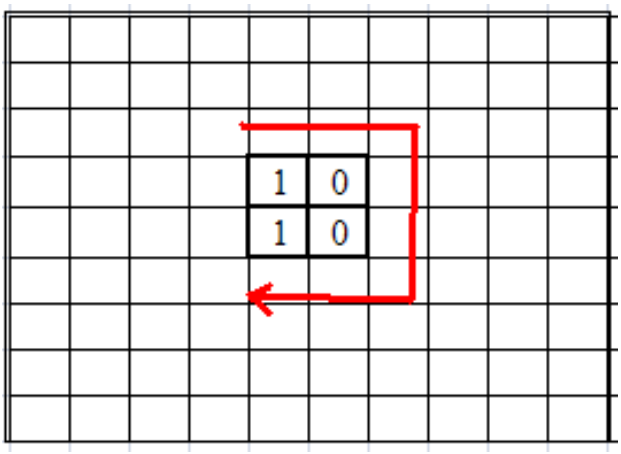


Fig. 1. Bits interpretation Sense

Four cells imply the use of four bits, thus the minimum value is 0000 (in binary) and the maximum value is 1111 (in binary and 15 in decimal). Each cell may contain a value of the set D = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15}. Therefore, to construct a reversible rule, we just need to take a bijective application $f: D \to D$.

Figure 2 explains the concept of the MARGOLUS neighborhood. First, we take the original automaton « A », then we partition into blocks of n cells (4 cells in this case). Second, we partition starting with the first index (even partitioning (B) on Figure 2).

Finally, we redo the odd partitioning starting with an odd index (C), see Figure 2.

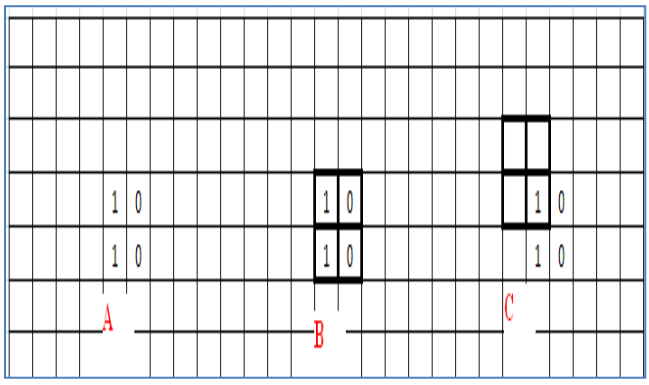


Fig. 2. Partitioning of a 2-D cellular automaton in MARGOLUS neighborhood basis.

IV. PROPOSED CRYPTO-SYSTEM

The proposed cryptosystem is based on two-dimensional reversible CA using MARGOLUS neighborhood, its secret key is the transition rule of the cellular automaton.

A. Description of the algorithm

Encryption step

- Divide the unencrypted data into blocks of $4T^2$-bits (for example 1024-bits)
- Arrange the $4T^2$-bits in a square matrix M of dimension 2 with a size of 2Tx2T
- Generate a reversible transition rule R (Rr is the inverse rule of R)
- Generate a key K={R, Rr}
- For i←1 until 6

  M=PRule(K,M)
  M=IRule(K,M)
  End

- Group the bits of the matrix M to obtain the encrypted data.

Decryption step

As already mentioned, the proposed system is symmetric; for the decryption, we use the same key as in the encryption step.

- Divide the encrypted data into blocks of $4T^2$-bits (for example 1024-bits)
- Arrange the $4T^2$-bits in a square matrix M of two dimension with a size of 2Tx2T

  ***For** i←1**util** 6*
  M=IRule(K,M)
  M=PRule(K,M)
  End

- Group the bits of the matrix M to obtain the unencrypted data.

B. Transformation PRule() and IRule()

Consider the following data presented in Figure 3:

Fig. 3. Illustration data

PRule(), IRule() are the functions that evolve the cellular automaton, they use the MARGOLUS neighborhood described above.

PRule() applies on the even blocks (the red blocks in the graph), in opposite, IRule() applies on the odd blocks (the green blocks in the graph).

C. Security of the algorithm

The security of this algorithm is based on the choice of CA parameters: we can increase or decrease the neighborhood size or/and the number of states.

Let N be the neighborhood size and S the number of states. The maximum possible number of keys is $(S^N)!$ which making impossible to apply a brute force attack with an exhaustive search of keys; for example, if S =2 and $N = 8$, then we have to try $2^8!$=256! $\approx 8.10^{506}$ combinations that are impossible to achieve within a reasonable time.

V. SIMULATION

For the simulation, we used a cellular automaton by blocks with the following data:

- Number of blocks per cell: 4 (2X2)
- Cardinal of the set of states : 2 (0 and 1)
- Transition rule:
  R={0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15} and its inverse Rr={3,2,5,8,9,7,0,4,1,13,6,14,10,15,11,12}

For the system parameters, we used the following data:

- Message = « Crypto-system based on the CAs. »
- K={R, Rr}
- 256-bits in each data block.

To make the encrypted data readable, we used the *Base64* encoding.

The clear message encoded by *Base64* is:
« Q3J5cHRvLXN5c3Q/bWUgYmFzPyBzdXIgbGVzIEFDcy4 = »
The encrypted message is:
« isFbk69OTBjYrZdRdOb46riq+5NlrrYZN/uuYE5sQF8= ».

A. Diffusion application

The diffusion calculates the influence to change the bits in the clear message (plain text) onto the encrypted message keeping constant the key. Figure 4 gives diffusion tests of CA cryptosystem and AES algorithm by taking a plain text of size 256-bits with a key of 256-bits. It is observed that diffusion levels obtained by CA cryptosystem are better than AES.

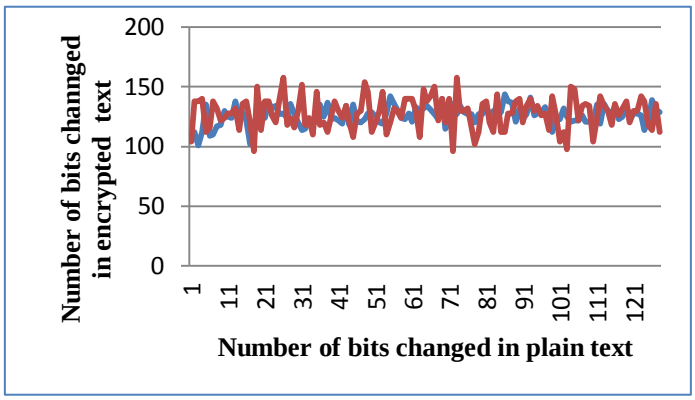


Fig. 4. Diffusion tests: Red curve corresponds to CA cryptosystem and blue curve corresponds to AES algorithm.

B. *Performance test*

We compare our CA cryptosystem with a block algorithm as for instance AES-256. This is illustrated in Figure 5. The comparison focuses on the performance tests (CPU time in ms). The encryption using CA is faster than AES algorithm

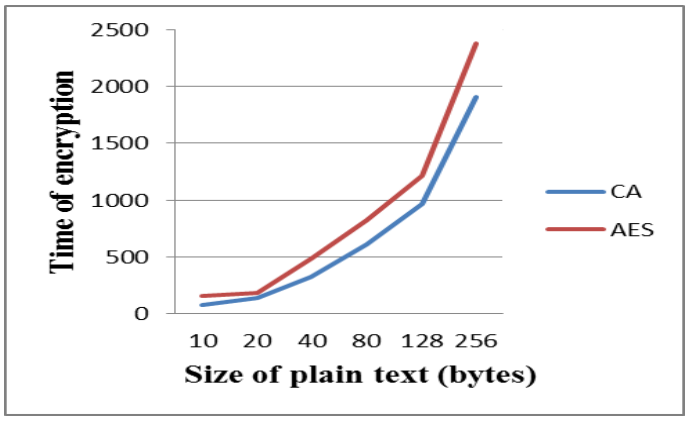


Fig. 5. Comparison of CPU times between CA cryptosystem (blue curve) and AES algorithm (red curve).

VI. CONCLUSION

In this paper, we presented a brief review of symmetric and asymmetric cryptographic methods; we introduced the concept of cellular automata as a promising approach to cryptography allowing creation of unpredictable cryptosystem keys born through a chaotic and complex system. We outlined also the neighborhood concept that is linked to the creation of cellular automaton and we described Margolus neighborhood which we selected to build our two-dimensional reversible CA. We constructed a fast secret key cryptosystem and we demonstrated its feasibility.

We realized a test of diffusion and a CPU time comparison with a block cipher algorithm; the comparison shows that the proposed algorithm is faster than AES-256. The security is proved and the implemented algorithm resists against a brute force attack. We remind that some new research works focus on network sensor security and use encryption algorithms to improve the security of data and energy consumption [16-18]; in this context, we aim to implement our CA cryptosystem as a lightweight security protocol.

REFERENCES

[1] S. Bruce Schneier, “Applied Cryptography”, John Wiley & Sons New York, 1996.

[2] T. Toffoli and N. Margolus, “Invertible Cellular automata”, Physica D 45, 1990.

[3] S. Wolfram, “A New Kind of science”, Wolfram Media, 2002.

[4] P. Sarkar, “A Brief History of Cellular Automata”, ACM Computing Surveys, 2000.

[5] J. L. Schiff, “Cellular automata: A discrete view of the world”, Wiley, 2008.

[6] J. Kari. “Reversible cellular automata”, Springer Berlin Heidelberg, 2005

[7] J. Kari, “Reversibility and Surjectivity problems of cellular automata”, Journal of Computer and system sciences 48, 1994.

[8] Marcin Seredynski and Pascal Bouvry, “Block Encryption Using Reversible Cellular Automata”, Lecture Notes in Computer Science, Volume 3305, pp 785-792, 2004.

[9] [9] Sambhu Prasad Panda, Madhusmita Sahu, Umesh Prasad Rout and Surendra Kumar Nanda, “Encryption and Decryption algorithm using two dimensional cellular automata rules in Cryptography”, International Journal of Communication Network & Security, Volume-1, Issue-1, 2011.

[10] Somanath Tripathy and Sukumar Nand, “LCASE: Lightweight Cellular Automata-based Symmetric-key Encryption” International Journal of Network Security, Vol.8, No.2, PP.243-252, Mar. 2009.

[11] Franciszek Seredynski, Pascal Bouvry, Albert Y. Zomaya, “Cellular automata computations and secret key cryptography Parallel Computing”, 2004.

[12] Joaquin Cerda, Rafael Gadea and Guillermo Paya. “Implementing a Margolus Neighborhodd Cellular Automata on a FPGA”, Lecture Notes in Computer Science, Volume 2687, pp 121-128, 2003.

[13] Stephen Wolfram, “Cryptography with cellular automata”, Lecture Notes in Computer Science. Volume 218, pp 429-432, 1986.

[14] Stephen Wolfram, “Cellular automata and complexity”, Addison-Wesley, 1994.

[15] Norman H, Packard and Stephen Wolfram, “Two dimensional cellular automata”, Journal of statistical Physics, 1985.

[16] Alvaro Araujo, Javier Blesa, Elena Romero and Daniel Villanueva, “Security in cognitive wireless sensor networks. Challenges and open problems”, EURASIP Journal on Wireless Communications and Networking, 2012.

[17] Nabil Ali Alrajeh, S. Khan and Bilal Shams, “Intrusion Detection Systems in Wireless Sensor Networks: A Review”, International Journal of Distributed Sensor Networks, 2013.

[18] Alexandros Fragkiadakis, Vangelis Angelakis and Elias Z. Tragos, “Securing Cognitive Wireless Sensor Networks: A Survey”, 2014.

Herbal Leave Recognition System Based on Dirichlet Laplacian Eigenvalues

Mahmoud Elgamal
The Custodian of the Two Holy Mosques
Institute for Hajj and Omra Research,
Umm Al -Qura University
Makkah, Saudi Arabia
Email: maelgamal@uqu.edu.sa

Mahmoud Youness R. Alaidy
College Of Computer
Computer Eng., Dept, Qassim Univ.,
Buryadh, Saudi Arbia
Email: aaiedy@qu.edu.sa

***Abstract*—Identifying and recognition of herbal plant green leaves is essential in botanical study. In[8] Thai herb leaf image recognition system used for recognition of leaves with accuracy of 93.29%, in this paper, we propose a recognition system of leaves based on the eigenvalues of Dirichlet Laplacian that used to generate three different sets of features for shape analysis and classification in binary images[4]. First leaf images are preprocessed to remove unwanted background, converted to binary form; used to build the images database, finally Queries made on the system. The correct classification rates without noise is 100% and with noise is $\sim 90\%$.**

Keywords: Eigenvalues, Finite difference method, Curve descriptor, Binary image classification, noise, leave recognition.

I. Introduction

Shape recognition is the field of computer vision which addresses the problem of finding out whether a query shape lies or not in a shape database, up to a certain invariance. Most shape recognition methods simply sort shapes from the database along some similarity measure to the query shape. Shape analysis is a key component in object recognition, matching, registration and analysis. A shape description method generates a feature vector that will uniquely characterize the silhouette of the object. This vector should in many cases, be translation-, rotation-, and size-invariant. Depending on the application at hand, a certain level of robustness and tolerance to shape deformation and noise is also required. As an important application of shape recognition, leave recognition which has significant attention in botanical study. However, by far the most popular classification in Loncaric [6] of shape techniques divides the different methods into two groups: boundary methods and global methods. Boundary methods treat the boundary or exterior points of the shape, while global methods deal with the interior points of the object. There is no clear consensus which method or category of methods works best. Each method seems to give a good result in some applications and fail in some others or in presence of noise. The method presented in this paper is a numerical non-preserving global method that attempts to use the ratios of eigenvalues of the Dirichlet Laplacian operator of a certain shape as the feature vector.

The paper structured as follows, a brief mathematical overview of the model and the evaluation of the eigenvalues in sections(II and III), feature set evaluation in section(IV), algorithmic implementation in section(V), and finally the simulation.

II. The Dirichlet Laplacian eigenvalues

Let Ω be a bounded domain in $\mathbb{R}^n$, $n \geq 2$. Consider the eigenvalue problem for the Laplace operator with Dirichlet boundary condition,

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega, \\ u|_{\partial\Omega} = 0. \end{cases} \tag{1}$$

Here in (1), $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$ is the Laplacian. As is well known, the Dirichlet Laplacian (Laplacian with Dirichlet boundary condition) has diverse applications in science and engineering, and we refer to Cureton and Kuttler [2] and Kuttler [5] on the detailed study of Dirichlet Laplacian in two dimensional polygons.

Let us denote the eigenvalues by $\lambda_1(\Omega), \lambda_2(\Omega), \cdots$, (we will sometimes omit explicit dependence on Ω when speaking about generic domain), where

$$0 < \lambda_1 < \lambda_2 \leq \lambda_3 \leq \cdots \longrightarrow \infty. \tag{2}$$

It is also well known that the eigenvalues of the Dirichlet Laplacian are preserved if the underlying domain Ω is translated or rotated (see Courant and Hilbert [1]). In the next section it will be discussed how to evaluate the eigenvalues.

III. Computation of the eigenvalues

A. Finite difference method

In order to evaluate the approximate numerical solution of (1), there are several methods. Among those we choose the finite difference scheme which was first proposed in Pólya [11]. The scheme is to replace (1) by the recursive formula

$$\frac{u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{i,j}}{h^2} = -\lambda u_{i,j}. \tag{3}$$

Here the domain Ω is divided into squares of side h, and u_{ij} is the value of the eigenfunction corresponding to λ at the lattice point (ih, jh) (see Figure 3.1). This scheme can be written in compact form as

$$\mathcal{L}u = \lambda u,$$

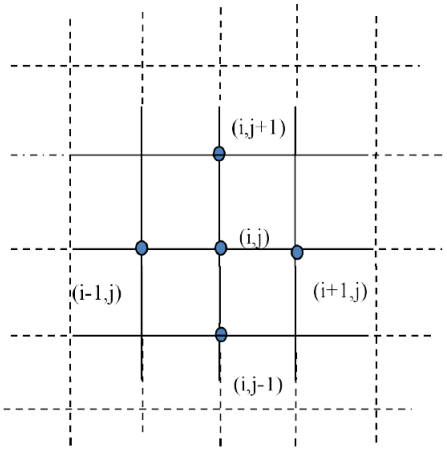


Fig. 3.1: five-stencil approximation for the Laplacian.

where

$$\mathcal{L} = \frac{1}{h^2}\begin{bmatrix} A & I_n & 0 & \cdots & 0 \\ I_n & A & I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A \end{bmatrix}_{n^2\times n^2}$$

and

$$A = \begin{bmatrix} -4 & 1 & 0 & \cdots & 0 \\ 1 & -4 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -4 \end{bmatrix}_{n\times n}.$$

Here, n is inversely proportional to h and accounts for the size of the domain Ω. The eigenvalues $\lambda'_1, \lambda'_2, \cdots$ of this finite dimensional problem provide, in general, lower bounds for $\lambda_1, \lambda_2, \cdots$ (cf. [10]).

IV. Features Generation and Evaluation

For a given binary image Ω, ([12] and [4]) proposed the following three feature sets based on the above described eigenvalues

$$F_1(\Omega) \equiv \left\{\left(\frac{\lambda_1}{\lambda_2}, \frac{\lambda_1}{\lambda_3}, \frac{\lambda_1}{\lambda_4}, \cdots, \frac{\lambda_1}{\lambda_n}\right)\right\}, \quad (4a)$$

$$F_2(\Omega) \equiv \left\{\left(\frac{\lambda_1}{\lambda_2}, \frac{\lambda_2}{\lambda_3}, \frac{\lambda_3}{\lambda_4}, \cdots, \frac{\lambda_{n-1}}{\lambda_n}\right)\right\}, \quad (4b)$$

$$F_3(\Omega) \equiv \left\{\left(\frac{\lambda_1}{\lambda_2} - \frac{d_1}{d_2}, \frac{\lambda_1}{\lambda_3} - \frac{d_1}{d_3}, \cdots, \frac{\lambda_1}{\lambda_n} - \frac{d_1}{d_n}\right)\right\} \quad (4c)$$

Here n counts the number of the desired features to be used for the recognition scheme, and $d_1 < d_2 \leq d_3 \leq \cdots \leq d_n$ are the first n eigenvalues (counting multiplicity) of a disk. All three features are obviously size-invariant [4]. The F_1 features were first proposed by Zuliani et al. [12]. The values of $F_1(\Omega)$ and $F_2(\Omega)$ are in the unit cube, while those of $F_3(\Omega)$ are between ± 1 a useful range when using neural networks. This later descriptor is a good measure of the deviation of Ω from a disk. The optimal number of features n depends on the problem being addressed and is determined experimentally.

To test the consistency of these feature sets for a given image class, their tolerance to noise, experiments are conducted and the simulation discussed in next section.

V. Implementation

In this section, an algorithm is developed based on the above discussion to evaluate F_1, F_2, and F_3 of a all images, also for the query image; compare to get the minimum value of the norm as explained in fig. (5.1). Figure (5.2) shows the flow diagram of the GUI

Input: query image i_q,
Output: the most similar image (i_{out}), from the image database(IDB).

Step 1:

a) Read the images ($i = 1, \cdots, n$) from the stored Database images(IDB)
b) Convert all the images in *IDB* to binary images.
c) Evaluate F_1^i, F_2^i, & F_3^i for each images of the *IDB*.
d) Calculate the norms of F_1^i, F_2^i, & F_3^i; $NF_1^i := \|F_1^i\|$, $NF_2^i := \|F_2^i\|$, and $NF_3^i := \|F_3^i\|$

Step 2: Let the query image q, repeat Step 1:(b - d) to calculate the norms of $NF_1^q := \|F_1^q\|$, $NF_2^q := \|F_2^q\|$, and $NF_3^q := \|F_3^q\|$

Step 3: For i := 1 to N; (N = no. of images),

a) Calculate $NdF_j^i := (NF_j^i - NF_j^q)$, $(j = 1, 2, 3)$.
b) Store the values of NdF_j^i in an array A

Step 4: Find $\min(A)$ and the corresponding index; which is the index of the retrieved image.

Step 5: Display the images of the query image; i_q and the retrieved image; i_{out} along side.

Step 6: Add noise to the images and apply steps 1-5.

Fig. 5.1: Leaf recognition system algorithm.

VI. Simulations

In this part we focus on testing the above described algorithm, it was implemented using Matlab version 8.0; figure(6.2). The leave images was refined by removing background using Adobe Photoshop as shown in figure() A particular feature should have a fairly constant value for all images from a particular class. The consistency of a feature can be measured using its standard deviation from the mean for that image class. To test the consistency of the three feature sets being used, experiments were conducted on different images with and without noise.

Example 1: image resolution at 256 pixels

a) noise = 0.0

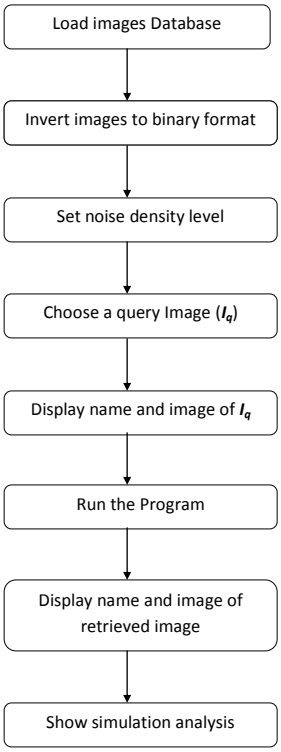
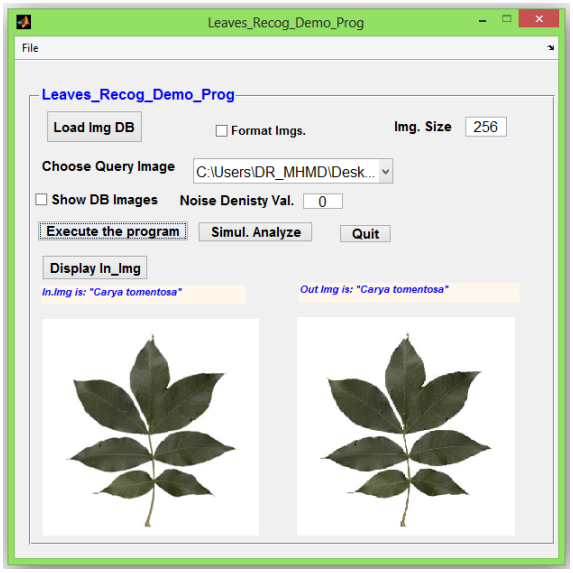


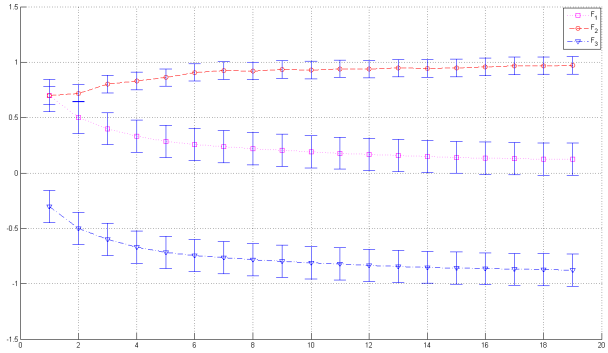
Fig. 5.2: flow-diagram of the algorithm GUI-interface.

Fig. 6.1: Leaf background removal.

Fig. 6.2: GUI of Leave recognition System.



(a) input and retrieved images.



(b) average and standard deviation of the first 25 features from F_1, F_2, and F_3.

Fig. 6.3: output at noise=0.0.

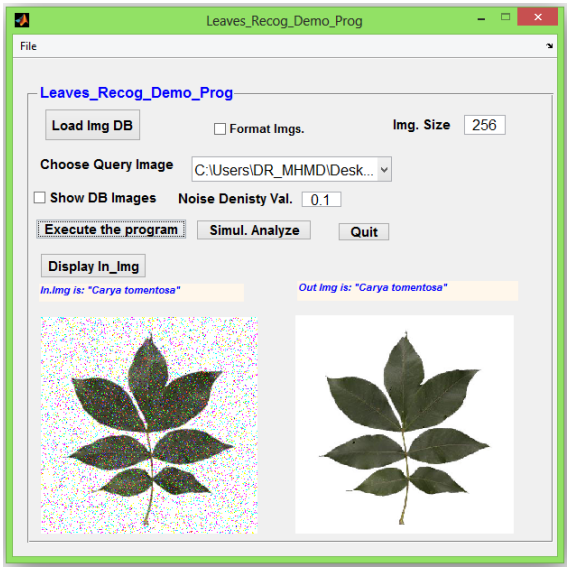
b) noise = 0.1, figure (6.4(b)) shows almost identical values of F_1 for the images without noise and the others with noise

Example 2: image resolution at 128 pixels

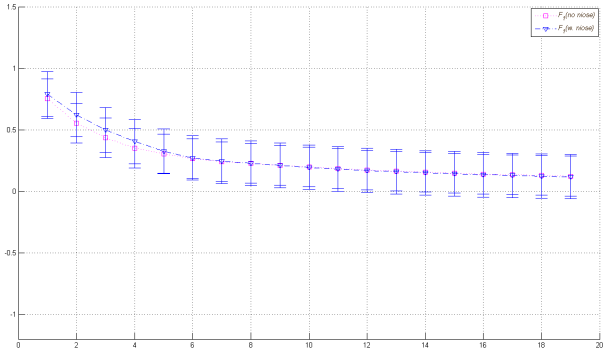
a) noise = 0.0

b) noise = 0.1, figure (6.6(b)) shows almost identical values of F_1 for the images without noise and the others with noise

Example 3:



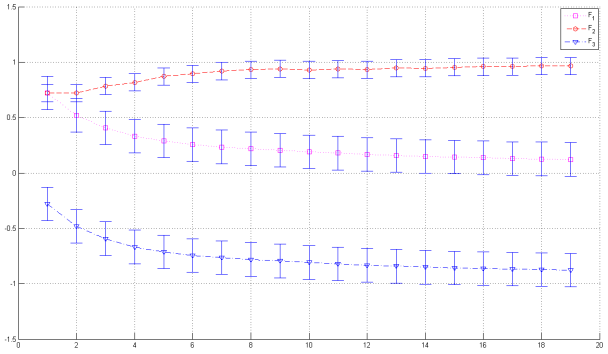
(a) input and retrieved images.



(b) average and standard deviation of F_1 for images with/without noise.

Fig. 6.4: output at noise=0.1.

(a) input and retrieved images.



(b) average and standard deviation of the first 25features from F_1, F_2, and F_3.

Fig. 6.5: output at noise=0.0.

A. Performance Evaluation

We evaluate the performance of the proposed method in terms of *precision*, *recall*, and *accuracy* see[7]. Image retrieval system has the goal to retrieve relevant images while not retrieving irrelevant ones. The measures of performance used in image retrieval borrowd from the field of *document information retrieval* and are based on two primary figures of merit: *precision* and *recall*.

- *Precision*(P) is the number of relevant documents retrieved by the system divided by the total number of documents retrieved(i.e., true positives plus false alarms).

$$P = \frac{TP}{TP + FP} \tag{5}$$

- *Recall*(R) is the number of relevant documents retrieved by the system divided by the total number of relevant documents in the data base(which should have been retrieved).

$$R = \frac{TP}{TP + FN} \tag{6}$$

Precision can be interpreted as a measure of exactness, whereas recall provides a measure of completeness.

- *Accuracy*(A) is the probability that the retrieval is correctly performed

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

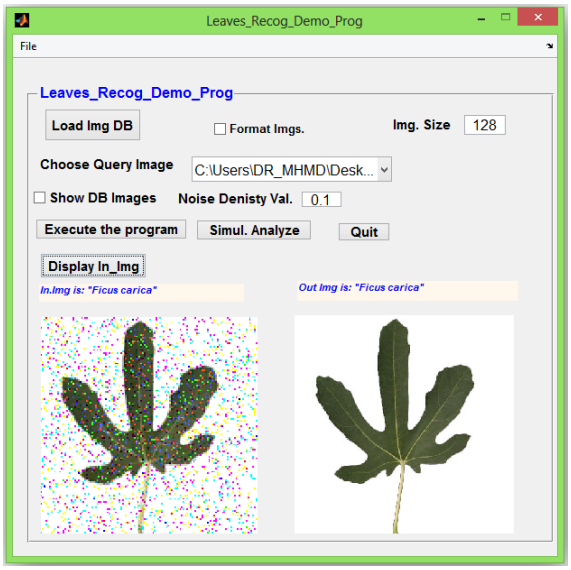
where,

TP(True Positive) - correctly classified positive,

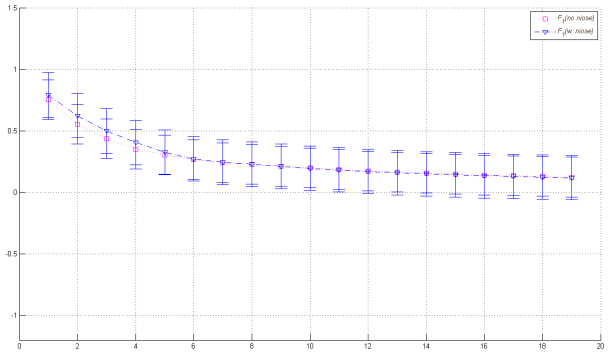
TN(True Negative) - correctly classified negative,

FP(False Positive) - incorrectly classified negative, and

FN(False Negative) - incorrectly classified positive.



(a) input and retrieved images.



(b) average and standard deviation of F_1 for images with/without noise.

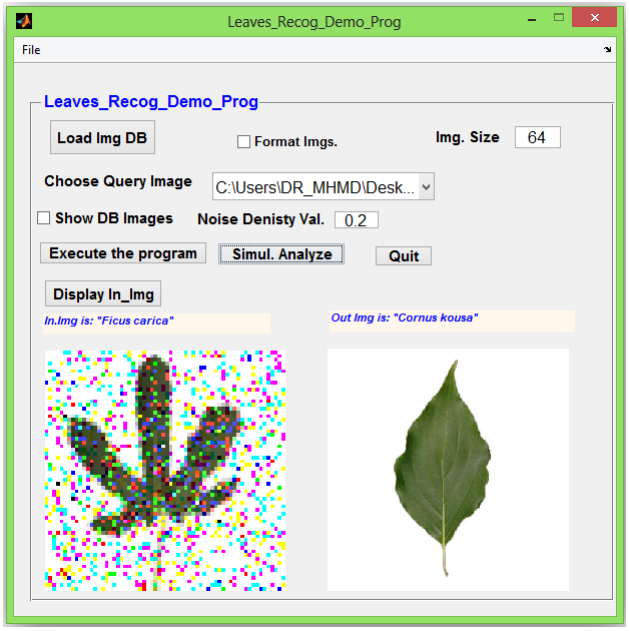


Fig. 6.7: Mis-recognition at 0.2 noise level and 64 image resolution.

	TP	*TN*	*FP*	*FN*	*P(%)*	*R(%)*	*A(%)*
Noise = 0	200	50	0	0	100	100	100
Noise = 0.1	200	50	15	11	93%	94.8%	90.6%

Fig. 6.8: Performance of the used techniques.

VII. Conclusion

The three sets of features based on the eigenvalues of Dirichlet Laplacian, was used to develop a user friendly leave recognition system. The system used successfully to classify images with a high degree of accuracy and using a relatively small number of features. At first it was run on leave database images for the purpose of recognition. Initially without noise and the obtained result was good and then a noise was add to the images but still showed a good result but when increasing the noise level the input and output was different.

References

[1] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Second ed., Interscience Publishers, New York, 1965.

[2] L. M. Cureton and J. R. Kuttler, Eigenvalues of the Laplacian on regular polygons and polygons resulting from their dissection, J. Sound Vib., **220**(1999), 83-98.

[3] Leafsnap: An Electronic Field Guide,"http://leafsnap.com/species/", 2013.

[4] M. A. Khabou, L. Hermi and M. B. H. Rhouma, Shape Recognition Using Eigenvalues of the Dirichlet Laplacian, Pattern Recognition, **40**(2007), 141-153.

[5] J. R. Kuttler and V. G. Sigillito, Eigenvalues of the Laplacian in two dimensions, SIAM Rev. **26**(2)(1984), 163-193.

[6] S. Loncaric, "A survey of shape analysis techniques", Pattern recognition **31**(1998), 983-1001.

[7] Olson, David L.; and Delen, Dursun (2008); "Advanced Data Mining Techniques", Springer, 1st edition (February 1, 2008), page 138, ISBN 3-540-76916-1.

[8] C. Pornpanomchai, S. Rimdusit, P. Tanasap and C. Chaiyod, "Thai Herb Leaf Image Recognition System (THLIRS)", Kasetsart journal(natural science), 45 : 551 - 562 (2011).

[9] S. Pramanik, S. K. Bandyopadhyay, D. Bhattacharyya, and T. Kim, Identification of Plant Using Leaf Image Analysis ,2010.

[10] G. Pólya, Sur une interprétation de la méthode des différences finies qui peut fournir des bornes supérieures ou infrieures, C. R. Acad. Sci. Paris **235**(1952), 995-997.

[11] G. Pólya, Esimates for eigenvalues, in: Studies in Mtahematics and Mechanics Presented to Richard von Mises, Academic Press, New York, (1954), 200-207.

[12] M. Zuliani, C. Kenny, S. Bhagavathy and B. S. Manjunath, Drums and curve descriptors, UCSB Visition Research Lab Preprint, 2004.

Solving for the RC4 stream cipher state register using a genetic algorithm

Benjamin Ferriman
School of Computer Science
University of Guelph
Guelph, ON N1L 1L9

Charlie Obimbo
School of Computer Science
University of Guelph
Guelph, ON N1L 1L9

***Abstract*—The RC4 stream cipher has shown to be quite resilient to cryptanalysis for the 26 years it has been around. The algorithm is still one of the most widely used methods of encryption over the Internet today being implemented through the Secure Socket Layer and Transport Layer Security protocols. Genetic algorithms are a sub-class of evolutionary algorithms that have been used to help solve many different problems of optimization in a variety of disciplines. In this paper we will examine the abilities of the genetic algorithm as a tool to help solve the permutation that is stored as the state register of the RC4 stream cipher. Finally, we will show that on average the genetic algorithm can solve 100% of the keystream in $2^{121.5}$ generations.**

I. Introduction

Over the past twenty years the Internet has evolved astronomically as a tool for education, pleasure, and economics, to name a few applications. In today's society there are very little tasks in one's daily life which are not facilitated by the Internet in some way, shape, or form. As the services available over the Internet continue to expand, new and old problems of security arise and must be accounted for in order to properly facilitate these applications.

One of the largest sectors which continues to grow is on-line banking and other electronic financial transactions (conveniently distinguished as E-Commerce). These two applications face many of the same problems as traditional physical banking, but also a new set of challenges that have amounted due to the use of the Internet. The most obvious contemporary issue is that of *communication*. Traditionally a customer simply communicated with a bank teller where the environment could be controlled as well as the manor of communication (i.e. whether something could simply be conveyed through speech or read privately by the customer). With the advent of on-line banking there is an unknown communication between the customer (client computer) and the teller (bank servers). The very fact that on-line banking improves the ease of use for a customer by virtually letting them do their banking anywhere with an Internet connection also hinders their ability to know specifically how their private communication with the bank system is being conducted.

Besides this, there it also the convenient and ubiquitous use of mobile computing. With advent of smart-phones, the main use of the Internet is quickly shifting to being used mainly in the mobile computing environment. According to PewResearch [**?**] as of May 2013, 63% of adult cell owners use their phones to go online and 34% of cell internet users go online mostly using their phones, and not using some other device such as a desktop or laptop computer. As can also be seen on Table I, obtained from the United States Census Bureau Data [**?**], the younger population, between the ages of 10 and 90 comprise over 60% of the population of the US, and according to PewResearch, as can be seen on Table II, about three-quarters of these have and use Smartphones.

TABLE I: Demographics of the US population, 2012

Age	Population	Percentage	Cummulative Percentage
All ages	308,827	100.0%	
Under 5	20,110	6.5%	6.5%
5 - 9	20,416	6.6%	13.1%
10 - 14	20,605	6.7%	19.8%
15 - 19	21,239	6.9%	26.7%
20 - 49	124,607	40.3%	67.0%
50 - 59	42,842	13.9%	80.9%
60 - 64	17,501	5.7%	86.6%
65 & older	41,506	13.4%	100.0%

TABLE II: Smartphone owners in 2014 [**?**]

	Have a smartphone
All Adults	**58%**
Gender	
a. Men	61%
b. Women	57%
Race	
a. White	53%
b. African American	69%
c. Hispanics	61%
Age Group	
a. 18 - 29	83%
b. 30 - 49	74%
c. 50 - 64	49%
d. 65+	19%

With new attacks on Internet-based encryption protocols coming to light in the past four months, a lot of focus has shifted from traditional forms of cryptanalysis to methods of circumvention to attack these ciphers. One cipher that is still widely used and investigated is the RC4 stream cipher. Due to it's simplicity and robustness (efficient for both software and hardware) [1], the RC4 stream cipher is one of the most implemented encryption schemes online and over computer networks. It's usage is seen in the *Secure Socket Layer (*SSL) [2] and *Transport Layer Security (*TLS) [3]

protocols as well as the now obsolete *Wired Equivalent Privacy (*WEP) [4] protocol. It is also used in *Wi-Fi Protected Access (*WPA and WPA 2) protocols (when TKIP is not selected by default).

In recent years, focus has been drawn to implementing *genetic algorithms* as a tool for cryptanalysis. These tools, from the evolutionary algorithms family, have been used in the past to help solve permutation problems such as the traveling salesman problem ([5], [6], [7], [8]). Since the state register of RC4 (see Section I-A1) is a permutation, the researchers would like to investigate the effectiveness of using a genetic algorithm to attempt and solve the permutation sequence of the RC4 state register.

This paper is organized into several sections to present our findings. We will first present the reader with a background of the RC4 stream cipher and genetic algorithms in Sections I-A and I-B respectively. Next, we will propose an implementation of a genetic algorithm in Section II. This will be followed up with an examination and discussion of the results of our experiments in Section III. Finally, we will conclude our findings and present any future avenues of research in Sections IV and V respectively.

The following paper proposes a genetic algorithm to try and solve for the state register permutation of the RC4 stream cipher. The operators that will be investigated include: partially mapped crossover, edge recombination crossover, swap mutation, and inversion mutation. In addition, an *adaptive mutation method* will be utilized in order to reduce the occurrence of a candidate solution becoming stuck in a local optimum over the vast search space. Finally, this paper will show that on average the genetic algorithm will be able to discover 100% of the keystream and replicate the state register in $2^{121.5}$ generations.

A. RC4 Overview

The RC4 stream cipher was invented by Ron Rivest in 1987 while working at RSA Security and designed as a *non-linear feedback shift register* (non-LFSR). It is a stream cipher meaning: given identical initialization keys, the algorithm will produce the same keystream for all parties involved in the communication. RC4 allows the initialization key $\mathcal{K}$ to be of length 40 to 2048 bits. The cipher produces a keystream z of word size n from a state register S of size 2^n consisting of all bit permutations of a n-bit word. Two word sized index pointers, i and j, are used to help perform permutations on the register S. Generally RC4 is implemented with n = 8 bit words.

There are two algorithms that make up RC4. The first algorithm is called the *key scheduling algorithm* (*KSA*) and is used in conjunction with $\mathcal{K}$ to initialize the shift register S into a pseudo-random ordering (see Algorithm 1). The second algorithm, the *pseudo-random generation algorithm* (*PRGA*), uses the register S to produce a pseudo-random keystream z of n-bit words during each iteration of the PRGA loop. The keystream generation process is witnessed in Algorithm 2.

Algorithm 1 Key Scheduling Algorithm (KSA) for RC4

```
Input: Shared Key K
Output: State Register S
for i = 0 → 2^n − 1 do
    S_i ← i
j ← 0
for i = 0 → 2^n − 1 do
    j ← j + S_i + K_i mod 2^n
    Swap(S_i, S_j)
return(S)
```

Encryption works by dividing the plaintext $\mathcal{P}$ into n-bit words and xor'ing them with the keystream z to produce the ciphertext $\mathcal{C}$. This can be expressed as:

$$\mathcal{E}(\mathcal{P}, z) = \mathcal{P} \oplus z = \mathcal{C}$$

Decryption is done by xor'ing each n-bit word of $\mathcal{C}$ with the keystream z producing $\mathcal{P}$. Decryption can be represented as:

$$\mathcal{D}(\mathcal{C}, z) = \mathcal{C} \oplus z = \mathcal{P}$$

Algorithm 2 Pseudo-random Generation Algorithm (PRGA) for RC4

```
Input: State Register S
Output: Keystream bytes z
i ← 0
j ← 0
while 1 do
    i ← i + 1 mod 2^n
    j ← j + S_i mod 2^n
    Swap(S_i, S_j)
    z ← S_{S_i + S_j mod 2^n}
    output(z)
```

1) RC4 State Register: The state register S is represented in the following way. The register as a whole is represented as an array S while each word in the array is represented as $S[i]$, where $i \in \{0, \ldots, 2^n - 1\}$ making up a unique permutation. Within each permutation value of $S[i]$, there exists a binary value denoted as $b_{(i)}$ where $i \in \{0, \ldots, n\}$. An illustration of this representation can be seen in Figure 1. The standard word size is $n = 8$ bits which implies that the total storage of a particular candidate solution in the algorithm is 2048 bits or 256 bytes; practical implementations of RC4 would require 2064 bits to account for the two index pointers (i and j).

B. Genetic Algorithms

Genetic Algorithms (GAs) were introduced as a viable optimization algorithm by John Holland in 1975 [9]. The algorithm utilizes methods found in biology to help evolve a set of candidate solutions (called *chromosomes*) to solve an optimization problem using a cost function over time ([10], [11], [12], [13]). GAs are members of the evolutionary family of algorithms due to their implementation of various evolutionary operators such as reproduction and mutation.

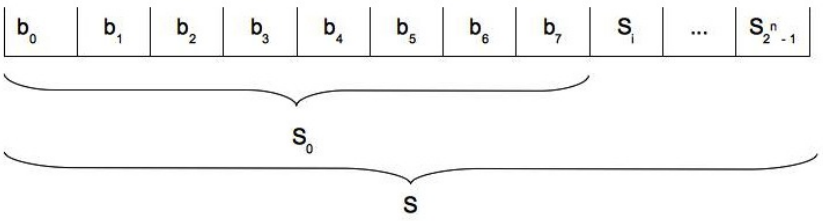


Fig. 1: State Register bit/byte Representation

Solution strength is measured by *fitness* and is determined by a *fitness function* $f(x)$ which quantifies the quality of the solution with a desired requirement [14]. A group of solutions or *population* are a small sample of all possible solutions, thus there is no guarantee that a best fit solution will be the global optimum for the problem. Reproduction or *crossover* is conducted each generation of the algorithm so the stronger candidate solutions have a better probability of entering the next generation of the algorithms life cycle. This is known as *selection*. To ensure that a population does not converge too quickly on a local optimal solution, a process called *mutation* is also introduced in the reproduction process. A basic algorithm of a GA can be seen in Algorithm 3.

Algorithm 3 General Genetic Algorithm

```
Input: Generate initial population p of size N randomly
while Max iteration not met or fitness not satisfactory do
  for i = 1 → N/2 do
    Select p1 and p2 from population
    (c1, c2) ← Crossover(p1, p2)
    Mutate(c1)
    Mutate(c2)
    Insert c1 and c2 into newpoulation
  Replace population with newpopulation
  for i = 1 → N do
    fitness(pi)
```

C. Related Work

One of the first uses of genetic algorithms to aid in the cryptanalysis of a cipher was in 1993 by Richard Spillman. He chose to attack the *Knapsack cipher* using a genetic algorithm [15]. A GA was utilized to decrypt cipher text letter by letter with an average decryption time of 84 seconds making the attack a practical use of a soft computing method. In [16], a genetic algorithm was used to attack the *Chor-Rivest Knapsack public key crypto system* with very promising results in all test cases. This also happened to be the first public attack on the crypto system. The algorithm was able to attack the cipher with a minimal search space as well as a very small number of generations. The encryption algorithm uses modular multiplication as well as logarithmic functions implying that a genetic algorithm should be able to tackle the far simpler arithmetic involved in the RC4 stream cipher. Brown et al. made use of a genetic algorithm to attack a substitution permutation network. While the study was intended to find weak keys, it was later noted that a GA would be a good tool to attack the scheme [17]. The research used SwM as a genetic operator to attack the network and it was also a method proposed to attacking other relevant crypto systems such as *AES* and *3DES*. Surprisingly, no work has been done on the cryptanalysis of RC4 using genetic algorithms.

In [18], it was found that the use of a genetic algorithm could be used to break a simplified implementation of *Data Encryption Standard* (DES) called *S-DES*. The research was successful in retrieving the key used for encryption as well as performing the task in less than 1/5 the time it would take to brute force the key (remember this is S-DES and is far easier to brute force than regular DES would be). The following paper will utilize several different methods than the research just mentioned. The first area is the method used as a crossover method; the previous paper used a ring crossover method that does not respect the properties of ordered chromosomes such as permutations. The second deviation is the use of letter frequency analysis as the fitness function which is unnecessary for our GA since we are solving for the keystream and not a decrypted plaintext.

Genetic Algorithms have had a recent introduction as a form of not only cryptanalysis, but steganography as well. In [19], researchers used a genetic algorithm to encode image information in a watermark that could then be hidden or kept in plain sight within the image. The method utilized by the researchers was also highly resilient to many common attacks of detection [19]. What makes this paper relevant is that the watermarking problem is encoded as a permutation problem. As a result, the authors examine several *ordered crossover* (OX) methods including *partially mapped crossover* (PMX) and *edge recombination crossover* (ER) as well as a method called *cycle crossover* (CX) that was not chosen for this task due to its larger computational needs over the other methods proposed. The authors also experimented with both the *swap mutation method* (SwM) and the *inversion mutation method* (InvM).

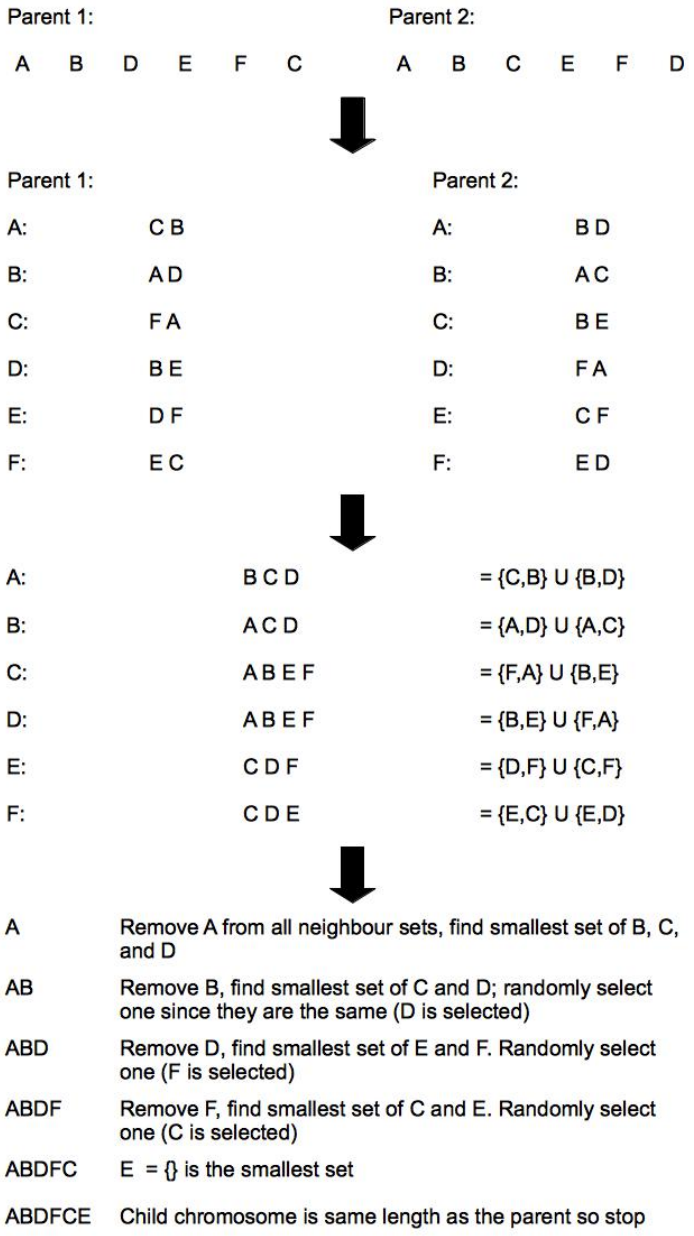


Fig. 2: Example of the steps involved in edge recombination crossover (ER)

II. Proposed Genetic Algorithm

The proposed GA will try to solve for the permutation that is represented as the state register S in RC4 using a variety of crossover and mutation methods. Due to the requirements of a permutation, special crossover and mutation algorithms must be implemented in order to not destroy the ordered property of a candidate solution. An attempt to *brute-force* the permutation would work out to $256! = 8.578\text{x}10^{506}$ possible permutations. Any improvement on this value can be seen as an improvement on such an approach.

The GA will evaluate the fitness of a candidate solution based on how well it can replicate the keystream that will be provided. Each keystream will be 256 bytes in length as suggested in [20] to find a unique solution. The fitness value will be an integer value from 0 to 256 representing how many bytes of the keystream were successfully replicated using a candidate state register.

Through various testing of different parameters for the GA, a population of size 5 has been selected to help evolve the best solution. For the proposed experiments, *tournament selection* has also been chosen as the selection method with a tournament size of 2. Finally, *elitism* will also be implemented which ensures that the best solution from a generation is carried over to the next one. This in turn does not allow the genetic operators to accidentally destroy the best solution found thus far by selection.

A. Proposed Crossover Methods

Traditional single-point and two-point crossover methods would destroy the permutation represented in the chromosome. Due to this constraint, ordered crossover methods are employed.

The first algorithm used for ordered crossover is *partially mapped crossover* (PMX). This method resembles two-point crossover in that two points are selected and the bits (bytes in our case) between them are exchanged to create two new children. It differs when the original parents are added to the child chromosomes, each value is added to the child in order of appearance until that value is already present in the child from the original crossover. At this point the duplicated value is replaced in the child chromosome by one of the values that was removed also due to the initial crossover.

The second method used is *edge recombination crossover* (ER). This method uses an adjacency matrix which is a list of each node and their respected neighbours. A master adjacency matrix is constructed by taking the union ($\cup$) of the two parent matrices. From the new matrix, a starting node is randomly selected and removed from all neighbouring sets and the node is appended to the empty child list. The next node appended is the smallest nodes set of the previous set. In the event which there are multiple sets that are the smallest, the set to use is randomly chosen. The process is repeated until the child list is the same length as the parent chromosome. A full example of ER is seen in Figure 2.

B. Proposed Mutation Methods

Similar to the crossover methods, the mutation methods employed must uphold the permutation property that exists for the candidate solution.

The first method for mutation that is called the *swap mutation method* (SwM). This form of mutation is very straight forward and is conducted by selecting two random indices in the state register and simply swapping their contents.

The second mutation method that is implemented is the *inversion mutation method* (InvM). This algorithm also preserves the permutation requirement of the chromosome. The process requires two indices in the chromosome to be selected with the sub-sequence between the two indices being simply reversed.

1) Adaptive Mutation Method: The GA will also use an adaptive mutation method. The default rate of mutation is set to 4% but there is a ceiling mutation rate also set at 15%. During the iterations, the best fitness is sampled at a predetermined rate proportional to the total amount of iterations. If the fitness appears to stagnate over these samples, the mutation rate is increased by 1% until it hits the ceiling rate in order to encourage more diversity in the population. Conversely if the fitness seems to improve over this period, the mutation rate will decrement by 1% until it is back to the original rate of mutation.

III. Experiments and Discussion

In the following section we will present the results of our experiments (see Section III-B) and evaluate the data collected (see Section III-C).

A. Equipment Used

All experiments were conducted on an Intel i7 dual-core CPU running at 2.8 GHz. The machine had 4 Gb of DDR3 RAM available to it. All three methods of exploration were programmed using C and were compiled with gcc version 4.2.1. Aside from certain programming optimizations including reducing any use of system functions that could be time intensive (i.e. *malloc()* and *free()*), optimizations at the compiler level were done using the *-O3* flag built into the gcc compiler.

B. Results

Several experiments were conducted on word sizes 6, 7, and 8. Of the results collected, PMX and SwM were shown to be the best operators for the state register problem. Further, adaptive mutation was shown to be a successful method of not allowing the candidate solutions to fall into a local optima too early. The results of utilizing both PMX and SwM with adaptive and non-adaptive mutation are exhibited in Table III.

	n					
	6		7		8	
Generations	non-adaptive	adaptive	non-adaptive	adaptive	non-adaptive	adaptive
10000	4.5	4.7	4.7	4.2	3	3.8
100000	8.4	10.6	11.6	11.3	10.5	10.5
1000000	11.8	11.8	17.4	16.9	18.1	20.7
10000000	12.8	16.2	21.4	23.9	29.5	26.2

TABLE III: Best average fitness for word size n using a genetic algorithm

The graph in Figure 4 shows that the fitness increases logarithmically and increases consistently for all sizes of the problem.

1) Comparison of Crossover Operators: Two crossover operators were examined. The operators were PMX and ER. While it was found that ER was more successful when solving the traveling salesman problem [21], when solving the RC4 permutation problem, PMX was found to be a far better technique. It is predicted that due to the requirement that the register produces the keystream consecutively, PMX disturbs the candidate solution far less than ER. Thus PMX maintains the integrity of higher fitness solutions in the crossover phase.

PMX was able to evolve a solution far better than ER by almost 10%. This is shown in Figure 5 where $n = 6$ and adaptive mutation is utilized. Both methods grow logarithmically, but the ER method just does not produce the results that the PMX achieves. Finally ER was far more time consuming when running the GA (especially for generations of 1 million and greater).

2) Comparison of Mutation Operators: The two ordered mutation operators chosen were SwM and InvM. Each of these approaches preserves the permutation property of the state register S. Experiments confirmed that SwM was a far better candidate for evolving fitness than InvM. The improvements that SwM improved the percentage fitness over InvM are seen in Figure 3. In fact, InvM seemed to reduce the best fitness found after about 1 million generations. This is believed to be the case because SwM make a number of small changes to the to the candidate solution while InvM makes a larger impact on the solution and can potentially mutate the entire solution if the indexes randomly selected were 0 and $2^n - 1$.

C. Discussion

From the results displayed in Figure 4, we were able to extrapolate an equation for the curve of $n = 8$. Using logarithmic regression we were able to derive Equation 1. This equation allowed us to determine on average how many generations it would take before 100% of the keystream could be recovered.

$$f(x) = 0.0131306222\ln(x) - 0.1065234375 \qquad (1)$$

Using Equation 1, it can be predicted that on average the keystream could be completely replicated after approximately $4.0\text{x}10^{36}$ or approximately $2^{121.5}$ generations. This attack would be a great improvements over other theoretical attacks such as [22] which had a complexity of 2^{241}. While this attack has a large complexity in terms of computer power necessary to conduct it, it is still a great improvement over $8.578\text{x}10^{506}$, which is what is needed to brute-force the state register permutation. Finally if the GA was implemented as a hardware chip, it could pump out one generation each cycle.

IV. Conclusions

In closing, we presented a genetic algorithm capable of evolving a candidate solution to try and solve the permutation of the state register S. Further, we added the capability for the GA to utilize adaptive mutation in order for the population to converge at a far slower rate since the search space is so vast to begin with ($8.578\text{x}10^{506}$). We also did a comparison of two common ordered crossover operators and two common ordered mutation operators and found that PMX out-performed ER for this type of problem while SwM proved to be a much better mutation operator than InvM for the same problem.

It has been shown that for different sizes of the permutation problem that the best fitness improves logarithmically over generations. Finally, it was derived through extrapolation that on average 100% of the keystream could be derived in about $2^{121.5}$ generations. This attack is far better than previously outlined theoretical attacks.

V. Future Work

Other operators for ordered problems such as cycle crossover should be investigated to confirm whether the ones examined in this paper were the best candidates for this particular problem. The GA could be parallelized to further improve the running time of the algorithm and allow higher amounts of generations to be tested in reasonable time.

It would also be interesting to see how well the GA fairs with other encryption algorithms which would also make an interesting study of the versatility of the GA when used for cryptanalysis.

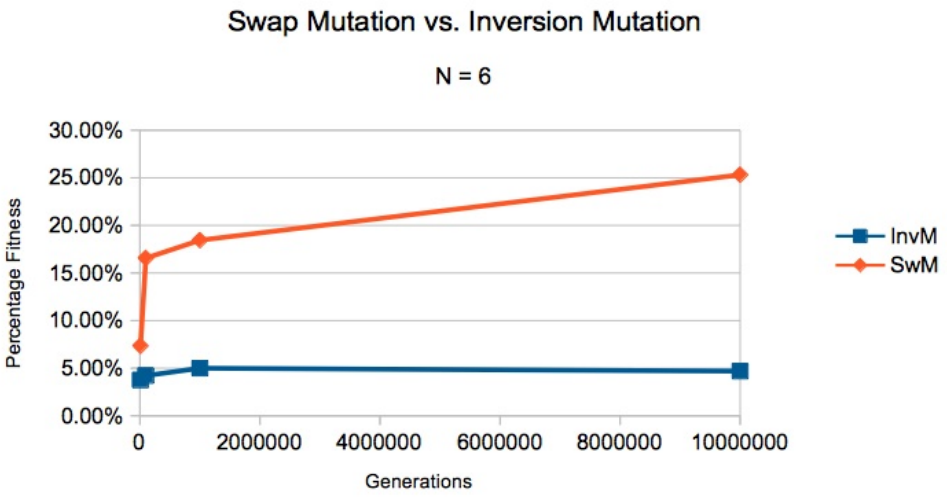


Fig. 3: Swap mutation vs. inversion mutation for $n = 6$ using adaptive mutation

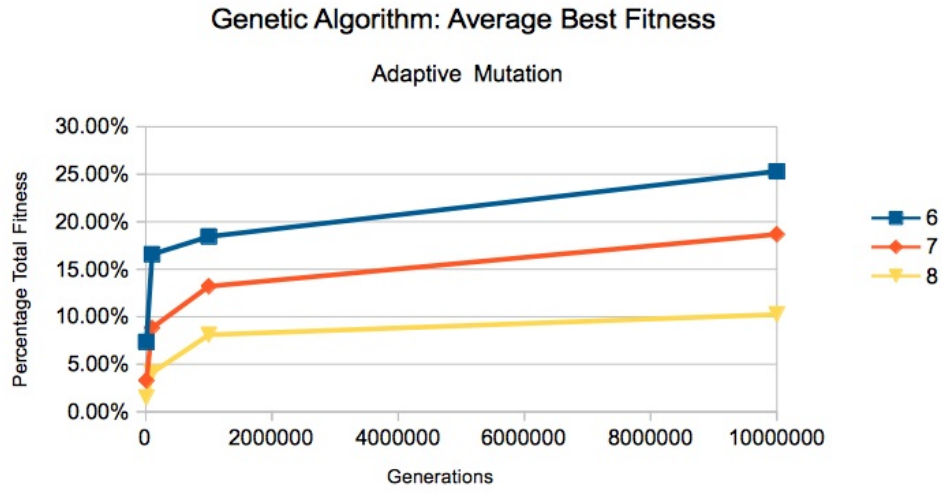


Fig. 4: Genetic algorithm best average fitness for various word sizes n

REFERENCES

[1] P. Kitsos, G. Kostopoulos, N. Sklavos, and O. Koufopavlou, "Hardware implementation of the RC4 stream cipher," in *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*, vol. 3. IEEE, 2003, pp. 1363–1366.

[2] F. A., K. P., and K. P., "RFC: 6101: The Secure Socket Layer (SSL) Protocol Version 3.0," *Internet RFC 6101*, 2011. [Online]. Available: http://tools.ietf.org/html/rfc6101.txt

[3] D. T. and A. C., "RFC: 2246: The TLS Protocol (v1)," *Internet RFC 2246*, 1999. [Online]. Available: http://www.ietf.org/rfc/rfc2246.txt

[4] A. H. and J. S., "RFC: EAP Mechanism using TLS and SASL (version 1) draft," *Internet RFC*, 2001. [Online]. Available: http://tools.ietf.org/html/draft-andersson-eap-tls-sasl-00

[5] P. J. Hancock, "Genetic algorithms and permutation problems: A comparison of recombination operators for neural net structure specification," in *Combinations of Genetic Algorithms and Neural Networks, 1992., COGANN-92. International Workshop on*. IEEE, 1992, pp. 108–122.

[6] D. R. Jones and M. A. Beltramo, "Solving Partitioning Problems with Genetic Algorithms." in *ICGA*, 1991, pp. 442–449.

[7] H.-F. Wang and K.-Y. Wu, "Hybrid genetic algorithm for optimization problems with permutation property," *Computers & Operations Research*, vol. 31, no. 14, pp. 2453–2471, 2004.

[8] P. Prinetto, M. Rebaudengo, and M. S. Reorda, "Hybrid genetic algorithms for the traveling salesman problem," in *Artificial Neural Nets and Genetic Algorithms*. Springer, 1993, pp. 559–566.

[9] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Oxford, England: U Michigan Press, 1975.

[10] J. J. Grefenstette, "Optimization of Control Parameters for Genetic Algorithms," *IEEE Transactions on System, Man, and Cybernetics*, vol. SMC-16, NO. 1, January 1986.

[11] Z. Michalewicz, "Genetic algorithms, numerical optimization, and constraints," in *Proceedings of the Sixth International Conference on Genetic Algorithms*, vol. 195. Morgan Kaufmann, San Mateo, CA, 1995, pp. 151–158.

[12] C.-Y. Lin and P. Hajela, "Genetic algorithms in optimization problems

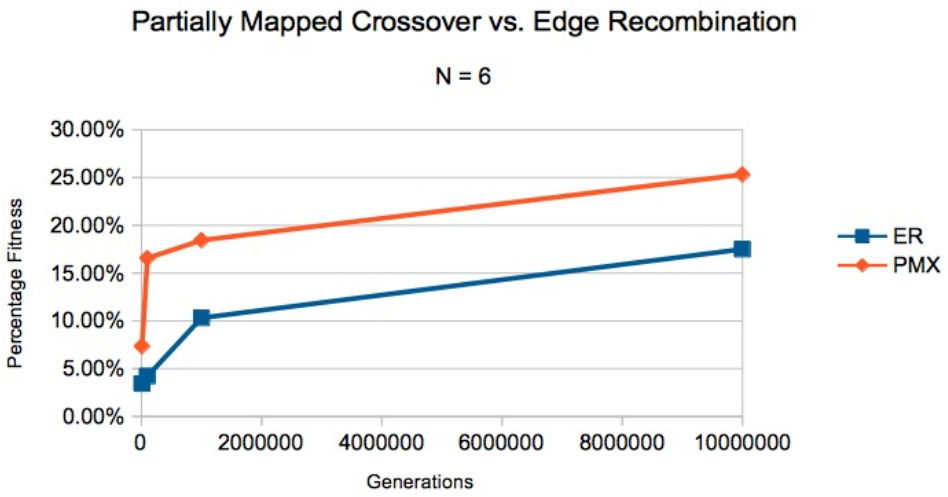


Fig. 5: Partially mapped crossover vs. edge recombination for $n = 6$ using adaptive mutation

with discrete and integer design variables," *Engineering Optimization*, vol. 19, no. 4, pp. 309–327, 1992.

[13] J. C. Bean, "Genetic algorithms and random keys for sequencing and optimization," *ORSA journal on computing*, vol. 6, no. 2, pp. 154–160, 1994.

[14] M. Mitchell, S. Forrest, and J. H. Holland, "The royal road for genetic algorithms: Fitness landscapes and GA performance," in *Proceedings of the first european conference on artificial life*. Cambridge: The MIT Press, 1992, pp. 245–254.

[15] R. Spillman, "Cryptanalysis of Knapsack Ciphers using Genetic Algorithms," *Cryptologia*, vol. 17:4, pp. 376–377, October 1993.

[16] I. Yaseen and H. V. Sahasrabuddhe, "A genetic algorithm for the cryptanalysis of chor-rivest knapsack public key cryptosystem (pkc)," in *Computational Intelligence and Multimedia Applications, 1999. ICCIMA '99. Proceedings. Third International Conference on*, 1999, pp. 81–85.

[17] J. Brown, S. Houghten, and B. Ombuki-Berman, "Genetic Algorithm Cryptanalysis of a Substitution Permutation Network," *IEEE Symposium on Computational Intelligence in Cyber Security*, pp. 115–121, March 2009.

[18] L. Sharma, R. Sharma, and B. K. Pathak, "Breaking Simplified Data Encryption Standard Using Genetic Algorithm," *Journal of Computer Science and technology*, vol. 12:5, pp. 55–59, 2012.

[19] V. Alvarez, J. A. Armario, M. D. Frau, F. Gudiel, M. B. Guemes, E. Martin, and A. Osuna, "GA based robust blind digital watermarking," *Dept. Matematica aplicada, Dept. Algebra*, pp. 376–377, 2012.

[20] G. Carter, E. Dawson, and K. Wong, "An Analysis of the RC4 Family of Stream Ciphers against Algebraic Attacks," *Proc. 8th Australian Information Security Conference (AISC 2010)*, 2010.

[21] P. Larranaga, C. Kuijpers, R. Murga, I. Inza, and S. Dizdarevic, "Genetic Algorithms for the Travelling Salesman Problem: A Review of Representations and Operators," *Artificial Intelligence Review*, vol. 13, no. 2, pp. 129 – 170, 1999.

[22] A. Maximov and D. Khovratovich, "New state recovery attack on RC4," in *Advances in Cryptology–CRYPTO 2008*. Springer, 2008, pp. 297–316.